Bo Li
Lei Shu
Deze Zeng (Eds.)



236

Communications and Networking

LNICST

12th International Conference, ChinaCom 2017
Xi'an, China, October 10–12, 2017
Proceedings, Part I

Part 1





Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 236

Editorial Board

Ozgur Akan
Middle East Technical University, Ankara, Turkey
Paolo Bellavista
University of Bologna, Bologna, Italy
Jiannong Cao
Hong Kong Polytechnic University, Hong Kong, Hong Kong
Geoffrey Coulson
Lancaster University, Lancaster, UK
Falko Dressler
University of Erlangen, Erlangen, Germany
Domenico Ferrari
Università Cattolica Piacenza, Piacenza, Italy
Mario Gerla
UCLA, Los Angeles, USA
Hisashi Kobayashi
Princeton University, Princeton, USA
Sergio Palazzo
University of Catania, Catania, Italy
Sartaj Sahni
University of Florida, Florida, USA
Xuemin Sherman Shen
University of Waterloo, Waterloo, Canada
Mircea Stan
University of Virginia, Charlottesville, USA
Jia Xiaohua
City University of Hong Kong, Kowloon, Hong Kong
Albert Y. Zomaya
University of Sydney, Sydney, Australia

More information about this series at http://www.springer.com/series/8197

Bo Li · Lei Shu
Deze Zeng (Eds.)

Communications and Networking

12th International Conference, ChinaCom 2017
Xi'an, China, October 10–12, 2017
Proceedings, Part I

Springer

Editors
Bo Li
Northwestern Polytechnical University
Xi'an
China

Deze Zeng
China University of Geosciences
Wuhan, Hubei
China

Lei Shu
Guangdong Pharmaceutical University
Guangzhou
China

ISSN 1867-8211 ISSN 1867-822X (electronic)
Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering
ISBN 978-3-319-78129-7 ISBN 978-3-319-78130-3 (eBook)
https://doi.org/10.1007/978-3-319-78130-3

Library of Congress Control Number: 2018937368

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The 12th International Conference on Communications and Networking in China (CHINACOM 2017) was successfully held in Xi'an, China, during October 10–12, 2017. The aim of CHINACOM is to bring together Chinese and international researchers and practitioners, from over the world, to exchange original ideas and share R&D advances in communications and networking.

ChinaCom 2017 consisted of 16 technical tracks and one workshop: Wireless Communications and Networking, Next-Generation WLAN, Big Data Network, Cloud Communications and Networking, Ad-Hoc and Sensor Networks, Satellite and Space Communications and Networking, Optical Communications and Networking, Information and Coding Theory, Multimedia Communications and Smart Networking, Green Communications and Computing, Signal Processing for Communications, Network and Information Security, Machine-to-Machine and Internet of Things, Communication QoS, Reliability and Modeling, Cognitive Radio and Networks, E-Health, and Workshop on Advances and Trends of V2X Networks. CHINACOM has become one of the major events in these areas in the Asia-Pacific region. It has been successful in encouraging interactions among participants, exchanging novel ideas, and disseminating knowledge.

Following the great success of the past CHINACOM events held during 2006–2016, CHINACOM 2017 received more than 178 submitted papers, out of which 112 papers were selected for presentation. The Technical Program Committee (TPC) did an outstanding job in organizing a diverse technical program consisting of 25 symposia that covered a broad range of research areas in information and communication technologies. Under the excellent leadership of TPC Co-chairs Prof. Pinyi Ren, Prof. Yunsong Li, and Prof. Chilian Chen, the TPC members handled the reviews of papers, with more than three reviews per paper on average.

The technical program featured five outstanding keynote speakers, who presented their vision of wireless communication systems in theory and practice: Prof. Jian-Nong Cao, Hong Kong Polytechnic University, Hong Kong, SAR China; Prof. Bob Li, University of Electronic Science and Technology of China, China; Prof. Jian-Dong Li, Xidian University, China; Prof. Ying-Chang Liang, University of Electronic Science and Technology of China, China; and Prof. Ping Zhang, Beijing University of Posts and Telecommunications, China.

We would like to thank the TPC co-chairs, TPC members, all the reviewers, the workshop co-chairs, the Web chairs, the publication chair, the local chairs, and all the members of the Organizing Committee, for their assistance and efforts to make the conference succeed. The continuing sponsorship by EAI and Springer is gratefully acknowledged. We also express our appreciation to the conference keynote speakers, tutorial speakers, paper presenters, and authors.

February 2018 Bo Li

Organization

Steering Committee

Imrich Chlamtac	EAI/Create-Net, Italy
Hsiao-Hwa Chen	National Cheng Kung University, Taiwan
Ya-Bin Ye	Huawei Europe Research Center
Zheng Zhou	Beijing University of Posts and Telecommunications, China
Bo Li	Hong Kong University of Science and Technology, SAR China
Andreas F. Molisch	University of South California, USA
Jun Zheng	Southeast University, China
Zhi-Feng Zhao	Zhejiang University, China

Organizing Committee

General Chair

Bo Li	Northwestern Polytechnical University, China

General Co-chairs

Pinyi Ren	Xi'an Jiaotong University, China
Yunsong Li	Xidian University, China
Chilian Chen	China Electronics Technology Group Corporation, China

TPC Chair

Lei Shu	EAI China Office, China/Guangdong University of Petrochemical Technology, China/University of Lincoln, UK

TPC Co-chairs

Shuai Wan	Northwestern Polytechnical University, China
Lin Zhang	Beijing University of Posts and Telecommunications, China
Song Xiao	Xidian University, China

Local Chair

Zhongjiang Yan	Northwestern Polytechnical University, China

Workshop Co-chairs

Lijun Qian	Prairie View A&M University, USA
Daji Qiao	Iowa State University, USA

Publicity and Social Media Chairs/Co-chairs

Der-Jiunn Deng	Changhua University of Education, Taiwan
Hongwei Zhao	Northwestern Polytechnical University, China

Publications Chair

Deze Zeng	China University of Geosciences, China

Web Chair

Mao Yang	Northwestern Polytechnical University, China

Panels Chair

Lei Wang	Dalian University of Technology, China

Sponsorship and Exhibits Chairs

Xiaoya Zuo	Northwestern Polytechnical University, China
Ming Hao	China Electronics Technology Group Corporation (CETC), 20th Institute, China
Luohui Su	Aviation Industry Corporation of China (AVIC), Xi'an Flight Automatic Control Research Institute, China
Li Yan	Guangdong University of Petrochemical Technology, China

Tutorials Chair

Weipeng Jing	Northeast Forestry University, China

Demos Chairs

Yuanfang Chen	Guangdong University of Petrochemical University, China
Jianhua Zhang	China Academy of Space Technology (CAST), Xi'an Branch, China

Poster and PhD Track Chairs

Mithun Muhkerjee	Guangdong University of Petrochemical University, China
Yunjie Yuan	China North Industries Group Corporation (CNIGC), 206th Institute

Industrial Exhibition Co-chairs

Chaoxing Su (Director)	Zhongke Industrial Park of Western Guangdong Province, China
Jun Feng (Deputy Director)	Zhongke Industrial Park of Western Guangdong Province, China
Meiquan Ou	Zhongke Industrial Park of Western Guangdong Province, China
Tianping Chen	Zhongke Industrial Park of Western Guangdong Province, China

Conference Manager

Dominika Belisova	European Alliance for Innovation (EAI)

Technical Program Committee

Track Chairs

Wireless Communications and Networking

Zheng Ma	Southwest Jiaotong University, China
Zhenyu Xiao	Beijing University of Aeronautics and Astronautics, China
Min Zhu	Xidian University, China

Special Session on the Next-Generation WLAN

Yingpei Lin	Huawei Technologies Co., China
Yunbo Li	Huawei Technologies Co., China

Big Data Network

Yong Li	Tsinghua University, China
Xiang Chen	Sun Yat-sen University, China

Cloud Communications and Networking

Chungang Yang	Xidian University, China
Yichen Wang	Xi'an Jiaotong University, China

Ad-hoc and Sensor Networks

Shihai Shao	University of Electronic Science and Technology of China
Chengbing He	Northwestern Polytechnical University, China

Satellite and Space Communications and Networking

Jinglun Shi	South China University of Technology, China
Mingwu Yao	Xidian University, China
Song Xiao	Xidian University, China

Optical Communications and Networking

Xiaoyan Pang	Northwestern Polytechnical University, China
Jie Lian	University of Virginia, USA

Information and Coding Theory

Jialiang Lu	Shanghai Jiao Tong University, China
Zhengchuan Chen	Singapore University of Technology and Design, Singapore

Multimedia Communications and Smart Networking

Fuzheng Yang	Xidian University, China
Hui Yuan	Shandong University, China

Green Communications and Computing

Jianchao Du	Xidian University, China
Guomei Zhang	Xi'an Jiaotong University, China
Yiming Liu	China Academy of Electronics and Information Technology, China

Signal Processing for Communications

Hongxing Guo	Huazhong University of Science and Technology, China
Wentao Shi	Northwestern Polytechnical University, China
Qinghe Du	Xi'an Jiaotong University, China

Network and Information Security

Yongqiang Hei	Xidian University, China
Chengkai Tang	Northwestern Polytechnical University, China
Pinyi Ren	Xi'an Jiaotong University, China

Machine-to-Machine and Internet of Things

Shaohui Mei	Northwestern Polytechnical University, China
Yongqian Du	Northwestern Polytechnical University, China

Communication QoS, Reliability, and Modeling Symposium

Zhong Shen	Xidian University, China
Guifang Li	Northwestern Polytechnical University, China

Cognitive Radio and Networks Symposium

Yong Niu	Beijing Jiaotong University, China
Li Sun	Xi'an Jiaotong University, China

e-Health

Zhongjin Liu	Coordination Center of China
He Guan	Northwestern Polytechnical University, China

Contents – Part I

Wireless Communications and Networking

A Throughput Enhanced JTIDS Waveform Design 3
Yingjie Song, Meng Zhang, and Ming Hao

Design and Application of Aperture-Coupled-Patch Frequency Selective Surfaces 13
Yu Wang

Radar Cross Section Reduction Based on Metasurface 21
Xing-Long Zhang, Meng Niu, Luo-Hui Su, and Ke-Pu Song

Joint Direction-of-Departure and Two-Dimensional Direction-of-Arrival Estimation with Automatic Pairing in Massive MIMO Systems. 27
Yuan Zhou, Xiang Ji, and Man Dai

Non-orthogonal Multi-carrier Technology for Space-Based Internet of Things Applications. 37
Mengdi Zhao, Hongxing Li, Yi Li, Li Fang, and Ping Chen

A Novel GPS Spoofing Algorithm Based on Modifying Navigation Message 46
Qiong Yang, Yi Zhang, and Chengkai Tang

Implementation of Narrowband Interference Suppression Technology for DSSS Systems in Time Domain and Frequency Domain. 54
Xiongfei Li, Xiangyang Meng, and Pan Liu

Spinal Codes Based Efficient Rateless Transmission Scheme for Massive MIMO System with Imperfect CSI 64
Liangliang Wang, Yin Sun, Xuming Lu, Xiang Chen, and Hongzhou Tan

Low-Cost Topology Design for Wireless Ad Hoc Networks with Connectivity Constraint 75
Mengmeng Xu, Qinghai Yang, and Hengzhou Xu

Waveform Design of UNB Modulation and Performance Analysis 84
Aoqin Duan, Qinghe Du, and Bo Sun

System-Level Simulation Platform for Device-to-Device Communications in 5G Networks 95
Zhihao Ding, Pinyi Ren, and Qinghe Du

Joint Algorithm for MIMO-OFDM System Based on the Idea of Layered Iteration 106
Juan Wei, Shixuan Feng, and Lihong Wang

Improved UWB Indoor Positioning Algorithms Based on BP Neural Network Model. 114
Haojie Liu, Zhonghua Liang, Danli Liu, and La'ning Ma

Performance of Transmitted Reference UWB Communication System with Time-Hopping Multiple Access 125
Danli Liu, Wenjing Wu, Zhonghua Liang, Jinjin Liu, and Xiaojun Yang

Improved Ultra-wideband Pulse Shaping Technique Based on Spectrum Shifted Gaussian Waveforms 135
Danli Liu, Zhonghua Liang, La'ning Ma, and Huansheng Song

An Uplink SCMA Scheme Combined with Interleaving. 146
Fei Yang and Zhisong Bie

Integer-Forcing Detection of Generalized Spatial Modulation 155
Guoquan Li, Ying Zhou, Xiangyun Zhou, Jinzhao Lin, Yu Pang, Jiacheng Wang, and Hongyu Zhu

Design of Dynamic TDMA Protocols for Tactical Data Link 166
Lei Liu, Yiming Liu, Zhaowei Wang, and Chunxu Liu

Estimation Under Staggered Frame Structure for Multi-cell Multi-user Massive MIMO Systems 176
Mengjun Zhang, Gang Xie, and Jinchun Gao

A Time-Domain Combined Anti-jamming Preprocess Method for DOA Estimation in DSSS System 186
Shuyun Li, Sheng Ke, and Yuan Chai

An Efficient Wireless Backhaul Algorithm for User-Centric Ultra-Dense Networks 195
Hong Feng, Xi Li, and Hong Ji

Stochastic Optimal Control for Power Allocation and Link Selection in Energy-Harvesting Three-Node Relay Network 206
Qianyun Gong, Qinbo Chen, Hui Cai, Maoxin Tian, and Sai Zhao

Parallel Symbol Timing Recovery Using FPGA for 600 Msps QPSK 216
Di Huang, Zhijie Wang, Jun Wang, and ZiYao Liu

Joint Fingerprinting and Curve Fitting Scheme with NLOS Identification for Indoor Localization 226
Ruiyang Yuan, Xiaohui Li, Yanbin Zhao, Bin Zhou, and Si Huang

Fairness-Aware Scheduling for Millimeter Wave Backhaul Networks 240
Danfeng Meng, Xiaohui Li, Yanbin Zhao, Bin Zhou, and Dantao Li

Lyapunov Optimization for NOMA Wireless Transmission with Energy Harvesting and Storage 254
Qinbo Chen, Qianyun Gong, Maoxin Tian, Hui Cai, and Xiancai Chen

SEA: SDN-Based Evolution Architecture for 5G Network 264
Chen He and Xiaofeng Qiu

A Method of Link Margin Analysis for Aeronautical Communication Systems in LOS Environment 274
Guofeng Jiang, Yangyu Fan, Hongbo Yuan, and Gang Zhou

Study on Handover of High Speed Railway by Rate-Splitting and Interference Cancellation 284
Bo Zhang, Hui Liu, Hai Zhu, and Hengzhou Xu

Satellite and Space Communications and Networking

A Novel Wavelet Denoising Algorithm Based on 2-Fold Cross Validation for BDS GEO Code Multipath Bias 297
Peng Wu, Baowang Lian, Zesheng Dan, and Qiong Yang

Multi Block Overlapping Zero Padding Algorithm for Weak Signal Acquisition 307
Jinhua Sun, Shanshan Lv, and Hao Wang

A Partially Repeated Scheme for Polar Codes 317
Daolong Wu

Performance Bound Analysis on Hamming-Weight-Analysis Algorithm for Blind Recognition of Linear Block Codes 323
Di Mao

An Efficient Algorithm for Refining Position and Velocity Outputs of Space Borne GNSS Receivers 332
Shuhao Chang, Xi Chen, and Menglu Wang

Fountain-Coding Aided Symmetrical Encryption for Secure Wireless Transmission 341
Chao Lyu, Pinyi Ren, Qinghe Du, and Li Sun

A New Anti-blackout Communication Method Based on Carrier Aggregation of OFDMA and Frequency Diversity 351
Jieqi Li, Xiaoya Zuo, Yong Zhang, Zhengxue Li, Shaohui Mei, and Zhongqi Tian

Real-Time Indoor Positioning System Based on Background Training Model Using IR-UWB Radar 361
Hui-Seon Gang, June-Buem Park, and Jae-Young Pyun

Bandwidth Adaptive Image Communication via Similarity Based Auto-Selection 372
Jieqi Li, Mingyang Ma, Yong Zhang, Zhengxue Li, Shaohui Mei, and Shuai Wan

Hyperspectral Image Fusion Based on Multistage Guided Filter 381
Wenqian Dong, Song Xiao, Jiahui Qu, and Lizhao Li

Adaptive Long-Term Object Tracking for Real-Time Visual Search in Multimedia Surveillance 388
Lei Li, Song Xiao, and Feiyang Tan

Multi-objective Optimization and Design of LEO Satellite Broadband Networks 398
Dawei Yan, Cong Liu, Peng You, and Shaowei Yong

Routing Protocol Design for LEO Satellite Networks with Uni-directional Links 409
Juan Wei, Lihong Wang, and Shixuan Feng

Measuring Quality of Experience of Novel 360-Degree Streaming Video During Stalling 417
Wenjie Zou and Fuzheng Yang

A Realizable and Efficient Approach for Building Software Defined Satellite Network 425
Wei Ma, Leifang Hui, Hanwen Sun, and Zhe Zhao

Proactive Edge Caching Strategy Based on Mobility Prediction in Dense Small Cell Networks 433
Tianming Song, Heli Zhang, Xi Li, Chunsheng Zhu, and Hong Ji

The Minimum of GDOP for Satellite Navigation System 443
Xuetao Yang, Jianchao Du, and Weibin Li

A SDN Security Transmission Service Using Multi-connection and Time-Slot Scheduling 448
Shiqi Zhou, Rongheng Lin, Shuquan Fan, Hua Zou, and Budan Wu

Construction of Quasi-Cyclic LDPC Codes with Diagonal Structure of Parity-Check Matrices 460
Hengzhou Xu, Huaan Li, Min Zhu, Bo Zhang, and Baoming Bai

Analyzing the Impact of Redundant Paths on Multistage Switch Performance 470
Gao Ya, Pan Weitao, and Zheng Ling

Big Data Network Track

Positive and Unlabeled Learning for Mobile App User and Server Interaction Prediction 481
Binbin Wang, Ke Yu, Xiaofei Wu, Fang Wei, Wan Jiang, and Di Pan

Game-Based Mobile Traffic Sharing with WiFi 492
Rong Wang, Zhifei Wang, Jibing Wu, Hongbin Huang, Su Deng, and Yahui Wu

Continuous Cell Zooming Algorithm Towards Energy Efficient in Random Heterogeneous Cellular Networks 502
Jie Sheng, Yang You, Dong Ma, Canyan Zhu, and Fan Xu

D2D Assisted MU-MIMO Precoding to Enhance LTE Network 512
Yanqiu Zhang, Shu Fang, Yuanchao Han, and Yu Zeng

High Concurrent Elastic Resource Allocation in Hadoop YARN 524
Yang Peng, Danyan Luo, Jian Dong, and Zhibo Wu

An Autoencoder-Based Learning Method for Wireless Communication Protocol Identification 535
Jie Ren, Zulin Wang, and Mai Xu

A Novel Path Selection Algorithm Based on Graph-Cut to Ensure Security in SDN 546
Shuquan Fan, Rongheng Lin, Shiqi Zhou, Hua Zou, Budan Wu, and Yun Su

Secure and Reliable Transmission for Dense HetNets Based on Secure Energy Efficiency Optimization 557
Yunjia Xu, Kaizhi Huang, and Yi Zou

Author Index 567

Contents – Part II

Multimedia Communications and Smart Networking

WSN Routing Algorithm Based on Energy Approximation Strategy 3
Yuexia Zhang, Hang Chen, and Tianrun Gao

SIFT Based Monocular SLAM with GPU Accelerated 13
Tonghui Wang, Guoyun Lv, Shikai Wang, Haili Li, and Baicen Lu

Combining Computer Graphics and Feature for 3D Camera Tracking Based on CAD Model 23
Linlin Wang, Guoyun Lv, Ningxin Zhang, and Yanggege Yu

Performance Analysis of CBRP, AODV and DSR Routing Protocols in VANETs Based on IDM-IM 33
Luqi Wu and Xia Wang

DOA Estimation for Far-Field Sources in Mixed Signals with Gain-Phase Error Array 41
Jiaqi Zhen and Yanchao Li

Sparse Reconstruction in Frequency Domain and DOA Estimation for One-Dimensional Wideband Signals 51
Jiaqi Zhen and Yanchao Li

Radar Burst Control Based on Constrained Ordinal Optimization Under Guidance Quality Constraints 61
Bo Li, Qingying Li, and Daqing Chen

An Overview of CCMANET: Content Centric MANET 72
Yuehua Huo, Weiqiang Fan, Yinlong Liu, and Dong Li

A Cooperative Broadcast Algorithm Based on the Successful Broadcasting Ratio and Residual Energy of Neighbor Nodes in Mobile Ad Hoc Networks 82
Yichen Deng, Yanping Yu, and Zhengqing Yan

Signal Processing for Communications

Joint User Association and ABS for Energy-Efficient eICIC in Heterogeneous Cellular Network 95
Jie Zheng, Ling Gao, Hai Wang, Jinping Niu, Xiaoya Li, and Jie Ren

Transmissions and Network Management of Multiple 100 GB/s Based on Stacking Technology 102
Xin Huang, Kaixiong Zhou, Chaoxiang Shi, Jianxin Chang, and Meng Gao

Topology-Transparent Scheduling in Mobile Ad Hoc Networks with Unidirectional Links 112
Yiming Liu, Zhi Fu, and Mao Yang

Carrier Synchronizator in Nonbinary LDPC Coded Modulation Systems 122
Qiang Zhang, Zhong-yang Yu, and Baoming Bai

An Assessment Method of Pilot Situation Awareness in Manned/Unmanned-Aerial-Vehicles Team 132
Jun Chen, Qilin Zhang, Xunjie Qiu, and Bo Hou

Minimum Interference Beam Selection for Millimeter Wave BeamSpace MIMO System 141
Dantao Li, Xiaohui Li, Bin Zhou, Yanbin Zhao, and Danfeng Meng

An Efficient Algorithm for Constructing Underwater Sensor Barrier 153
Weiqiang Shen, Chuanlin Zhang, Min-Rong Chen, Jinglun Shi, and Guo-Qiang Zeng

Distributed Coverage Hole Detection Algorithm Based on Čech Complex . . . 165
Wang Yuchen, Lu Jialiang, and Philippe Martins

An Energy-aware Routing Protocol Based on Network Coding for Wireless Sensor Networks 176
Panpan Cui, Song Xiao, and Lei Li

Detecting Hierarchical and Overlapping Community Structures in Social Networks Using a One-Stage Memetic Algorithm 182
Chun-Cheng Lin, Der-Jiunn Deng, Jung-Chao Wu, and Liang-Yi Lu

Optimal ZF Precoder Under per Antenna Power with Conjugate Beamforming for MU Massive MIMO Systems 189
James Kweku Nkrumah Nyarko and Christian Ango Mbom

Butterfly-Flow-Graph Based MAP Decoding Algorithm for Channel Quality Information in 3GPP-LTE 199
Qi Li, He Wang, Yunchuan Yang, Bin Yu, and Chengjun Sun

Group-Based Layered Scheduling of ADMM Decoding for LDPC Codes . . . 207
Xing-Long Zhang, Meng Niu, Luo-Hui Su, and Ke-Pu Song

LUT-Based Efficient Impulse Shaping for Direct Synthesizing Digital Communication Signals at Arbitrary Symbol Rate 215
Ziyao Liu, Zhijie Wang, Jun Wang, Di Huang, and Nangen Zhang

An Adaptive Step-Size Prediction Joint OMP Algorithm for Beam Tracking in Millimeter Wave Systems 227
Xu Yang, Xiaohui Li, Bin Zhou, Yanbin Zhao, and Ruiyang Yuan

Priority-Based Multi-carrier Access Schemes for Safety Message Transmission in Vehicular Networks 239
Huichao Wang, Longjiang Li, Fan Wu, Ke Zhang, and Supeng Leng

V2V Resource Allocation Schemes for Non-safety Service in Cellular Vehicular Networks 250
Lei Jiang, Fan Wu, Ke Zhang, and Lixiang Ma

Adaptive Slot Assignment for TDMA Based Dynamic Airborne Ad Hoc Networks 260
Yueyan Qian, Mingwu Yao, and Liang Zhang

On Low-Pass Phase Noise Mitigation in OFDM System for mmWave Communications 271
Xiaoming Chen, Wei Fan, and Anxue Zhang

Network and Information Security

Secure Algorithm via Hybrid Relaying Scheme and Resource Allocation for OFDM Networks 283
Xianwen Zhou, Pinyi Ren, and Qinghe Du

Dynamic Group Behavior Analysis and Its Application in Network Abnormal Behavior Detection 292
Yan Tong, Jian Zhang, Wei Chen, Mingdi Xu, and Tao Qin

A Simplifying Logic Approach for Gate Level Information Flow Tracking . . . 302
Yu Tai, Wei Hu, Dejun Mu, Baolei Mao, Lantian Guo, and Maoyuan Qin

A Cloud-Based Distance Bounding Protocol for RFID Conforming to EPC-C1 G2 Standards 312
Zhenjiang Dong, Xinluo Wang, Miao Lei, Wei Wang, and Hui Li

Research on the Monitoring Method of the Road Communication Network Quality Based on Vehicle Borne Internet of Things 323
Baihua Ji, Xiao Liu, Tenghui Ke, Rongjie Kuang, Zibin Gao, and Daoce Wang

A Negotiation-Based Collision Avoidance Scheme for Autonomous Mobile Robots 331
Zhaoxuan Chen, Pin Lv, Siyu Pan, and Jia Xu

Intelligent Access Scheme for Internet of Things Supported by 5G Wireless Network 341
Yingshuan Song, Heli Zhang, Xi Li, Chunsheng Zhu, and Hong Ji

Fast Group Paging Algorithm for Large-Scale MTC Systems 352
Si Huang, Xiaohui Li, Bin Zhou, Yanbin Zhao, and Ruiyang Yuan

A Honeyfarm Data Control Mechanism and Forensic Study 362
Wei Yin, Hongjian Zhou, and Chunlei Yang

Environment-Related Information Security Evaluation for Intrusion Detection Systems 373
Ran Cheng, Yueming Lu, and Jiefu Gan

A Novel Accurate Source Number Estimation Method Based on GBSA-MDL Algorithm 383
Taha Bouras, Di He, Fei Wen, Peilin Liu, and Wenxian Yu

Communication QoS, Reliability and Modeling

A Green Load Balancing Algorithm for Dynamic Spatial-Temporal Traffic Distribution in HetNets 395
Jichen Jiang, Xi Li, and Hong Ji

Research on the Construction of Radio Environment Map (REM) Based on Spatial Interpolation 405
Ran Zi, Jun Chang, Rong Zong, Ruonan Wang, and Guiwen Liao

Resource Allocation in OFDM-Based Cognitive Radios Under Proportional Rate Constraint 416
Hongzhi Wang, Yongfei Yan, and Mingyue Zhou

Capacity Analysis of Secondary User System in Cognitive MIMO Networks Based on NOMA 426
Han Liao

A Novel Energy Aware Coded Opportunistic Routing for Social Cognitive Radio Networks 440
Shoubai Xiao, Renhao Lu, and Xiaoxiong Zhong

Research on Anti PUE Attack Based on CAF Spectrum and Repeated-Game 450
Hong Chang and Yong Li

Multi-modal Transmission Strategies with Obstacle Avoidance for Healthcare Applications 463
Tiong Hoo Lim

A BP Neural Network Based Self-tuning for QoS Support in AVB Switched Ethernet 474
Dong Chen and Ang Gao

Optimal Power Splitting in a Full-Duplex Wireless Powered Network with a Bidirectional Relay 485
Yingting Liu, Hongwu Yang, Chunman Yan, and Li Cong

Advances and Trends of V2X Networks

Design and Implementation of an Intelligent Shipborne Terminal System 497
Wenqiu Chen, Rongbin Yao, Jincai Ye, Xin Tang, and Xiaohuan Li

A Positioning Method Based on RSSI and Power Spectrum Waveform Distinction 509
Yuyang Lin, Zunwen He, Jiang Yu, and Yan Zhang

A Network Coding Collaborative Download Scheme for Platoon-Based VANETs 519
Jincai Ye, Xuewen Huang, Xiaohuan Li, Guan Wang, Rongbin Yao, and Xin Tang

Analysis of Conditional Connectivity Based on Two Lanes for VANETs 532
Bin Pan and Hao Wu

Optimized Workload Allocation in Vehicular Edge Computing: A Sequential Game Approach 542
Dongdong Ye, Maoqiang Wu, Jiawen Kang, and Rong Yu

Navigation Route Based Stable Clustering for Vehicular Ad Hoc Networks 552
Zhiwei Yang, Weigang Wu, Yishun Chen, Xiaola Lin, and Xiang Chen

A Social-Aware Routing Protocol for Two Different Scenarios in Vehicular Ad-Hoc Network 563
Zhixiao Zhu, Binjie Hu, Haina Song, and Xiaotao Liu

Author Index 577

Wireless Communications and Networking



A Throughput Enhanced JTIDS Waveform Design

Yingjie Song(✉), Meng Zhang, and Ming Hao

China Electronics Technology Group Corporation (CETC), 20th Institute, Xi'an 710068, China
yingjiesongll@163.com

Abstract. JTIDS is the active service Data-Link of the USA and is playing an important role in modern warfare. However, the low throughput of 115 kbps becomes its bottle neck. In this paper, a throughput enhanced JTIDS waveform with max throughput of 1.1 Mbps is proposed. Multi-rate LDPC and GMSK are adopted to achieve the throughput. To prove its advantages, the performance of different waveforms is compared by simulation. The results show that the throughput enhanced JTIDS waveform is excellent.

Keywords: JTIDS waveform · Throughput enhance · Multi-rate LDPC GMSK · N-SISO

1 Introduction

Joint Tactical Information Distribution System (JTIDS), which is currently employed by the United States Army, is an integrated system of communication, navigation and identification. In modern warfare, information warfare is critical while the maximum data rate of JTIDS is 115 kbps, which is far lower than the request of image and video service. To improve the data throughput of JTIDS, One example is Link-16 Enhanced Throughput (LET), which works by replacing the spread spectrum and Reed Solomon (RS) encoding of the original JTIDS waveform with a RS/convolution coding scheme which can adapt to required link capability [1–3]. However, it increases in data throughput at the expense of both jamming resistance and transmission range. [4] employed MSK and LDPC coding to improve data throughput, but the performance is not good enough due to the method of MSK demodulation. [5, 6] proposed Partitioned Cyclic Code Shift Keying (PCCSK), which was combined with various-length RS codeword (adaptive channel coding) to increase the data throughput for JTIDS. Although [5, 6] improve the performance of JTIDS in some way, the data throughput is still lower than the expected.

In this paper, a design of throughput enhanced JTIDS waveform is proposed. To be compatible with active service JTIDS waveform, the message synchronization and

Y. Song—Female, senior engineer of China Electronics Technology Group Corporation (CETC), 20th Institute, major in telecommunication.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 3–12, 2018.
https://doi.org/10.1007/978-3-319-78130-3_1

message header of the throughput enhanced JTIDS waveform are kept the same. For the message data, QC-LDPC and GMSK are adopted as channel coding and modulation respectively. To improve the performance, Noncoherent Soft In Soft Out (N-SISO) algorithm is employed. In addition, a new frame structure is constructed. Two Decision-Directed (DD) symbols are inserted into the message data of each hop to further improve the demodulation performance. This design can not only greatly improve the data throughput, but also perform very well in anti-interference.

2 Active Service JTIDS Waveform

The number of JTIDS data pulses is dependent upon the "packing structure". There are three types of data packing that can be used in each time slot: Standard (3 words), Packed-2 (6 words), and Packed-4 (12 words). Among these three data packing, Packed 4 Single Pulse (P4SP) have 372 pulses of data, which gets maximum throughput of 115 kbps.

JTIDS waveform uses (31,15) RS encoding and cyclic code-shift keying (CCSK) modulation, where each encoded symbol consists of five bits and is spread into 32 chips per symbol for transmission using Minimum-Shift Keying Modulation (MSK).

As seen in Fig. 1 the top branch is the model of a JTIDS transmitter [7]. The major processes involved to generate a JTIDS waveform includes bit-to-symbol mapping, RS encoding, pseudo-random symbols interleaving, 32-ary CCSK baseband symbol modulation (each 5-bit coded symbol is converted into a 32-chip non-orthogonal sequence), PN sequence scrambling, MSK waveform modulation on a chip-by-chip basis, and Frequency Hopping (FH). After up-conversion, the signal is amplified, filtered, and transmitted over the channel.

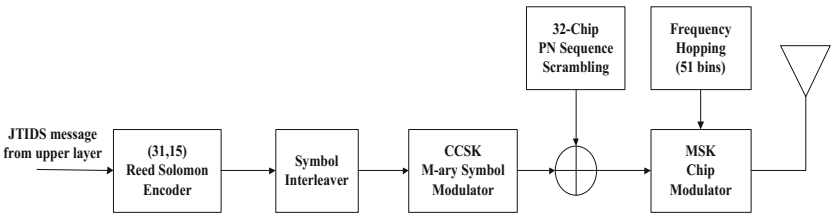


Fig. 1. JTIDS system model

3 The Proposed Throughput Enhanced JTIDS Waveform

To be compatible with active service JTIDS waveform, the message synchronization and message header of the throughput enhanced JTIDS waveform are kept the same. For the message data, LDPC is chosen to replace the RS code and CCSK, and GMSK is adopted to replace MSK. For the throughput enhanced JTIDS waveform, the number of symbol which is used to modulated by GMSK is also 32. Unlike JTIDS, each modulated symbol of proposed waveform is consisted of 30 bit coded symbols and 2

known Decision-Directed (DD) symbols. Simply, the DD symbol can be set to be 0. By adding DD symbol in each hop, the performance of GMSK demodulation is improved. To increase the data throughput, the maximum number of data pulse in each time slot is increased from 372 to 432, and three data rate, 276 kbps, 552 kbps, 1.1 Mbps are proposed to fit different channel conditions. The frame structure of hop data of the proposed throughput enhanced JTIDS waveform is illustrated in Fig. 2.

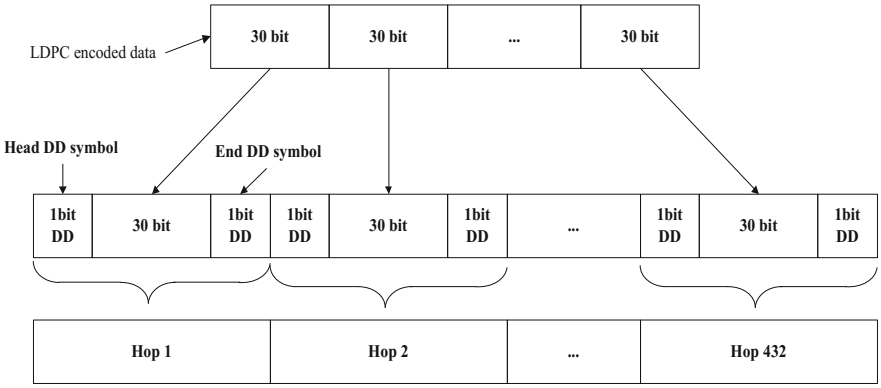


Fig. 2. Frame structure of the proposed waveform

3.1 Multi-rate LDPC

The reliable communication in a variable or interfered channel needs a flexible LDPC code. High rate code is employed in channel of good conditions to improve the efficiency. On the contrary, low rate code is used to make sure the reliability. To decrease the complexity of the system, a multi-rate coding with fix code length is designed in this paper.

In the proposed waveform, we choose a series of LDPC codes whose parity-check matrices process approximate left lower triangular form to encode systematically using the parity-check matrices directly [8]. First, a high rate (2/3) irregular repeat accumulate code is constructed as master code. Then based on the master code, code rate is decreased by decreasing the information length and increasing the check length at the same time. Finally, the multi-rate (2/3, 1/3, 1/6) QC-LDPC of fix code length is illustrated in Table 1.

Table 1. Parameters of multi-rate LDPC

Rate (R)	Code length N (bit)	Information K (bit)	Highest row weight of parity-check matrix
2/3	12960	8640	11
1/3	12960	4320	6
1/6	12960	2160	5

Likelihood ratio normalized minimum sum algorithm (LLR-Normalized MSA) is chosen in this paper as decode algorithm.

3.2 Nonrecursive GMSK Modulator

GMSK adds a Gaussian filter before the MSK modulation and results in a smoother phase path and faster attenuation of side lobe of power spectrum. The parameters of this paper is chosen as follow: modulation index $h = 0.5$, normalized 3 dB bandwidth $BT_b = 0.5$, memory length $L = 2$.

GMSK is a kind of CPM (Continuous Phase Modulation) with binary modulation $M = 2$. There are two decomposition of CPM modulator, recursively and nonrecursively. It has been shown that the nonrecursive decomposition can significantly weaken the error correlation in the demodulator and improve the demodulation performance [9]. It is also exhibited that LDPC coded nonrecursive binary CPM can obtain about 2 dB gains in E_b/N_0 when compared with recursive binary CPM [8]. Therefore, we consider using nonrecursive GMSK to reduce the propagation of errors in the demodulator.

Figures 3 and 4 give recursive and nonrecursive decomposition of GMSK modulator respectively.

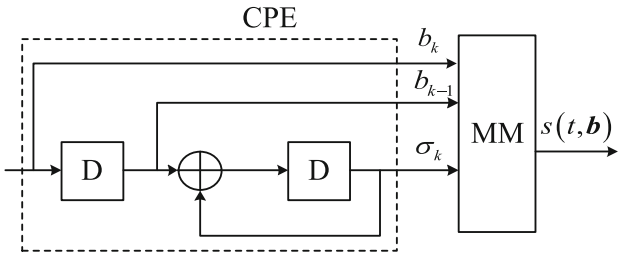


Fig. 3. Recursive decomposition of GMSK modulator

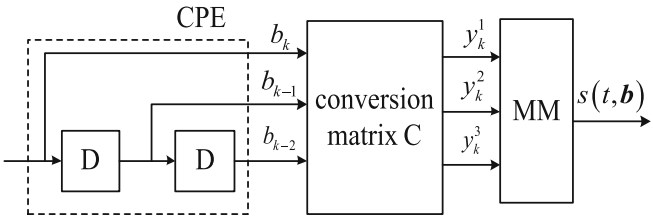


Fig. 4. Nonrecursive decomposition of GMSK modulator

In Fig. 3, the generator polynomial of CPE is $G_R = [1, D, D^2/(1 - D)]$. Suppose that for MM, the input is $\left(y_k^1, y_k^2, y_k^3\right) \Leftrightarrow (b_k, b_{k-1}, \sigma_k)$, then the nonrecursive form is defined as $(1 - D)G_R = [1 - D, D - D^2, D^2] = G_{NR}C$. And in Fig. 4, the generator polynomial of nonrecursive CPE is $G_{NR} = [1, D, D^2]$, so the conversion matrix is

$$C = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \tag{1}$$

According to

$$\left[y_k^1, \quad y_k^2, \quad y_k^3 \right] = \left[b_k, \quad b_{k-1}, \quad b_{k-2} \right] \bullet C \tag{2}$$

y_k^1, y_k^2 and y_k^3 is given by

$$y_k^1 = (b_k - b_{k-1}) \bmod 2,\ y_k^2 = (b_{k-1} - b_{k-2}) \bmod 2,\ y_k^3 = b_{k-2} \tag{3}$$

Then, according to $\left(y_k^1, y_k^2, y_k^3\right) \Leftrightarrow (b_k, b_{k-1}, \sigma_k)$, the tilted phase of nonrecursive GMSK is given by:

$$\varphi(\tau + kT, \mathbf{b}) = R_{2\pi}\left[\begin{array}{l} 2\pi h b_{k-2} + 4\pi h(((b_k - b_{k-1}) \bmod 2)q(\tau) + ((b_{k-1} - b_{k-2}) \bmod 2)q(\tau + T)) \\ \quad + \pi h(M-1)\tau/T - 2\pi h(M-1)(q(\tau) + q(\tau + T)) + (L-1)(M-1)\pi h \end{array}\right] \tag{4}$$

The input symbol sequence $a_k \in \{-1, +1\}$, $b_k = a_k/2$ and $q(t)$ is the phase response.

3.3 N-SISO Demodulation of GMSK

Note that the JTIDS waveform is noncoherently detected at the chip level due to the fast hopping rate. As the initial phase of each hop pulse is not certain, noncoherent detection makes itself an attractive strategy in practice without requiring explicit phase estimation. Moreover, by adjusting the parameters, the noncoherent detection algorithm can achieve near coherent performance [10]. The paper develops an improved algorithm of Noncoherent Soft In Soft Out (N-SISO) with near optimum performance and lower complexity.

Let the state vector $S \triangleq (b_{k-2}, b_{k-1})$ in k-th interval. Noncoherent detection of GMSK trellis diagram is shown in Fig. 5.

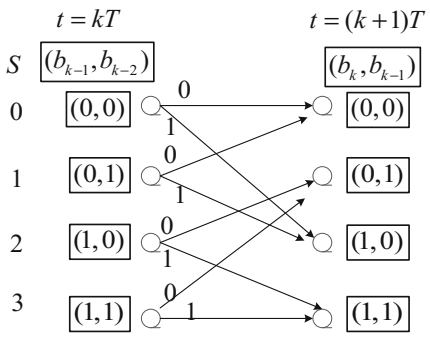


Fig. 5. GMSK trellis diagram

The simplified N-SISO algorithm is summarized below [10].

Step1: Initialization

Forward accumulated metric $\alpha_k(S_k)$ is set to be

$$\alpha_0(S_0 = S) = \begin{cases} const & S = 0 \\ 0 & others \end{cases} \quad const \gg 0 \tag{5}$$

Backward accumulated metric $\beta_N(S_N)$ is set to be

$$\beta_N(S_N = S) = \begin{cases} const & S = 0 \ or \ 2 \\ 0 & others \end{cases} \quad const \gg 0 \tag{6}$$

Phase reference symbol $q_k(S_k)$ is set to be.

$$q_0(S_0 \in S) = 0. \tag{7}$$

Step2: Forward Recursion (k from 0 to $N - 1$)

For all possible transitions $S_k \rightarrow S_{k+1}$, calculate all branch metrics $\gamma_k(S_k, S_{k+1})$, Then update $\alpha_k(S_k)$ and $q_k(S_k)$, where

$$\begin{aligned} \gamma_k(S_k, S_{k+1}) &= \left|q_k(S_k) + \mathbf{r}_k \mathbf{s}_k^H\right| - |q_k(S_k)| \\ a_{k+1}(S_{k+1}) &= \max_{S_k}(a_k(S_k) + \gamma_k(S_k, S_{k+1})) \\ q_{k+1}(S_{k+1}) &= \alpha \bullet q_k(S_k) + F_0 \mathbf{r}_k \mathbf{s}_k^H \end{aligned} \tag{8}$$

Here $\alpha \in (0, 1)$ is the so-called forgetting factor. F_0 is used to improve performance of demodulation. When $kT \leq t < (k + 1)T$, $\mathbf{r}_k$ means the received data and $\mathbf{s}_k$ means GMSK signal. $()^H$ denotes complex conjugation.

Step3: Backward Recursion (k from $N - 1$ to 0)

Update backward-accumulated metric $\beta_k(S_k)$:

$$\beta_k(S_k) = \max_{S_{k+1}}\left(\beta_{k+1}(S_{k+1}) + \gamma_k(S_k, S_{k+1})\right) \tag{9}$$

Step 4: Output Soft Metric (k from 0 to $N - 1$)

$$P(O, a_k) = \max_{(S_k, S_{k+1})}\left(\alpha_k(S_k) + \gamma_k(S_k, S_{k+1}) + \beta_{k+1}(S_{k+1})\right) - SF * \ln p(I, a_k)\Big) \tag{10}$$

Where $p(I, a_k)$ is prior information, here $\ln p(I, a_k) = 0$, $SF \in (0, 1)$ is optimized by trial and error.

4 Simulation and Comparison

In this section, the performance of the proposed waveform scheme and traditional one are compared by simulation. The simulations were performed under AWGN channel and burst-error channels.

Figure 6 gives the performance curse of the throughput enhanced waveform proposed in this paper under AWGN channel, and compares with the performance of P4SP mode of JTIDS, rate 1.1 Mbps mode of Tsinghua university and the rate 1.18 Mbps mode of MIDS-LVT (Multifunctional Information Distribution System-Low Volume Terminals). It can be observed from Fig. 6 that the proposed waveform is 3.4 dB loss in E_b/N_0 at $BER = 10^{-5}$ compared with JTIDS-P4SP, but outperform the Tsinghua's waveform by 1.6 dB in E_b/N_0, and the MIDS-LVT by 3.1 dB in E_b/N_0.

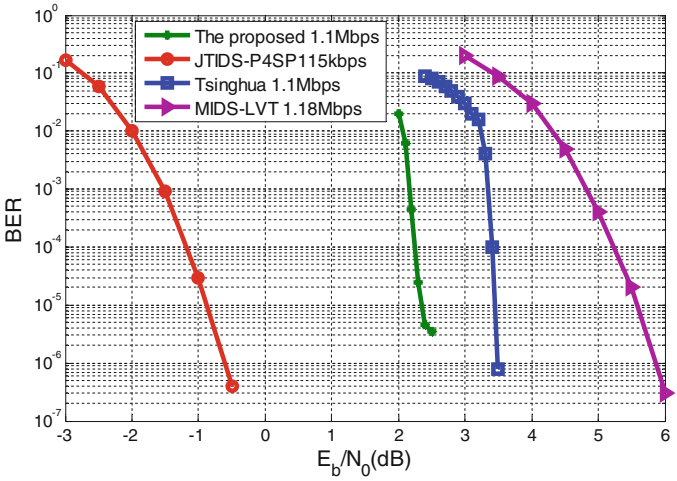


Fig. 6. Performance of the proposed waveform scheme and traditional one

As an important way of spread spectrum, FH communication is always applied in anti-interference system and burst channel [8]. According to the property of FH communication, let ε be the proportion of erased symbols, in the paper, we suppose, as soon as burst errors occur, 432 * ε hops in random positions are erased.

Figures 7, 8 and 9 illustrates the performance of rate 1.1 Mbps, 550 kbps, 270 kbps waveform proposed in this paper over burst-error channels when the proportions of erased symbols are 0, 10%, 20%, 50%, 66.7%.

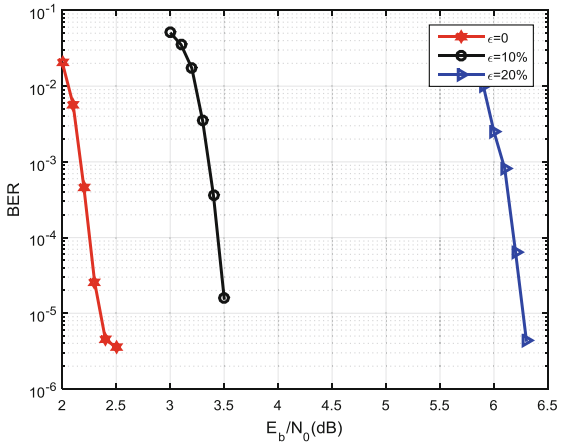


Fig. 7. Performance of the proposed 1.1 Mbps with different erased symbol rate

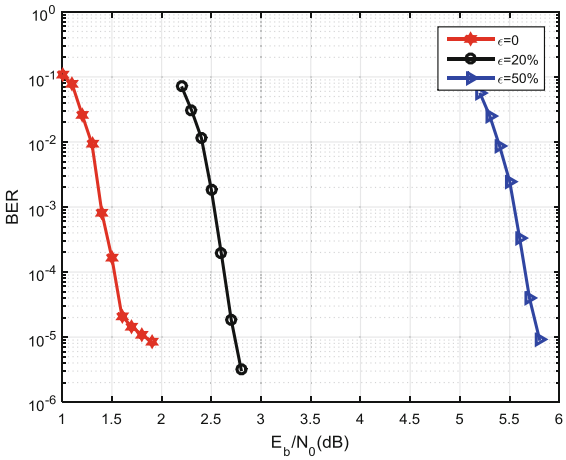


Fig. 8. Performance of the proposed 550 kbps with different erased symbol rate

Table 2 illustrates the performance of the rate 1.1 Mbps, 550 kbps, 270 kbps waveform over AWGN channel and burst-error channels. Analyzing from the simulation results, the lower of the rate result in better performance over burst-error channels.

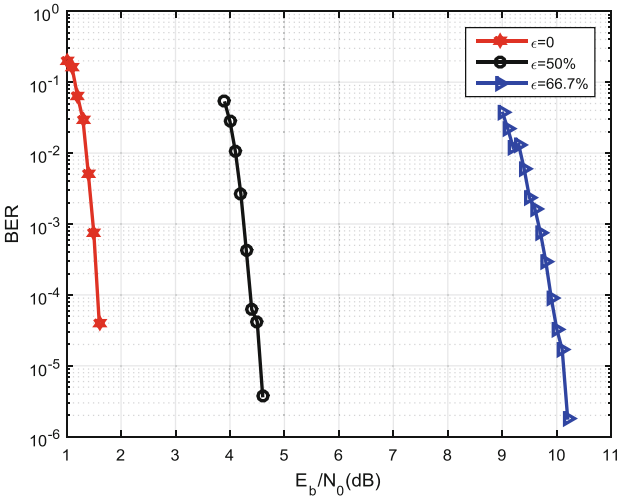


Fig. 9. Performance of the proposed 270 kbps with different erased symbol rate

Table 2. Performances of the propose waveform over AWGN and burst-error channels

Channel	Rate	Erasure proportion	E_b/N_0
Gaussian	1.1Mbps	0	2.35 dB
Gaussian	550 kbps	0	1.8 dB
Gaussian	270 kbps	0	1.6 dB
Erasure	1.1 Mbps	10%	3.5 dB
Erasure	1.1 Mbps	20%	6.25 dB
Erasure	550 kbps	20%	2.7 dB
Erasure	550 kbps	50%	5.75 dB
Erasure	270 kbps	50%	4.5 dB
Erasure	270 kbps	66.7%	10.1 dB

5 Conclusion

A throughput enhanced JTIDS waveform is proposed in this paper, which supports the rate of 1.1 Mbps, 550 kbps, and 270 kbps. As for power spectral, the side lobe of power spectrum attenuates fast for the GMSK. For modulation, nonrecursive GMSK significantly weaken the error correlation in the demodulator. As for demodulation, noncoherent demodulation algorithm of N-SISO improves the performance. When constructing each data frame, the special frame structure of adding one Head DD and one End DD makes a considerable improvement in correcting channel errors. Simulation results show that throughput enhanced waveform offers a better BER performance under AWGN channel and burst-error channel. As future work, we consider a scheme for interference recognizing and suppressing to solve the performance deterioration problem of proposed throughput enhanced JTIDS in the presence of partial-band barrage jamming.

References

1. Kopp, C.: Network Centric Warfare Fundamentals—Part 3: JTIDS/MIDS. Defense TODAY Magazine. 12–19 (2005)
2. Gonzales, D.M., Norton, D., Hura, M.: Multifunctional Information Distribution System (MIDS) Program Case Study, Project AIR FORCE, RAND, DB-292-AF (2000)
3. Asenstorfer, J., Cox, T., Wilksch, D.: Tactical Data Link Systems and the Australian Defense Force (ADF)-Technology Developments and Interoperability Issues, Information Networks Division, Information Science Laboratory, DSTOTR-1470 (2004)
4. Zhao, W., Yu, Z., Jian, S.: Rate-enhanced JTIDS waveform based on LDPC. Tsinghua Univ. (Sci. Technol.) **55**, 520–525 (2015)
5. Noh, H., Lim, J.: Partitioned cyclic code shift keying for JTIDS. In: Proceedings of IEEE MILCOM (2012)
6. Noh, H., Yu, J.: Code selection approach for partitioned cycliccode shift keying to improve multinet capability. In: Waveforms and Signal Processing, IEEE MILCOM (2015)
7. Chi-Han, K., Robertson, C.: An improved link-16/JTIDS receiver in pulsed-noise interference. In: Waveforms and Signal Processing, IEEE MILCOM (2011)
8. Chunhui, S., Bing, L., Baoming, B.: Design of LDPC coded CPM over burst-error channels. In: IEEE 83rd Vehicular Technology Conference (2016)
9. Sun, J.H., Li, M.L., Wu, X.J.: The design of serial concatenated system of Turbo product code with continuous phase modulation. J. Xi'an Jiaotong Univ. **46**, 88–93 (2012)
10. Bing, L., Baoming, B., Mengyu, H.: A robust noncoherent iterative detection algorithm for serially concatenated CPM. In: 6th International Symposium on Turbo Codes & Iterative Information Processing (2010)

Design and Application of Aperture-Coupled-Patch Frequency Selective Surfaces

Yu Wang(✉)

China Electronics Technology Group Corporation (CETC),
20th Institute Key Laboratory of Technology on Datalink,
Xi'an 710068, Shaanxi, China
nicolas66wang@163.com

Abstract. In this paper, the theories, features, and design methods to the aperture-coupled-patch frequency selective surfaces (FSSs) were discussed. A novel wide bandwidth FSSs and a kind of multi-bands FSSs with reduced geometry were designed in this paper due to the theory we introduced. Based on the FSSs proposed in this paper, a 24-GHz flat lens antenna was designed as an application example to certify the theory. A prototype of lens antenna was fabricated and measured, and measurement results well agree with the calculated ones.

Keywords: Frequency Selective Surfaces (FSSs) · Minkowski fractal shape
Wide band · Dual band · Flat lens

1 Introduction

Frequency selective surface (FSS) is a kind of periodic structure with one or two dimensions. It was firstly discovered in optical field by Rittenhouse in 1786 [1]. At the beginning of the discovery, FSSs were just considered as a filtering phenomenon in the optical area. With the development of the microwave technologies, FSSs were introduced into microwave field gradually. The researches on FSSs were starting from 1960's, and after several decades' development the technologies for FSSs get now more and more ripe.

The carefully designed single layered FSS can satisfy a lot of frequency characteristics. Moreover, with the development of microwaves, a lot of new technologies were introduced into the FSS design, such as fractal shape for multiple frequency bands, active FSS for frequency characteristic controlling, etc. However, there are still some limits in designs if using just single layer, so that multiple layered FSSs are required.

Aperture-coupled-patch FSS is also a kind of multiple layered FSS [2]. The basic structure and its Equivalent circuit were shown in Fig. 1. As illustrated in Fig. 1(a), we assume the incident wave propagates along the z axis, and the top layer of patch can be regarded as reception. The incident wave will excite inductive currents on the top layer, and the inductive current will generate scattering field then. Such scattering field will be

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 13–20, 2018.
https://doi.org/10.1007/978-3-319-78130-3_2

coupled by the inside slot and radiated by the bottom patch. Figure 1(b) illustrates the equivalent circuit. Z_0 stands for the impedance of free-space, which is 377Ω. Z_1 means the impedance of substrate, and L is the thickness of substrate. L_1 and C_1 are the equivalent inductance and capacitance of the slot layer respectively. L_2 is the parasitic inductance between adjacent slot elements. C_2 is the parasitic capacitance between adjacent patch elements.

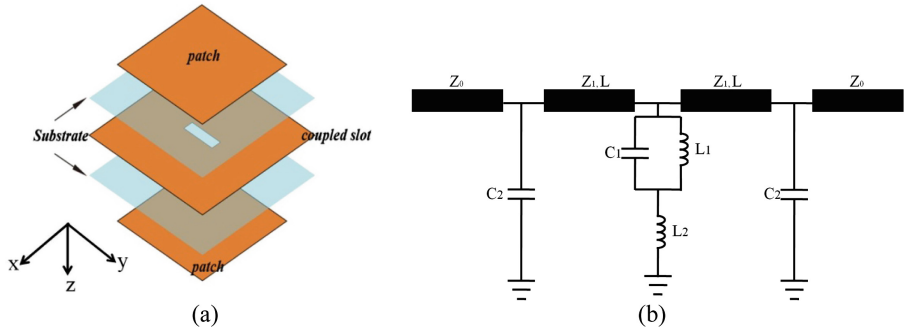


Fig. 1. Aperture-coupled-patch FSSs: (a) Structure, (b) Equivalent circuit.

2 A Four-Pole Pass-Band FSSs Design

A kind of aperture-coupled-patch FSS with four-pole pass band is proposed in this section. The classic resonant method is used here to analyze and design the FSSs frequency response.

The proposed novel FSSs element shown in Fig. 2 consists of just three layers of screens and two layers of 1-mm substrate with dielectric parameter εr = 2.8 and tanδ = 0.002. The frequency response of the proposed FSS is shown in Fig. 3. It is obvious in Fig. 3 that this structure can generate four resonant points in pass band. The first and second resonant points are generated by the patch layers respectively, while the third resonant point is generated by the coupled slot layer. As for the fourth

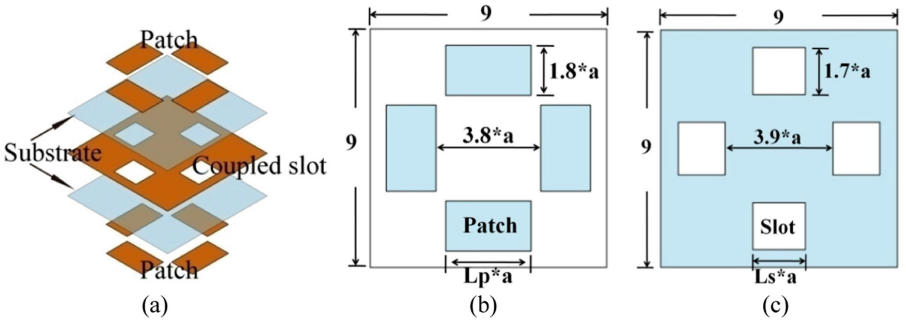


Fig. 2. Geometry of proposed element (unit: mm): (a) 3D View, (b) Patch Layer, (c) Slot Layer.

resonant point, it is a similar phenomenon to the Fabry-Perot (FP) resonance, which is controlled by the inter-screen space. The simulation results show that a perfect roof-top-like pass band is obtained as well as a wide range of transmission phase shift. A uniform array arranged by the proposed element is fabricated and the characteristic of S21 efficiency is measured in free-space. The measurement result well agreed with the calculated one. Compared with the other aperture-coupled patch FSS designs, the characteristic of four-pole pass-band is seldom reported in the previous works. However, such characteristic in this design will scarify the profile a little.

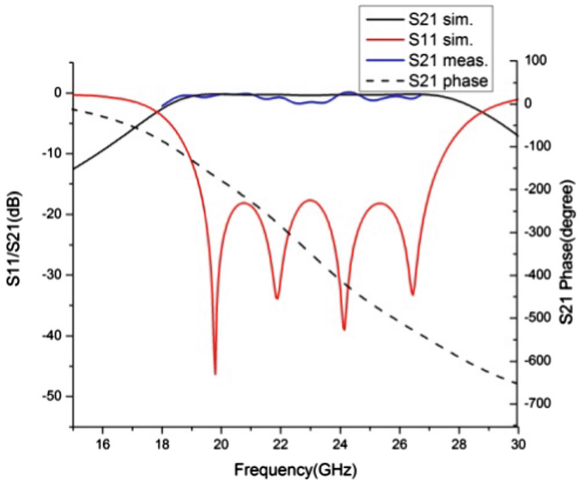


Fig. 3. Frequency response of the proposed FSS (a = 1.08, Lp = 3.1 mm, Ls = 1.9 mm).

3 Dual Band and Geometry Reduction Design

Minkowski fractal shape is a common used antenna miniaturization method [3–5]. A 3rd-order Minkowski fractal narrow slot is used here, it is transformed from the common square slot (1st-order). There are three steps of transformation from 1st-order to 3rd-order, illustrated in Fig. 4. The width of the slot is 0.2 mm.

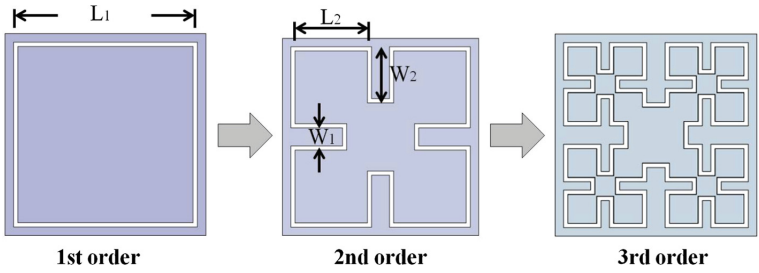


Fig. 4. Transformation of Minkowski fractal narrow slot.

In every transformation step, the original square loop iterates to a small scale. The initial geometry is decided by the following equations:

$$W_1^{n+1} = a_1 \times L_n \tag{1}$$

$$W_2^{n+1} = a_2 \times L_n \tag{2}$$

Here, *Ln* is the length of the square ring in *n*-th order, W_1^{n+1} is the middle segment width in $(n + 1)$-th order, and W_2^{n+1} is the indentation width, as shown in Fig. 4. a_1 and a_2 are the scale ratios.

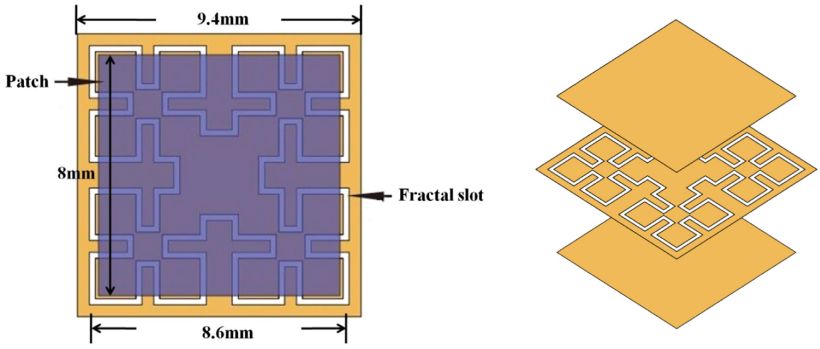


Fig. 5. Structure of proposed FSS element: (a) Geometry, (b) 3-D View.

Using the Minkowski fractal slot shown in Fig. 4 as the coupled slot, and using common square patch as the outside patch layers, the geometry of the proposed FSS element is shown in Fig. 5. The initial geometry parameters chosen here are L1 = 8.6 mm, $a_1 = 0.1$, and $a_2 = 0.15$. The three layers of screen of the proposed element are supported by two layers of substrate with thickness of 1.9 mm and electrical parameter $\varepsilon r = 3.2$, $\tan\delta = 0.003$.

The frequency response of this proposed structure is shown in Fig. 6. It can be seen that this structure can support double pass-bands, and three poles exist in the second operation band. A perfect roof-top liked transmission band as well as double sharp edges are obtained in the second operation band. Because the Minkowski fractal narrow slot used here increases the electrical path, the period of the FSS element is reduced to be just around $0.2\lambda_0$ (For most common structures it should be around $0.5\lambda_0$).

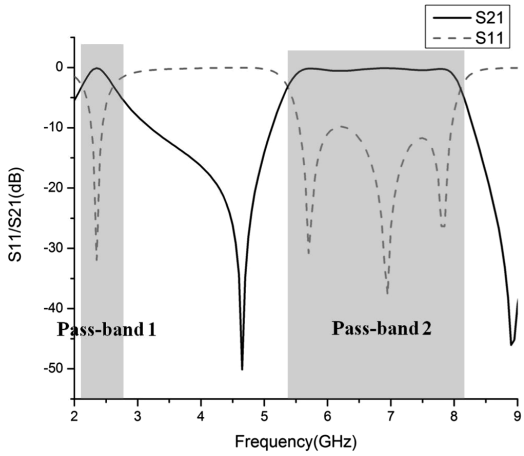


Fig. 6. Frequency response of proposed structure

4 Application to Flat Lens Antenna

In this section, a novel broadband lens antenna is presented. The original aperture-coupled patch FSS element proposed in Sect. 2 is chosen here to make up a phase shifter array.

The scale parameter a is chosen from 0.87 to 1.13 for S21 > −1dB and S11 < −10dB. The S21 phase ranges from −178.6° to −467.2° when a changes from 0.87 to 1.13, shown in Fig. 7. A 15 * 15 phase shifter array is designed, and the total size is 135 mm * 135 mm. Then we can use Eq. (3) to calculate the desired compensated phase [6]:

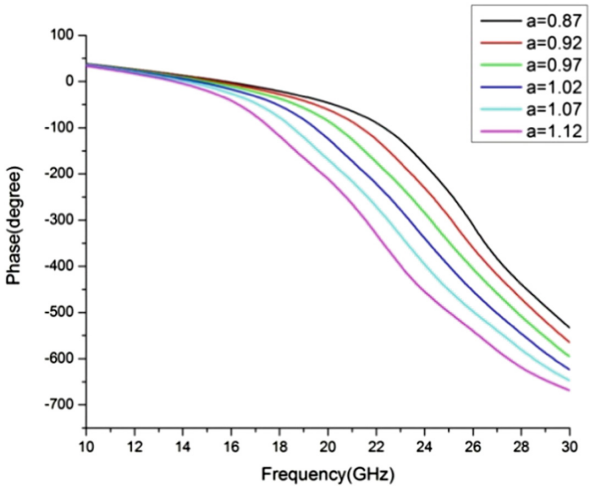


Fig. 7. Transmission phase shift against scale parameter.

$$\varphi_i = \frac{2\pi}{\lambda_0}(R_i - F) \pm 2\pi N + \varphi_0 \tag{3}$$

Here, φ_0 is the phase of centre cell on lens, and φ_i is the phase of i-th cell on lens. λ_0 is the wavelength at the designed frequency. N is just a integer number to make sure that: $0 < \varphi_i \leq 2\pi$. Usually for convenience, we assume φ_0 equals 0 because just relative phase differences between each cell are focused on.

After calculating the desired compensation phase, then we can obtain the phase distribution of the incident wave and transmission wave at different frequencies. Figure 8(a) and (b) show the phase of incident wave and transmission wave at the centre frequency 24 GHz, respectively. In Fig. 8, we can observe that the incident wave is spherical because obvious phase differences exit. After equipping the lens, homogeneous phase distribution is obtained.

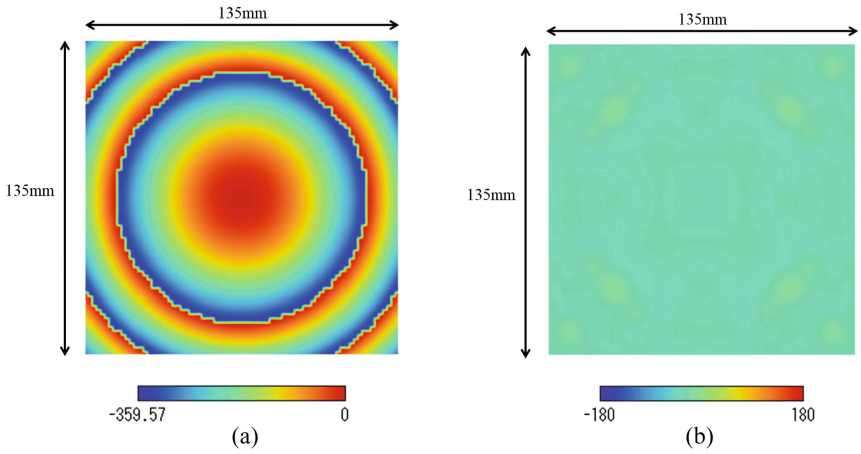


Fig. 8. Phase distribution at 24 GHz: (a) Incident wave, (b) Transmission wave.

The calculated results of phase distribution show that the incident wave will be compensated and plane wave will be obtained. Then we fabricate a calculation model and prototype to certify the theories further. The proposed lens antenna is calculated in HFSS and the prototype is measured in free-space. In the calculation, E-M symmetric boundaries are used to reduce the total size of the whole lens system so that the meshes and calculation time is reduced but without effect on the results.

Figure 9 shows the calculated and measured radiation patterns. It is obvious to see that the high gain and low side lobe level characteristics are obtained, and measurement results well agree with the calculated ones, which verify the introduced principles and theories. The peak gains at different frequencies were also plotted in Fig. 10, good results can be observed. It is obvious that the gain is enhanced after equipping lens compared with just using the standard horn. Figure 10 also shows that the element proposed in this paper can operate in a wide band.

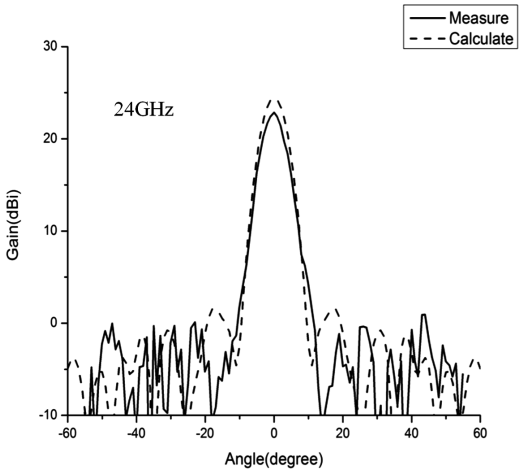


Fig. 9. Measured and calculated radiation patterns.

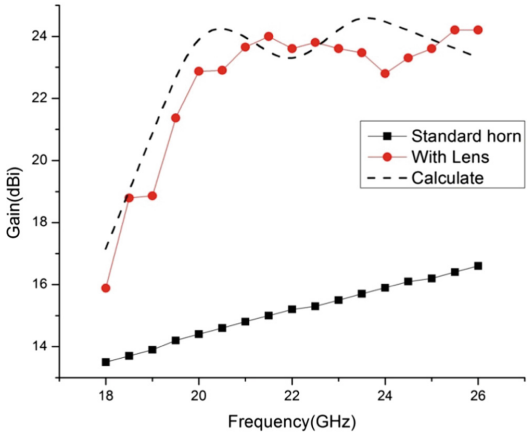


Fig. 10. Measured and calculated peak gain against frequency.

5 Conclusion

Two newly kinds of aperture-coupled-patch FSSs element are presented in this paper. Pass-band with four-pole resonant behavior is obtained in the first design, while geometry is miniaturized in the second design by using fractal narrow slot as the coupled slot. Then the element proposed in Sect. 2 is chosen to make up a phase shifter array for flat lens use as a design example to certify the introduced theory. The prototype was measured in free-space, and there were good agreements between measurement and calculation results. The results show that antenna gain is enhanced after equipping the lens, and side lobe level is depressed. The results also indicate that the proposed antenna can operate in a very wide band.

References

1. Rittenhouse, D.: An optical problem, proposed by Mr. Hopkinson, and solved by Mr. Rittenhouse. Trans. Am. Philos. Soc. **2**, 201–206 (1786)
2. Shibayama, R., Deguchi, H., Tsuji, M.: Flat thin polarizer-lens based on multiple resonance behavior. In: IEEE Antennas and Propagation Society Symposium (APSURSI), pp. 1–4 (2010)
3. Lee, E., Hashim, N., Vandenbosch, G., Volski, V., Mirza, H., Aziz, M.: Design and fabrication of a flexible Minkowski fractal antenna for VHF applications. In: Proceedings of the 5th European Conference on Antennas and Propagation, pp. 521–524 (2011)
4. Luo, Q., Salgado, H., Pereira, J.: Fractal monopole antenna design using Minkowski island geometry. In: Antennas and Propagation Society International Symposium, pp. 1–4 (2009)
5. Padhi, S., Swiegers, G., Bialkowski, M.: A miniaturized slot ring antenna for RFID applications. In: 15th International Conference on Microwaves, Radar and Wireless Communications, vol. 1, pp. 318–321 (2004)
6. Pozar, D.M.: Flat lens antenna concept using aperture coupled microstrip patches. Electron. Lett. **32**(23), 2109–2111 (1996)



Radar Cross Section Reduction Based on Metasurface

Xing-Long Zhang(✉), Meng Niu, Luo-Hui Su, and Ke-Pu Song

AVIC Xi'an Flight Automatic Control Research Institute, Xi'an 710065, China
zxilong@163.com

Abstract. A novel polarization insensitive metasurface for backscattering Radar Cross Section (RCS) reduction is designed and studied. The proposed metasurface consist of carefully arranged unit cells, which enables estimated uniform diffusion of incoming electromagnetic (EM) energy. A theoretical model based on reflect array theory is introduced to demonstrate the operating principle of the metasurface. An over 20 dB RCS reduction characteristics is observed in the vicinity of 9.5 GHz, indicating our metasurface will provide potential applications for future stealth technologies.

Keywords: Metasurface · Radar Cross Section (RCS) reduction
Stealth technology

1 Introduction

Recently, an increasing number of articles are reported for RCS reduction technology. Among them, the employment of metamaterial is now attracting a growing attention of research. Utilizing the loss components of the effective medium parameters, conventional metamaterials absorbers (MAs) transform the incoming electromagnetic energy into heat at the resonant frequency of the structure [1, 2]. Most MAs are single layer structures, which promise a dramatically reduced contour. Nevertheless, due to the strong resonant property, unit cells employed in metamaterial absorbers ultimately limit their operating bandwidth.

The newly reported electromagnetic cloaking devices, which mainly is based on optical transformation technology, provide an efficient way to bend the incident waves along the boundary of a region to be hidden without perturbing the exterior fields [3–5]. Therefore theoretical perfect invisibility can be realized without backward or forward RCS. Unfortunately, most cloakings are complex, multilayered and fragile structures, which make such devices quite difficult to fabricate and be hindered in practical applications.

Another promising method stems from the commonly used scattering cancellation approach, where artificial magnetic conductor (AMC) and perfect electric conductor (PEC) are arranged in chessboard conformation as super unit cells [6–8]. The reflected waves from the two parts experience a 180° phase difference at the resonance of AMC, hence the destructive interference between the two waves dramatically reduce the

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 21–26, 2018.
https://doi.org/10.1007/978-3-319-78130-3_3

scattering field level at mirror direction. The bandwidth of such coating is also limited since it is usually impractical to design ultra broadband AMC structures.

Recently, a series of structures named random meta-surfaces are reported to realize backscattering RCS reduction [9, 10]. Similar to the chessboard structures, the randomly distributed reflection phase promotes a diffused scattering wave. Comparatively, most currently reported random meta-surfaces perform an isotropic RCS reduction characteristics within the upper half space. However, for most recent searchers, they do not focus on the scattering effect of their metamaterials.

Different from other searchers, in this paper, we develop a polarization independent metasurface with reduced RCS in the vicinity of 9.5 GHz. The purposed RCS reduction comes from the diffusion instead of pure absorption of incoming wave The reflection phase of each unit tend to be random. The reflected waves from each part of the metasurface will experience destructive inference, resulting in diffused scattering waves. The operating principle of the design is demonstrated in detailed by reviewing the traditional reflect array theory. A full wave simulation utilizing HFSS is conducted to achieve the metasurface performance. In fact, our metasurface will be applied in future combat platforms and greatly improve their safety under the threaten of enemy air defend systems.

2 Unit Cell Design

The unit cell with a period of P = 10 mm is composed of a main circle ring together with a parasitic cross printed patch on a grounded FR4 dielectric layer with a relative permittivity of 4.4 and a thickness of 3 mm (shown in Fig. 1).

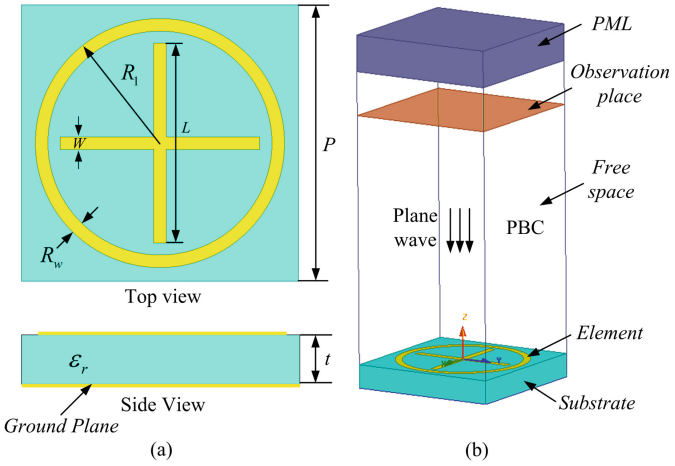


Fig. 1. (a) The unit cell of proposed metasurface (b) The simulation conformation assignment.

Other geometry parameters and their relationships are set as follows: R is the outer radius of the ring. $R_w = 0.1 \times R$ is the width of ring. $L = 1.6 \times R$ is the length of cross and $W = 0.05 \times R$ is the width of cross. The unit cell is symmetric with respect to x-axial and y-axial, leading to a polarization independent response respect to normal incidence. When a plane wave normally impinges on the metasurface, the phase of the reflected wave varies when the size of the patches is modified.

A fairly accurate model using the HFSS simulation was performed to analyze the reflection phase characteristics of the element, as shown in Fig. 1b. In this model, periodical boundary conditions (PBC) are assigned along the lateral directions to imitate the behavior of the infinite periodic array structure. An observation plane was set in the far-field zone of the unit cell surface to calculate the reflection phase of the incident plane wave. A perfectly matched layer (PML) was positioned to absorb the reflected waves. The scattered E-field on the observational plane was recorded, and the reflection phase on this plane was calculated using

$$\eta = (\int_A \text{Phase}(E_{scattered})ds)/(\int_A \hat{s} \bullet d\vec{s}) + 2\pi h/\lambda \tag{1}$$

where A is the evaluated observation plane and $\hat{s}$ is its normal unit vector. The E-field data can be directly available by utilizing the "plot field" function in HFSS. h is the distance between the observation plane and the surface of patches. λ is the free space wavelength at the operating frequency.

Figure 2 depicts the reflection phase and amplitude of the unit cell versus the outer radius of the ring-shaped patch when the plane wave is illuminated normal to the surface. A 355° phase shift at 9.5 GHz is achieved. The amplitude curve remains unchanged except for a slight dip less than 3 dB at R = 3.1 mm, where a weak resonance emerges and part incoming energy is absorbed [1–4].

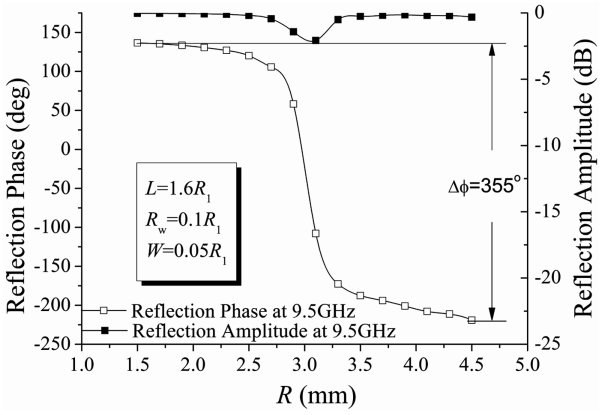


Fig. 2. The reflection phase of the unit cell versus R

3 Metasurface Design

Let us consider a two dimensional metasurface in Fig. 3, which is composed of $M \times N$ units (for the proposed metasurface in this paper, $M = N = 25$). When illuminated by an incoming plane wave, the total scattering field from the metasurface can be regarded as the superposition of the scattering wave from each basic unit cell. According to the reflect array theory, the field scattered by the array can be approximately written as

$$E_s = \sum_{m=1}^{M} \sum_{n=1}^{N} A(m,n) \exp(j2\pi\phi/\lambda) \tag{2}$$

$$\phi = \sin(\theta) \times [x_{mn} \cos(\varphi) + y_{mn} \sin(\varphi)] \tag{3}$$

where $A(m, n)$ represents the scattering field of the m,n-th unit cell, which is mainly affected by its geometry parameters. An infinite periodic array model depict in Fig. 1b is adapted to obtain the scattering field of each unit cell. In this approximate model, every unit cell is assumed to be surrounded by the same cells when considering the influence of mutual coupling. By carefully assign the geometry parameters of each unit cell, an arbitrary scattering field pattern of the entire metasurface can be theoretically realized.

Here 25×25 random number which varies from 0 to 360° are assigned as the reflection phase of each unit cell. From the relationship between R and the reflection phase at the operating frequency of 9.5 GHz (Fig. 2), the geometry parameters of the each patch are achieved. Finally, a metasurface composed of 25×25 elements is shown in Fig. 3.

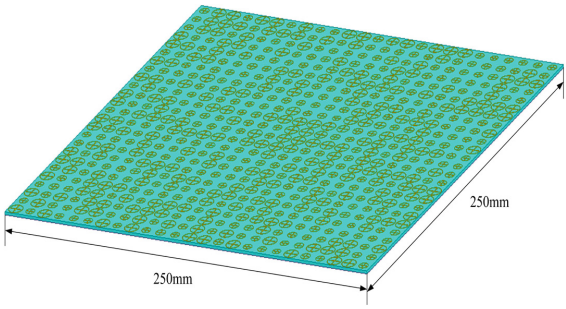


Fig. 3. The proposed metasurface with dimensional varied unit cells

4 Metasurface Performance

A full wave simulation by means of HFSS is carried out to investigate the backscattering and double static RCS behavior of the metasurface. For comparison, the scattering characteristics of an equal sized square perfect electric conductor (PEC) sheet

and a regular array with identical basic unit cells (shown in Fig. 1(b) and their R = 4.5 mm) are presented as well.

As presented in Fig. 4, a noticeable dip takes place in the vicinity of 9.5 GHz, corresponding to the mentioned RCS reduction. The slight frequency shift mainly comes from the mutual coupling of unit cells.

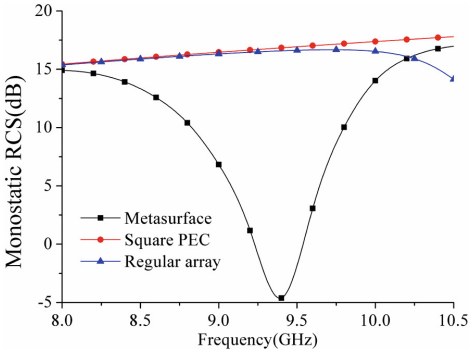


Fig. 4. The backscattering RCS of the metasurface

The 2-D double static RCS patterns at 9 GHz are illustrated in Figs. 5 and 6, respectively. Compared to the square PEC and regular array, the main beams of the metasurface double static RCS are dramatically suppressed. The side beams of RCS pattern are slightly raised, especially at end-fire directions. Obviously, the scattering energy is redistributed uniformly into the upper half space, resulting in thus a considerable backscatter RCS reduction. The scattering patterns predicted by the analytical model are computed utilizing MATLAB 2009 and are plotted as well. Both results seem to agree well, which validates the aforementioned analytical design method.

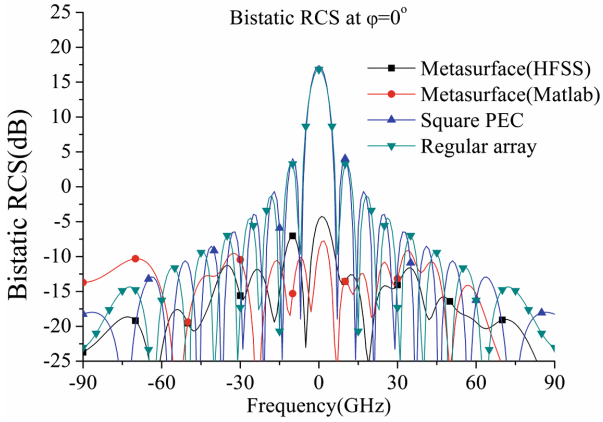


Fig. 5. The double static RCS of proposed metasurface at $\varphi = 0°$

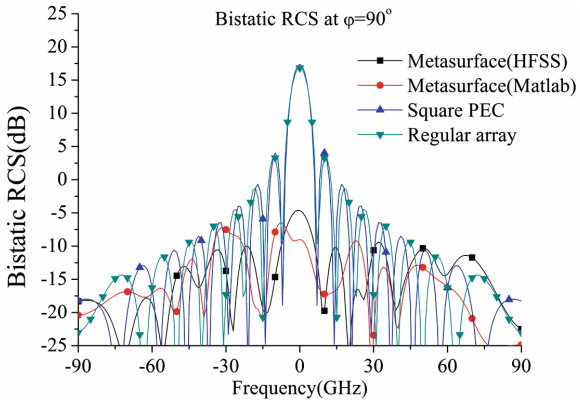


Fig. 6. The double static RCS of proposed metasurface at $\varphi = 90°$

5 Conclusion

In conclusion, a metasurface for backward RCS reduction is studied. The simulated results illustrate that the metasurface can effectively decrease the backward scattering in the vicinity of 9.5 GHz. Our low RCS metasurface will greatly improve the safety of combat platform under the threaten of enemy air defend systems.

References

1. Landy, N.I., Sajuyigbe, S., Mock, J.J., et al.: Perfect metamaterial absorber. Phys. Rev. Lett. **100**(20), 207402 (2008)
2. Ra'di, Y., Simovski, C.R., Tretyakov, S.A.: Thin perfect absorbers for electromagnetic waves: Theory, Design, and Realizations. Phys. Rev. Appl. **3**(3), 037001 (2015)
3. Schurig, D., et al.: Metamaterial electromagnetic cloak at microwave frequencies. Science **314**, 977–980 (2006)
4. Pendry, J.B., Schurig, D., Smith, D.R.: Controlling electromagnetic fields. Science **312**, 1780–1782 (2006)
5. Liu, R., et al.: Broadband ground-plane cloak. Science **323**, 366–369 (2009)
6. Paquay, M., Iriarte, J.C., Ederra, I., Gonzalo, R., de Maagt, P.: Thin AMC structure for radar cross-section reduction. IEEE Trans. Antennas Propag. **55**, 3630–3638 (2007)
7. Zhang, Y., Mittra, R., Wang, B.Z.: AMCs for ultra-thin and broadband RAMdesign. Electron. Let. **45**, 484–485 (2009)
8. de Cos, M.E., Alvarez, Y., Las-Heras, F.: A novel approach for RCS reduction using a combination of artificial magnetic conductors. Prog. Electromagn. Res. **107**, 147–159 (2010)
9. Yang, X.M., Zhou, X.Y., Cheng, Q., Ma, H.F., Cui, T.J.: Diffuse reflections by randomly gradient index metamaterials. Opt. Lett. **35**, 808–810 (2010)
10. Cheng Q., Tang W., Ma H., et al.: A metasurface for RCS reduction. In: 8th European Conference on Antennas and Propagation (EuCAP). IEEE, pp. 276–277 (2014)



Joint Direction-of-Departure and Two-Dimensional Direction-of-Arrival Estimation with Automatic Pairing in Massive MIMO Systems

Yuan Zhou(✉), Xiang Ji, and Man Dai

RCDCT, School of Information and Electronics,
Beijing Institute of Technology, Beijing, China
zhouyuanbit@163.com, {jixiang,dma}@bit.edu.cn

Abstract. In this paper, we focus on the problem of joint direction-of-departure (DOD) and two-dimensional (2D) direction-of-arrival (DOA) estimation in massive multiple-input multiple-output (MIMO) systems. A novel method for angle estimation is proposed with automatic pairing. The method transforms the element space into the beamspace by employing the property of conjugate centrosymmetric array, such that the angle estimation operates only on the real-valued computation. Then, we utilize the real-valued rotational invariance relationships in beamspace to estimate DOD and 2D DOA in a fresh fashion. Numerical results show that the proposed algorithm provides reduced computational complexity without the loss of estimation accuracy owing to real-valued processing, automatic pairing and computation dimension reduction in massive MIMO systems.

Keywords: Two-dimensional direction-of-arrival (DOA)
Direction-of-departure (DOD) · Uniform rectangle array
Massive multiple-input multiple-output

1 Introduction

The fifth generation (5G) wireless network will serve as a key enabler in meeting the ever-increasing demand for high-rate multimedia wireless services driven by smart phones and the mobile Internet [1]. As one of the flexible solutions in 5G, massive multiple-input multiple-output (MIMO) technique has been proposed to provide a high degree of spatial freedom to substantially enhance the system's throughput, energy efficiency and link reliability [2,3] by exploiting a large number of antennas at each base station (BS). Owing to the reciprocity between the uplink (UL) channel and the downlink (DL) channel for time-division-duplexing

This work was supported in part by Ericsson and by the National Natural Science Foundation of China under Grant No. 61371075.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 27–36, 2018.
https://doi.org/10.1007/978-3-319-78130-3_4

(TDD) based massive MIMO systems, direction-of-arrival (DOA) estimation for the UL is capable of providing the BS with the required DL channel knowledge to perform its DL transmit preprocessing.

Parameter estimation of the paths such as direction of departure (DOD) and direction of arrival (DOA) estimation is one of the most significant aspects in MIMO systems [4,5]. Numerous powerful methods have been developed for MIMO systems. To be more specific, the authors of [6] derived a reduced-dimension multiple signal classification (MUSIC) algorithm, which only required one dimension search and avoided the high computational cost within two-dimensional MUSIC (2D MUSIC) algorithm. In [7], an rotational invariance technique (ESPRIT) method by exploiting the invariance property of both the transmit array and the receive array was proposed, which avoided the spectrum searching process. In [8], the authors proposed a beamspace ESPRIT algorithm where the rotational invariance structure was reconstructed in the beamspace transformation, however an additional pairing process of DOD and DOA was needed. Recently, DOA estimation algorithms have also been conceived for massive MIMO systems. Note that the article [9] solved the automatic paring problem, the computation dimension was very high when it was directly introduced into the massive MIMO systems. For a very large MIMO array, a 2D MUSIC approach, which needed two dimensional spectrum searching process, was proposed in [10] with high computational complexity. All these aforementioned works considered the joint DOD and one-dimensional (1D) DOA estimation, however, the antenna arrays of the massive MIMO array are expected to be implemented in more than one dimension due to the constraint concerning the array aperture. Responding to this, the authors in [11] studied the joint DOD and 2D DOA estimation, nevertheless, the authors did not discuss how to pair the DOD and 2D DOA for each target.

In this paper, a novel method of joint DOD and 2D DOA estimation in the massive MIMO system is discussed, which not only just needs real-valued processing as well as avoids the pairing process, but also reduces the computation dimension.

Notation: Throughout the paper, lower- and upper-case boldface symbols denote vectors and matrices, respectively. $\mathbf{I}_K$ is the $K \times K$ identity matrix. $\otimes$ and $\odot$ stand for Kronecker product and Khatri-Rao product (column-wise Kronecker product), respectively. $(\cdot)^T$ and $(\cdot)^H$ denote transpose and conjugate transpose, respectively. j is the imaginary unit and $\text{diag}\{a_1, a_2, ..., a_M\}$ is the diagonal matrix with $a_1, a_2, ..., a_M$ being its diagonal elements. $\mathbb{E}$ denotes the expectation.

2 DOD and DOA Estimation for URA in Massive MIMO Systems

2.1 Data Model

Consider a massive MIMO system with a W-element uniform linear array (ULA) in the transmitter and a large uniform rectangular array (URA) consisting of

MN elements at the receiver. The URA is located in the X-Z plain. Assume that there are P noncoherent sources located at the same range, the output of the entire matched filters at the receiver can be written as [7]

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t). \tag{1}$$

The steering matrix $\mathbf{A} = \mathbf{A}_g \odot \mathbf{A}_u \odot \mathbf{A}_v = [\mathbf{a}_1, ..., \mathbf{a}_P] \in \mathbb{C}^{WMN \times P}$ and the pth column is the steering vector of the pth source, which can be expressed as $\mathbf{a}_p = \mathbf{a}_{g,p} \otimes \mathbf{a}_{u,p} \otimes \mathbf{a}_{v,p}$. Symbol $\mathbf{A}_g$ is the transmitting steering matrix of ULA with $\mathbf{a}_{g,p}$ being its pth column and $\mathbf{A}_u$, $\mathbf{A}_v$ are the receiving steering matrixes in two directions of URA with $\mathbf{a}_{u,p}, \mathbf{a}_{v,p}$ being their pth columns. Employing the center of the ULA as the phase reference, the array vector $\mathbf{a}_{g,p} \in \mathbb{C}^{W \times 1}$ is conjugate centrosymmetric. Considering the number of elements comprising the ULA W is odd, the steering vector of transmitted antennas is

$$\mathbf{a}_{g,p} = [e^{-j(\frac{W-1}{2})g_p}, ..., e^{-jg_p}, 1, e^{jg_p}, ..., e^{j(\frac{W-1}{2})g_p}], \tag{2}$$

where $g_p = 2\pi \frac{d}{\lambda} \cos\phi_p$ with λ equal to the wavelength, d equal to the interelement spacing and ϕ_p equal to the DOD [12]. Considering the array centroid is located at the origin of the URA, the steering vector of received URA antennas can be expressed the Kronerker product of two ULA responses as

$$\mathbf{a}_{rp} = \mathbf{a}_{u,p} \otimes \mathbf{a}_{v,p}, \tag{3}$$

where $\mathbf{a}_{u,p} \in \mathbb{C}^{M \times 1}$ and $\mathbf{a}_{v,p} \in \mathbb{C}^{N \times 1}$ are as follow

$$\begin{aligned} \mathbf{a}_{u,p} =& [e^{-j(\frac{M-1}{2})u_p}, ..., e^{-ju_p}, 1, e^{ju_p}, ..., e^{j(\frac{M-1}{2})u_p}], \\ \mathbf{a}_{v,p} =& [e^{-j(\frac{N-1}{2})v_p}, ..., e^{-jv_p}, 1, e^{jv_p}, ..., e^{j(\frac{N-1}{2})v_p}], \end{aligned} \tag{4}$$

with $u_p = 2\pi \frac{d}{\lambda} \cos\theta_p$, $v_p = 2\pi \frac{d}{\lambda} \sin\theta_p \cos\varphi_p$ and φ_p, θ_p are the corresponding azimuth DOA and elevation DOA, respectively [13]. The vector $\mathbf{s}(t) = [s_1(t), ..., s_P(t)]^T$ represents the transmit signals, of which $s_p(t) = \beta_p e^{j2\pi f_p t}$, β_p and f_p being the amplitude and Doppler frequency, respectively. The vector $\mathbf{n}(t)$ represents an $WMN \times 1$ complex Gaussian white noise vector with zeros mean and covariance matrix $\sigma^2 \mathbf{I}_{WMN}$.

2.2 Beamspace ESPRIT for Joint DOD and 2D DOA Estimation

Define $WMN \times L_W L_M L_N$ beamspace transformed matrix as $\mathbf{B} = \mathbf{B}_g \otimes \mathbf{B}_u \otimes \mathbf{B}_v$, where $\mathbf{B}_g \in \mathbb{C}^{W \times L_W} (L_W < W)$ is the transmitting beamspace transformed matrix and $\mathbf{B}_u \in \mathbb{C}^{M \times L_M} (L_M < M)$, $\mathbf{B}_v \in \mathbb{C}^{N \times L_N} (L_N < N)$ are receiving beamspace transformed matrixes in two directions. As $\mathbf{B}_g$ for example, the lth row of $\mathbf{B}_g^H$ is formulated as

$$\begin{aligned} \mathbf{b}_{g,l}^H =& e^{j(\frac{W-1}{2})(l-1)\frac{2\pi}{W}} \cdot \\ & [1, e^{-j(l-1)\frac{2\pi}{W}}, \ldots, e^{-j(W-1)(l-1)\frac{2\pi}{W})}], 1 \le l \le L_W. \end{aligned} \tag{5}$$

Applying the conjugate centrosymmetrized version of each row, the complex-valued transmitting steering vector $\mathbf{a}_{g,p}$ is transformed to the real-valued beamspace steering vector as $\mathbf{d}_{g,p} = \mathbf{B}_g^H \mathbf{a}_{g,p}$. Then the transformation from WMN dimensional element space to an $L_W L_M L_N$ dimensional beamspace can be described as [8]

$$\mathbf{D} = \mathbf{B}^H \mathbf{A} = \mathbf{B}_g^H \mathbf{A}_g \odot \mathbf{B}_u^H \mathbf{A}_u \odot \mathbf{B}_v^H \mathbf{A}_v. \tag{6}$$

That means the transmitting beamspace steering matrix $\mathbf{D}_g = \mathbf{B}_g^H \mathbf{A}_g = [\mathbf{d}_{g,1}, \ldots, \mathbf{d}_{g,P}]$, the receiving beamspace steering matrices in two directions are $\mathbf{D}_u = \mathbf{B}_u^H \mathbf{A}_u = [\mathbf{d}_{u,1}, \ldots, \mathbf{d}_{u,P}]$ and $\mathbf{D}_v = \mathbf{B}_v^H \mathbf{A}_v = [\mathbf{d}_{v,1}, \ldots, \mathbf{d}_{v,P}]$, respectively. Then the real-valued beamspace steering matrix

$$\mathbf{D} = \mathbf{D}_g \odot \mathbf{D}_u \odot \mathbf{D}_v = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_P]. \tag{7}$$

Then we can have

$$\mathbf{d}_p = \mathbf{d}_{g,p} \otimes \mathbf{d}_{u,p} \otimes \mathbf{d}_{v,p}. \tag{8}$$

The transmitting beamspace steering matrix $\mathbf{D}_g$ is discussed first. Applying the conjugate centrosymmetrized version of each row, the pth component of $\mathbf{D}_g$ is $L_W \times 1$ real-valued beamspace steering vector can be expressed as

$$\mathbf{d}_{g,p} = \mathbf{B}_g^H \mathbf{a}_{g,p} = [d_{g,p}(1), \ldots, d_{g,p}(L_W)]^T, \tag{9}$$

where the lth and $l+1$th element are

$$\begin{aligned} d_{g,p}(l) &= \frac{\sin[\frac{W}{2}(g_p - (l-1)\frac{2\pi}{W})]}{\sin[\frac{1}{2}(g_p - (l-1)\frac{2\pi}{W})]}, \\ d_{g,p}(l+1) &= \frac{\sin[\frac{W}{2}(g_p - l\frac{2\pi}{W})]}{\sin[\frac{1}{2}(g_p - l\frac{2\pi}{W})]}. \end{aligned} \tag{10}$$

Thus, the two adjacent components of $\mathbf{d}_{g,p}$ are related as

$$\begin{aligned} &\sin\left[\frac{1}{2}\left(g_p - (l-1)\frac{2\pi}{W}\right)\right] d_{g,p}(l) + \\ &\sin\left[\frac{1}{2}\left(g_p - l\frac{2\pi}{W}\right)\right] d_{g,p}(l+1) = 0. \end{aligned} \tag{11}$$

Trigonometric manipulations lead to

$$\begin{aligned} \tan\left(\frac{g_p}{2}\right)\left\{\cos\left((l-1)\frac{\pi}{W}\right) d_{g,p}(l) + \cos\left(l\frac{\pi}{W}\right) d_{g,p}(l+1)\right\} = \\ \sin\left((l-1)\frac{\pi}{W}\right) d_{g,p}(l) + \sin\left(l\frac{\pi}{W}\right) d_{g,p}(l+1). \end{aligned} \tag{12}$$

All $L_W - 1$ equations lead to a real valued invariance relationship for the $\mathbf{d}_{g,p}$ as follows:

$$\tan\left(\frac{g_p}{2}\right) \mathbf{K}_{g,1} \mathbf{d}_{g,p} = \mathbf{K}_{g,2} \mathbf{d}_{g,p}, \tag{13}$$

$$\mathbf{K}_{g,1} = \begin{bmatrix} 1 & \cos(\frac{2\pi}{W}) & 0 & \dots & 0 & 0 \\ 0 & \cos(\frac{\pi}{W}) & \cos(\frac{2\pi}{W}) & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \cos((L_W-1)\frac{\pi}{W}) & \cos(L_W - \frac{\pi}{W}) \end{bmatrix} \in \mathbb{C}^{(L_W-1)\times L_W} \tag{14}$$

$$\mathbf{K}_{g,2} = \begin{bmatrix} 0 & \sin(\frac{2\pi}{W}) & 0 & \dots & 0 & 0 \\ 0 & \sin(\frac{\pi}{W}) & \sin(\frac{2\pi}{W}) & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \sin((L_W-1)\frac{\pi}{W}) & \sin(L_W - \frac{\pi}{W}) \end{bmatrix} \in \mathbb{C}^{(L_W-1)\times L_W} \tag{15}$$

where $\mathbf{K}_{g,1}$ and $\mathbf{K}_{g,2}$ are $(L_W - 1) \times L_W$ matrices as (14) and (15), respectively.

Based on the similar rotational invariance relationships of subarrays of URA in two directions, the expressions can be obtained as

$$\begin{aligned} \tan\left(\frac{u_p}{2}\right)\mathbf{K}_{u,1}\mathbf{d}_{u,p} &= \mathbf{K}_{u,2}\mathbf{d}_{u,p}, \\ \tan\left(\frac{v_p}{2}\right)\mathbf{K}_{v,1}\mathbf{d}_{v,p} &= \mathbf{K}_{v,2}\mathbf{d}_{v,p}, \end{aligned} \tag{16}$$

where $\mathbf{K}_{u,1}, \mathbf{K}_{u,2}$ and $\mathbf{K}_{v,1}, \mathbf{K}_{v,2}$ are defined similarly to $\mathbf{K}_{g,1}, \mathbf{K}_{g,2}$ with W replaced by M and N. Then we can have the following expressions

$$\begin{aligned} \tan\left(\frac{g_p}{2}\right)\widetilde{\mathbf{K}}_{g,1}\mathbf{d}_p &= \widetilde{\mathbf{K}}_{g,2}\mathbf{d}_p, \\ \tan\left(\frac{u_p}{2}\right)\widetilde{\mathbf{K}}_{u,1}\mathbf{d}_p &= \widetilde{\mathbf{K}}_{u,2}\mathbf{d}_p, \\ \tan\left(\frac{v_p}{2}\right)\widetilde{\mathbf{K}}_{v,1}\mathbf{d}_p &= \widetilde{\mathbf{K}}_{v,2}\mathbf{d}_p, \end{aligned} \tag{17}$$

where $\widetilde{\mathbf{K}}_{g,1}$, $\widetilde{\mathbf{K}}_{g,2}$ are the $(L_W-1)L_ML_N \times L_WL_ML_N$ matrices, $\widetilde{\mathbf{K}}_{u,1}$, $\widetilde{\mathbf{K}}_{u,2}$ are the $L_W(L_M-1)L_N \times L_WL_ML_N$ matrices and $\widetilde{\mathbf{K}}_{v,1}$, $\widetilde{\mathbf{K}}_{v,2}$ are the $L_WL_M(L_N-1) \times L_WL_ML_N$ matrices as

$$\begin{aligned} \widetilde{\mathbf{K}}_{g,1} &= \mathbf{K}_{g,1} \otimes \mathbf{I}_{L_M} \otimes \mathbf{I}_{L_N}, \quad \widetilde{\mathbf{K}}_{g,2} = \mathbf{K}_{g,2} \otimes \mathbf{I}_{L_M} \otimes \mathbf{I}_{L_N}, \\ \widetilde{\mathbf{K}}_{u,1} &= \mathbf{I}_{L_W} \otimes \mathbf{K}_{u,1} \otimes \mathbf{I}_{L_N}, \quad \widetilde{\mathbf{K}}_{u,2} = \mathbf{I}_{L_W} \otimes \mathbf{K}_{u,2} \otimes \mathbf{I}_{L_N}, \\ \widetilde{\mathbf{K}}_{v,1} &= \mathbf{I}_{L_W} \otimes \mathbf{I}_{L_M} \otimes \mathbf{K}_{v,1}, \quad \widetilde{\mathbf{K}}_{v,2} = \mathbf{I}_{L_W} \otimes \mathbf{I}_{L_M} \otimes \mathbf{K}_{v,2}. \end{aligned} \tag{18}$$

For all P sources, we consider the $WMN \times P$ DOA matrix $\mathbf{D} = [\mathbf{d}_1, \cdots \mathbf{d}_P]$, (17) demonstrate that $\mathbf{D}$ satisfies

$$\begin{aligned} \widetilde{\mathbf{K}}_{g,1}\mathbf{D}\boldsymbol{\Omega}_g &= \widetilde{\mathbf{K}}_{g,2}\mathbf{D}, \\ \widetilde{\mathbf{K}}_{u,1}\mathbf{D}\boldsymbol{\Omega}_u &= \widetilde{\mathbf{K}}_{u,2}\mathbf{D}, \\ \widetilde{\mathbf{K}}_{v,1}\mathbf{D}\boldsymbol{\Omega}_v &= \widetilde{\mathbf{K}}_{v,2}\mathbf{D}, \end{aligned} \tag{19}$$

where $\mathbf{\Omega}_g = \text{diag}\left\{ \tan\left(\frac{g_1}{2}\right), \cdots, \tan\left(\frac{g_P}{2}\right) \right\}$, and $\mathbf{\Omega}_u, \mathbf{\Omega}_v$ have the similar expressions with g replaced by u and v.

Looking back on the beamspace received signal $\mathbf{y}(t)$ as

$$\mathbf{y}(t) = \mathbf{B}^H \mathbf{x}(t) = \mathbf{D}^H \mathbf{s}(t) + \mathbf{B}^H \mathbf{n}(t). \tag{20}$$

$\mathbf{Y}$ denotes the $L_W L_M L_N \times L$ beamspace data matrix containing L snapshots as columns. The appropriate signal eigenvectors for the algorithm may be computed as the P "largest" left singular vectors of the real-valued matrix $[\Re\{\mathbf{Y}\}\ \Im\{\mathbf{Y}\}]$ [12], where $\Re$ and $\Im$ are the real and imaginary parts. In the ideal noise-free situation, the signal subspace $\mathbf{E}_s$ is related to the real-valued steering matrix $\mathbf{D}$ as $\mathbf{E}_S = \mathbf{DT}$, where $\mathbf{T}$ is an unknown $P \times P$ full rank real-valued matrix. Substituting $\mathbf{D} = \mathbf{E_s T}^{-1}$ into (19)

$$\begin{aligned}
\widetilde{\mathbf{K}}_{g,1}\mathbf{E}_s\mathbf{\Psi}_g &= \widetilde{\mathbf{K}}_{g,2}\mathbf{E}_s \quad \text{with} \quad \mathbf{\Psi}_g = \mathbf{T}^{-1}\mathbf{\Omega}_g\mathbf{T}, \\
\widetilde{\mathbf{K}}_{u,1}\mathbf{E}_s\mathbf{\Psi}_u &= \widetilde{\mathbf{K}}_{u,2}\mathbf{E}_s \quad \text{with} \quad \mathbf{\Psi}_u = \mathbf{T}^{-1}\mathbf{\Omega}_u\mathbf{T}, \\
\widetilde{\mathbf{K}}_{v,1}\mathbf{E}_s\mathbf{\Psi}_v &= \widetilde{\mathbf{K}}_{v,2}\mathbf{E}_s \quad \text{with} \quad \mathbf{\Psi}_v = \mathbf{T}^{-1}\mathbf{\Omega}_v\mathbf{T}.
\end{aligned} \tag{21}$$

Equations (21) can be solved by the least squares (LS) or the total least squares (TLS) algorithm. Automatic pairing of u_p and v_p can be obtained by the fact that all of the quantities in (21) are real values. Thus, $\mathbf{\Psi}_u + j\mathbf{\Psi}_v$ may be spectrally decomposed as

$$\mathbf{\Psi}_u + j\mathbf{\Psi}_v = \mathbf{T}^{-1}\{\mathbf{\Omega}_u + j\mathbf{\Omega}_v\}\mathbf{T}. \tag{22}$$

Accordingly, u_p and v_p $(p = 1, \cdots P)$ can be estimated based on the eigenvalues η_p of (22) from $\hat{u}_p = 2\tan^{-1}\left(\Re\{\eta_p\}\right)$ and $\hat{v}_p = 2\tan^{-1}\left(\Im\{\eta_p\}\right)$. From the definition of the spatial frequency, we can get the final result

$$\begin{aligned}
\hat{\theta}_p &= \arccos(\hat{u}_p \frac{\lambda}{2\pi d}), \\
\hat{\varphi}_p &= \arccos(\frac{\hat{v}_p}{\sin\hat{\theta}_p}\frac{\lambda}{2\pi d}).
\end{aligned} \tag{23}$$

Similarly, automatic paring of u_p and g_p spatial frequency estimates can be used by the fact that all of the quantities in (21) are real values. Thus, $\mathbf{\Psi}_u + j\mathbf{\Psi}_g$ may be spectrally decomposed as

$$\mathbf{\Psi}_u + j\mathbf{\Psi}_g = \mathbf{T}^{-1}\{\mathbf{\Omega}_u + j\mathbf{\Omega}_g\}\mathbf{T}. \tag{24}$$

Accordingly, u_p and g_p $(p = 1, \cdots P)$ can be estimated based on the eigenvalues μ_p of (24) from $\hat{u}_p = 2\tan^{-1}\left(\Re\{\mu_p\}\right)$ and $\hat{g}_p = 2\tan^{-1}\left(\Im\{\mu_p\}\right)$. From the definition of the spatial frequency, we can get the final result

$$\begin{aligned}
\hat{\theta}_p &= \arccos(\hat{u}_p \frac{\lambda}{2\pi d}), \\
\hat{\phi}_p &= \arccos(\hat{g}_p \frac{\lambda}{2\pi d}).
\end{aligned} \tag{25}$$

Then automatic paring of $\hat{\phi}_p$, $\hat{\theta}_p$ and $\hat{\varphi}_p$ has been obtained. The proposed estimation algorithm is summarized as Algorithm 1.

Algorithm 1. Estimation of the DOD and 2D DOA without pairing

1: Calculate the beamspace received signal $\mathbf{Y} = \mathbf{B}^H\mathbf{X}$.
2: Compute the signal subspace $\mathbf{E}_s$ via P largest left singular vectors of $[\Re\{\mathbf{Y}\}, \Im\{\mathbf{Y}\}]$.
3: Compute $\boldsymbol{\Psi}_g$, $\boldsymbol{\Psi}_u$ and $\boldsymbol{\Psi}_v$ as the solution to the equations (21).
4: Compute the eigenvalues η_p for $p = 1, \cdots, P$ of the $P \times P$ matrix $\boldsymbol{\Psi}_u + \mathrm{j}\boldsymbol{\Psi}_v$.
5: Compute spatial frequency estimates $\hat{u}_p = 2\tan^{-1}\left(\Re\{\eta_p\}\right), \hat{v}_p = 2\tan^{-1}\left(\Im\{\eta_p\}\right)$.
6: Estimate $\hat{\varphi}_p$ and $\hat{\theta}_p$ using (23).
7: Compute the eigenvalues μ_p for $p = 1, \cdots, P$ of the $P \times P$ matrix $\boldsymbol{\Psi}_u + \mathrm{j}\boldsymbol{\Psi}_g$.
8: Compute spatial frequency estimates $\hat{u}_p = 2\tan^{-1}\left(\Re\{\mu_p\}\right)$ and $\hat{g}_p = 2\tan^{-1}\left(\Im\{\mu_p\}\right)$.
9: Estimate $\hat{\phi}_p$ and $\hat{\theta}_p$ using (25).

2.3 Complexity Analysis

For the element ESPRIT [7] and the unitary ESPRIT algorithm [9], the computational complexity of eigendecomposition is $O(M^3N^3W^3)$, which is very heavy. Unlike the element ESPRIT and the unitary ESPRIT algorithm, the proposed beamspace unitary ESPRIT algorithm transform the original data vector into lower-dimensional beamspace. If only $L_W < W$ transmitting beams and $L_M L_N < MN$ receiving beams are selected, the computational complexity could be reduced to $O(L_M^3 L_N^3 L_W^3)$. It is important to note that after the beamspace transformation, the matrices become real matrices. This will significantly reduce the computational complexity. Moreover, DOD and 2D DOA can be paired automatically, which need to be paired additionally in [7,8].

3 Simulation Results

In this section, two sets of simulations compared with the method in [7,9] are provided. Assume that there exist $P = 2$ uncorrelated stationary targets, which are located at angles $(\phi_1 = 12°, \varphi_1 = 30°, \theta_1 = 15°)$, $(\phi_2 = 25°, \varphi_2 = 40°, \theta_2 = 30°)$ and the number of snapshots is $L = 500$. We investigate the performance of the proposed estimator through 200 Monte-Carlo simulation experiments. The root mean squared error (RMSE) of overall angle estimation is defined as

$$\mathrm{RMSE} = \frac{1}{P}\sum_{p=1}^{P}\sqrt{\mathbb{E}[(\hat{\phi}_p - \phi_p)^2 + (\hat{\varphi}_p - \varphi_p)^2 + (\hat{\theta}_p - \theta_p)^2]}, \tag{26}$$

where $\hat{\phi}_p$, $\hat{\varphi}_p$ and $\hat{\theta}_p$ are the estimated DOD, estimated azimuth DOA and estimated elevation DOA for the pth path.

Figure 1 shows the RMSEs with SNR, where a receive URA with $M = N = 7$, $M = N = 9$ and a transmit ULA with $W = 5$ are used. Let the beamspace $L_M = M - 2$, $L_N = N - 2$ and $L_W = W - 2$. It is indicated in Fig. 1 that the RMSEs of the approaches decrease when the SNR increases and the proposed algorithm has approximately the same performance compared with [7,9].

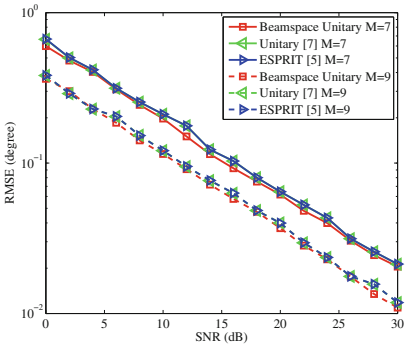


Fig. 1. RMSEs for overall angle estimation versus SNR, M = N = 15

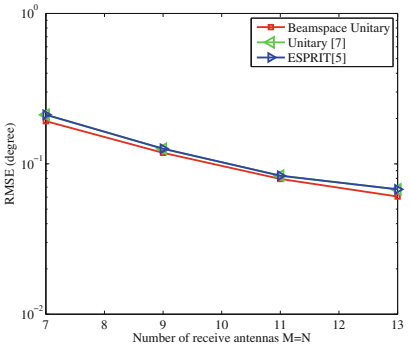


Fig. 2. RMSEs for overall angle estimation versus the number of antennas

Figure 2 shows the RMSEs versus the number of antenna $M = N$ with $W = 5$ and SNR = 10 dB. It can be seen that the RMSEs decrease rapidly as M increases. In other words, in the massive MIMO system, the transmitted power can be significantly reduced due to an appropriately large number of the BS antennas.

We present an evaluation of the computational complexity using TIC and TOC instruction in MATLAB. TIC and TOC instruction can be used to calculate

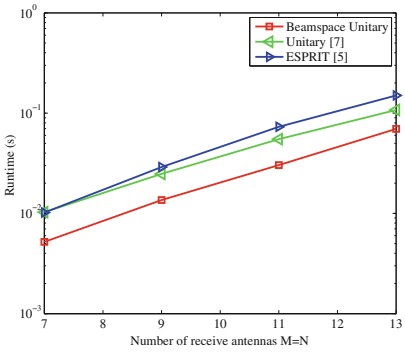


Fig. 3. Runtime for overall angle estimation versus the number of antennas

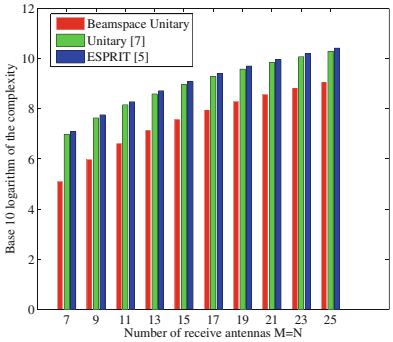


Fig. 4. Computational complexities versus the number of antennas

the runtime of an algorithm. Figure 3 shows the runtime against the number of array $M = N$. From Fig. 3, we can see that the runtime of the proposed approach is much smaller than the compared algorithm in [7,9].

In Fig. 4, the computation complexities are compared. For the element ESPRIT [7] and the unitary ESPRIT algorithm [9], the computational complexities of eigendecomposition are $O(4M^3N^3W^3)$ and $O(M^3N^3W^3)$. Unlike the element ESPRIT and the unitary ESPRIT algorithm, the beamspace unitary ESPRIT algorithm transforms the original data vector into lower-dimensional beamspace. If only $L_W < W$ transmitting beams and $L_M L_N < MN$ receiving beams are selected, the computational complexity could be reduced to $O(L_M^3 L_N^3 L_W^3)$.

4 Conclusions

In this paper, we have addressed the joint estimation of DOD and 2D DOA in massive MIMO systems. The conjugate centrosymmetrized beamspace transformed vectors are exploited to transform the received data into beamspace.

Then the new rotational invariance of the transmitting beams and the receiving beams is applied to estimate the DOD and the 2D DOA of different paths. The proposed beamspace unitary algorithm involves only real-valued computation and does not need pairing process, which leads to the reduction of computational complexity. Moreover, the simulation results shows that the proposed method can effectively estimate DOD and 2D DOA with reduced complexity and automatic pairing, while not debasing the performance of angle estimation.

References

1. Hossain, E., Hasan, M.: 5G cellular: key enabling technologies and research challenges. IEEE Instrum. Measur. Mag. **18**(3), 11–21 (2015)
2. Marzetta, T.L.: Noncooperative cellular wireless with unlimited numbers of base station antennas. IEEE Trans. Wirel. Commun. **9**(11), 3590–3600 (2010)
3. Rusek, F., Persson, D., Lau, B.K., et al.: Scaling up MIMO: opportunities and challenges with very large arrays. IEEE Sig. Process. Mag. **30**(1), 40–60 (2013)
4. Yao, B., Wang, W., Yin, Q.: DOD and DOA estimation in bistatic non-uniform multiple-input multiple-output radar systems. IEEE Commun. Lett. **16**(11), 1796–1799 (2012)
5. Chan, K.W., So, H.C., Huang, L., Huang, L.T.: Parameter estimation and identifiability in bistatic multiple-input multiple-output radar. IEEE Trans. Aeros. Electro. Syst. **51**(3), 2047–2056 (2015)
6. Zhang, X., Xu, L., Xu, L., Xu, D.: Direction of departure (DOD) and direction of arrival (DOA) estimation in MIMO radar with reduced-dimension MUSIC. IEEE Commun. Lett. **14**(12), 1161–1163 (2010)
7. Chen, D., Chen, B., Qin, G.: Angle estimation using ESPRIT in MIMO radar. Electron. Lett. **44**(12), 770–771 (2008)
8. Guo, Y.D., Zhang, Y.S., Tong, N.N.: Beamspace ESPRIT algorithm for bistatic MIMO radar. Electron. Lett. **47**(15), 876–878 (2011)
9. Chen, J., Gu, H., Su, W.: Angle estimation using ESPRIT without pairing in MIMO radar. Electron. Lett. **44**(24), 1422–1423 (2008)
10. Li, J., Jiang, H., Wei, Y.: DOD-DOA-polarization estimation in large MIMO radar system based on random matrix theory. In: IEEE International Conference on Computational Intelligence and Communication Technology (CICT), Ghaziabad, India, pp. 334–338, February 2015
11. Jiang, H., Wang, D., Liu, C.: Estimation of DOD and 2D-DOA and polarizations for bistatic MIMO radar. In: IEEE 19th Annual Wireless and Optical Communications Conference (WOCC), Shanghai, China, May 2010
12. Zoltowski, M.D., Haardt, M., Mathews, C.P.: Closed-form 2D angle estimation with rectangular arrays in element space or beamspace via unitary ESPRIT. IEEE Trans. Sig. Process. **44**(2), 316–328 (1996)
13. Zhu, Y., Liu, L., Wang, A., Sayana, K., Zhang, J.: DoA estimation and capacity analysis for 2D active massive MIMO systems. In: IEEE International Conference on Communications (ICC), Budapest, Hungary, January 2013



Non-orthogonal Multi-carrier Technology for Space-Based Internet of Things Applications

Mengdi Zhao[1(✉)], Hongxing Li[2], Yi Li[1], Li Fang[1], and Ping Chen[1]

[1] Beijing University of Posts and Telecommunications, Beijing 100876, China
{zhaomengdi,liyi,fangli,rab1106}@bupt.edu.cn
[2] Shanghai Aerospace Electronic Technology Institute, Shanghai 201109, China
shulhx@sina.com

Abstract. The Internet of things (IoT) technology plays an important role in shipping, vessel tracking, marine engineering and other fields. The space-based Internet of things (S-IoT) is an extension of the IoT in the space field. It is the integration of satellite communications and emerging IoT technology, which will greatly promote the development of the earth integrated network. In this paper, a new non-orthogonal multi-carrier modulation scheme is designed for the machine-to-machine (M2M) communication system to make it applied to space-based networking. This scheme can effectively improve the transmission rate and reduce the sensitivity to frequency offset, so it is suitable for a wide range of data communication environment. The simulation results indicate that the scheme proposed can effectively restrain inter-carrier interference (ICI) caused by non-orthogonal subcarriers, and realize the high speed data transmission rate, which provides a new way for information collection and ship monitoring.

Keywords: Space-based Internet of things
Non-orthogonal multi-carrier modulation
Machine-to-machine communication · Inter-carrier interference

1 Introduction

The Internet of things (IoT) could be conceptually defined as a dynamic global network based on standard and communication protocols where physical things and smart objects are integrated into the information network and can interact and communicate among themselves [1]. IoT achieves smart connectivity with existing networks anytime and anywhere not only by anyone but also by anything. Nowadays, with the rapid development of the Internet and communication technology, the IoT has been widely supported and applied [2]. The application of IoT relying on the ground network is becoming more and more mature [3]. However, in some large-scale, cross-regional, harsh environment and other applications, due to space, environment and other restrictions, the IoT based on the

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 37–45, 2018.
https://doi.org/10.1007/978-3-319-78130-3_5

ground network is not able to meet the needs of data exchange. Especially in the marine management, ship monitoring and other fields, the IoT based on the ground network has not been effectively supported. Therefore, the development of space-based Internet of things (S-IoT) based on space-based communication network has its real needs and prospects. However, at present, the protocols, technologies and documents of the IoT are mainly based on the ground network, and the research of S-IoT has just started. The integration of satellite communication technology and the IoT and big data is very essential for realizing the interconnection of all things [4,5]. At present, the academic and industrial circles of various countries are actively considering the integration of space-based network and ground network [6,7], so that users under different application scenarios can access the IoT system at any time in different ways, such as satellite, Wi-Fi, 4G, 5G, etc.

The S-IoT is the extension of the IoT in the space-based field. The S-IoT can make connection to objects that cannot be connected to the ground network by using space network, which is a typical space network application. Generally speaking, the S-IoT integrates space-based navigation, remote sensing, communications and other services, and is a barrier-free interactive information service platform. As shown in Fig. 1, the stratified reference model of the S-IoT can be divided into three layers: application layer, network layer and sensor layer.

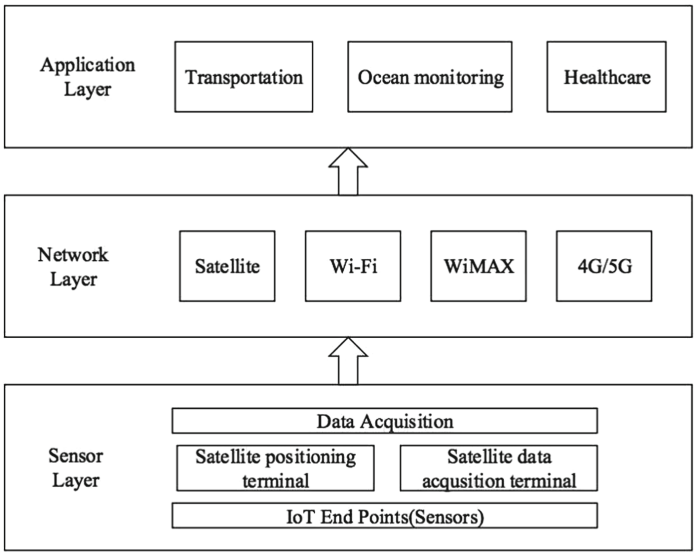


Fig. 1. S-IoT reference model.

Satellite communication plays an important supporting role in the three layers of IoT. At the sensor layer, satellite positioning terminals, satellite data acquisition terminals and other devices are used to collect the location, speed,

movement time and other information. At the network layer, satellite based connectivity is one of the possible solutions. At the application layer, the useful information transmitted through the satellite communication network becomes an important component of the information resources of various application platforms. IoT connectivity through the gateway has become one of several access methods for satellite networking [5].

Maritime satellite communications can provide users with speed, location, time and other basic information. It has the characteristics of all-weather, global coverage, and high precision. The application of satellite technology in S-IoT can support and complement the terrestrial network system, which can provide accurate time information, high-precision location information and navigation and monitoring services for S-IoT applications and plays an irreplaceable role in S-IoT. However, machine-to-machine (M2M), the key technology for IoT, is also experiencing tremendous growth in the satellite communications market. In fact, satellite based IoT type solutions are deployed now for industrial M2M data communication [5]. Moreover, the coverage of S-IoT is very wide, and this results in the generation of enormous amounts of data which have to be transferred and stored [8]. In order to achieve high-speed transmission, multi-carrier modulation technology can be adopted as one of the M2M communication technologies for S-IoT, but the existing multicarrier-code division multiple access (MC-CDMA) system and orthogonal frequency division multiplexing (OFDM) technology are not applicable.

In order to solve this problem, this paper proposes a scheme combining the non-orthogoal multi-carrier modulation technology with S-IoT, which uses satellite channels to transmit the data of S-IoT to satisfy the requirement of transmission capacity and bandwidth, and provides a new implementation for M2M communication. The scheme can effectively improve the transmission rate, and the non-orthogonality between subcarriers can reduce the sensitivity to frequency offset, so it is very suitable for maritime satellite communications in the ship automatic identification, tracking, positioning and other applications. The simulation results show that the proposed scheme has a certain feasibility and can ensure the reliability of the communication system, so the goal of high-speed transmission can be achieved.

2 Non-orthogonal Multi-carrier Modulation Technology

The Doppler effect of the low earth orbit (LEO) satellite communication system leads to the generation of frequency offset. The Doppler frequency shift f_d can be calculated as follows:

$$f_d = \frac{\nu}{\lambda} \cos\theta \tag{1}$$

In low-orbit satellite systems, the satellite orbit height is between 500 Km and 2000 km. The ship transmits signals through the VHF maritime frequency band, and the operating bands are 161.975 MHz and 162.025 MHz [9]. According to the wavelength formula $\lambda = c/f$,velocity of light $c = 3 \times 10^8$ m/s and the frequency

f is calculated at 162 MHz, so the carrier wavelength λ is estimated to be about 1.85 m. Ship speed is negligible compared to satellite speed, so the relative speed ν of the satellite and the user is about 24807−27371 km/h.

The angle θ between the connection between the satellite and the ship and the direction of the satellite's speed ν is:

$$\cos\theta = \frac{\overrightarrow{AB}\cdot\overrightarrow{OE}}{\left|\overrightarrow{AB}\right|\cdot\left|\overrightarrow{OE}\right|} = \frac{R\sin\theta_2\cos\varphi\cos\theta_2 - R\cos\theta_2\sin\theta_2}{\left|\overrightarrow{AB}\right|} \tag{2}$$

When $\theta = 0°$, the maximum Doppler frequency:

$$f_d = \frac{\nu}{\lambda}\cos\theta = \frac{24807 - 27371}{1.85\times 3.6}\cos 0° = 3.72 - 4.11\,\text{kHz} \tag{3}$$

Due to the large frequency offset of LEO satellite communication system, it is not suitable for the use of OFDM. The reason is that OFDM technology is particularly sensitive to the frequency offset. That is to say, any frequency offset will destroy the orthogonality among subcarriers, which will inevitably lead to inter-carrier interference (ICI). In addition, one of the major drawbacks of the OFDM system is the high peak to average power ratio (PAPR), which requires high dynamic level of power amplifier in transmitter and raises the cost of hardware implementation. Therefore, the non-orthogonal multi-carrier modulation mechanism is a better choice.

In order to utilize bandwidth effectively, 16QAM modulation is chosen as the modulation mode in this paper. The multi-carrier modulation system has 32 non-orthogonal subcarriers as shown in Fig. 2. Note form the figure that this waveform is comprised of 32 multi-carriers. Each carrier is modulated by 16QAM to generate 4-bit symbols.

The subcarrier center frequency shall be defined by the following formula [10]:

$$f_m = (0.5615 - (M/2 - m)\times 1.125)/T \tag{4}$$

where T is the symbol duration, M is the number of subcarriers, and $m = 0, 1, 2, \ldots, M-1$. This leads to a carrier spacing Δf of 2.7 kHz. The modulated signal, at carrier frequency f_c, shall be given by:

$$M(t) = \text{Re}\{s(t)\exp(j(2\pi f_c t + \phi_0))\} \tag{5}$$

where ϕ_0 is an arbitrary phase, and $s(t)$ is the complex envelope of the modulated signal defined as:

$$s_m(t) = \sum_{n=1}^{N} S_m(n)g(t - t_n)e^{j(\omega_m t + \varphi_m)} \tag{6}$$

$$s(t) = \sum_{m=0}^{M-1} S_m(t) = \sum_{m=0}^{M-1}\sum_{n=1}^{N} S_m(n)g(t - t_n)e^{j(\omega_m t + \varphi_m)} \tag{7}$$

where M is the number of subcarriers, N is the number of modulated symbols on each subcarrier in one slot, T is the symbol duration on each subcarrier, $t_n = nT$ is the symbol time corresponding to modulation symbol $S_m(n)$, $S_m(n)$ is the modulation symbol at time t_n on subcarrier m, $\omega_m = 2\pi f_m$ is subcarrier angular frequency, φ_m is the phase control for subcarrier m during the slot, and $g(t)$ is the ideal symbol waveform [10].

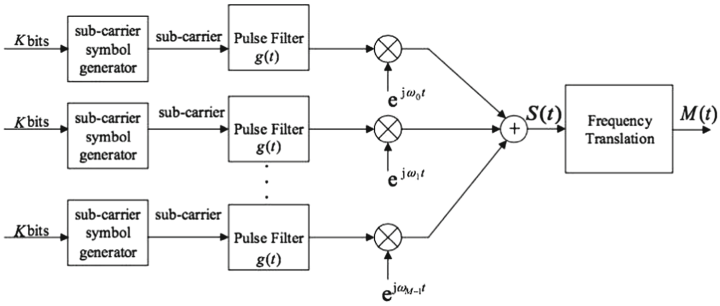


Fig. 2. Block diagram of the modulation process.

3 S-IoT Based on Non-orthogonal Multi-carrier Modulation Technology

3.1 Scheme Implementation

According to the above requirements, in this paper a multi-carrier modulation scheme is designed based on non-orthogonal subcarriers. The proposed scheme can be applied to S-IoT, which can improve the transmission rate and meet the requirement of wide range communication through satellite transmission.

Since the subcarrier frequencies are non-orthogonal, the non-orthogonal multi-carrier modulation scheme is realized by the idea of non-orthogonal multiple modulation (NOMM) [11,12].

It is known x is the K dimensional real random vector, and y is a linear mapping of x to the $L^2(R)$ space:

$$y = Gx \tag{8}$$

where $x = [x_0, x_1, \ldots, x_{K-1}]^{\mathrm{T}}$, and $G = [g_0, g_1, \ldots, g_{K-1}]$. The column vectors of G are linearly independent, but not necessarily orthogonal. In this system, y is called a composite wave, and x is called a signal vector.

The signal points of NOMM are distributed on the K dimensional constellation based on the vector of G matrix. G can be transformed into a standard orthogonal basis. In this way, NOMM's constellation design and traditional modulation methods can be unified.

If $R = G^{\mathrm{T}}G$, then R is a positive definite matrix. Thus, there is a K-order real matrix P, which makes

$$PRP^{\mathrm{T}} = I \tag{9}$$

As a result, the column vector of G is the standard orthogonal basis. Thus, (8) becomes

$$y = GP^{\mathrm{T}}x \tag{10}$$

The demodulator structure at the receiver is shown in Fig. 3. In the multiplier, the input vector is multiplied by the P matrix. And the judgment is made in the decision device.

At the transmitter and receiver, the P matrix is used to make a linear transformation of the non-orthogonal G matrix, so that crosstalk or interference between different symbols can be effectively eliminated.

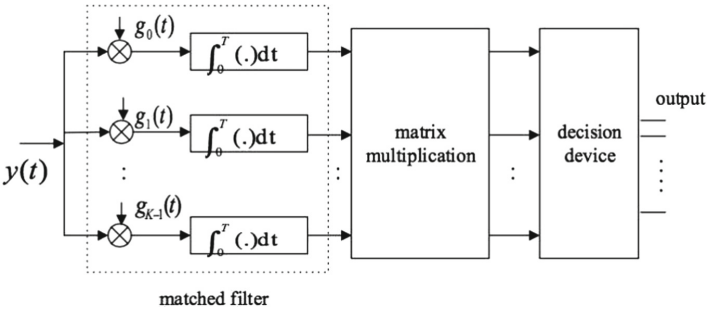


Fig. 3. Structure of NOMM demodulator.

NOMM can be extended to the non-orthogonal multi-carrier system. According to the basic principle of OFDM modulation and demodulation in [13], the serial bit stream is mapped to a serial symbol after passing the data encoder, and each symbol modulates one of the subcarriers. Then, for non-orthogonal multi-carrier technology, each subcarrier is modulated by the 16QAM symbol, that is to say, the real and imaginary parts of the 16QAM equivalent baseband signal are multiplied by the cosine and sine components of the corresponding non-orthogonal subcarriers respectively, and the addition of the modulated subcarriers constitutes the multi-carrier symbol.

From the above ideas, the row vector of x as the signal vector in (8) represents the 16QAM signal after the serial-to-parallel conversion, and the column vector of G corresponds to the subcarrier frequency. Since the 16QAM signal is a complex signal and contains the real part and the imaginary part, the real and imaginary parts of the 16QAM signal are processed separately. The real part is multiplied by the $\cos(2\pi f_m t)$ component of the subcarrier, and the imaginary part is multiplied by the $\sin(2\pi f_m t)$ component.

According to this idea, the first 32 column vectors of the G matrix are $\cos(2\pi f_m t)$ frequency components, while the last 32 column vectors of the G matrix are the $\sin(2\pi f_m t)$ frequency components:

$$g_k = \begin{cases} \cos(2\pi f_k t), & k = 1, 2, \ldots, 32 \\ \sin(2\pi f_k t), & k = 33, 34, \ldots, 64 \end{cases} \tag{11}$$

Correspondingly, the first 32 row vectors of signal vector x are the real parts of the 16QAM signal, and the last 32 row vectors are the imaginary parts of the 16QAM signal. In this way, in the noiseless channel, error-free demodulation can be achieved at the receiver. The non-orthogonal multi-carrier modulation technology can suppress ICI effectively. Therefore, the proposed scheme applied to the S-IoT can achieve high-speed data interactions between ships in the S-IoT.

3.2 Performance Results

This section uses computer simulations to test the bit error rate (BER) performance of the proposed scheme. In the simulation, the modulation mode is 16QAM with 32 non-orthogonal subcarriers. The subcarrier spacing is 2.7 kHz, and each subcarrier is modulated by 16QAM. The modulated multi-carrier symbols are transmitted through the AWGN channel, and the signal-to-noise ratio (SNR) is in the range of 0 dB–16 dB.

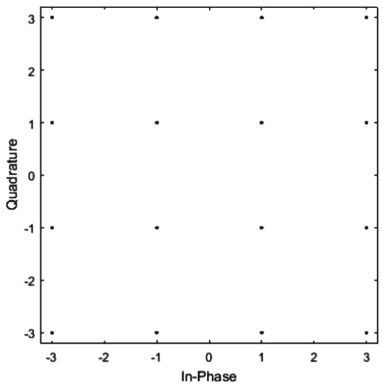


Fig. 4. Multi-carrier demodulation constellation (non-noise).

Figure 4 shows the signal constellation of 32 non-orthogonal subcarriers after modulation in the absence of noise. Because it has not been demodulated by 16QAM, the constellation is 16QAM constellation. Obviously, in the channel without noise, although the subcarriers are non-orthogonal, its non-orthogonality does not affect the performance of the system, which can be visually reflected from the constellation diagram in Fig. 4. Therefore, the scheme designed in this

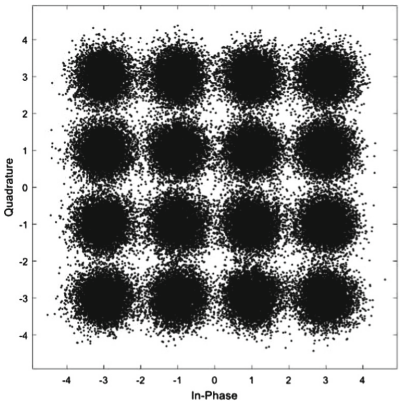


Fig. 5. Multi-carrier demodulation constellation (SNR = 10 dB).

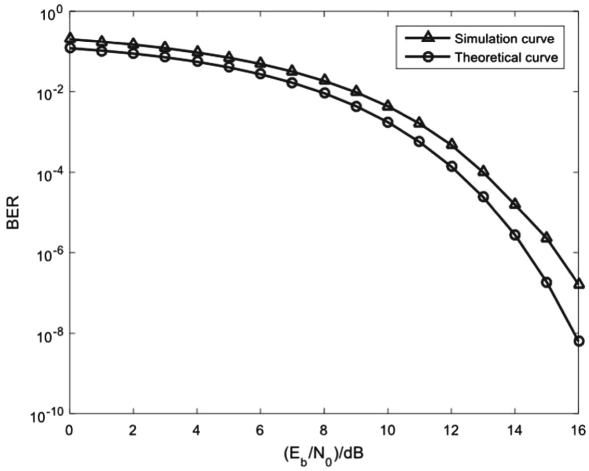


Fig. 6. Bit error rate performance.

paper can effectively eliminate the subcarrier interference caused by the non-orthogonality of subcarriers, so as to obtain the performance when orthogonal. Figure 5 shows the signal constellation of 32 non-orthogonal subcarriers after modulation when SNR is 10 dB.

In digital satellite communication systems, E_b/N_0 is usually used to evaluate the performance of the system, where E_b represents the energy of the unit bit data signal, and N_0 represents the power of the noise within the unit bandwidth. Figure 6 shows the relationship between the BER of the non-orthogonal multi-carrier system and E_b/N_0, and the comparison with the theoretical curve of the 16QAM BER. It is easy to see that the BER decreases with the increase of E_b/N_0, and is close to the theoretical curve to meet the target BER.

4 Conclusions

In this paper, the traditional satellite communication technology is combined with the IoT, and a new non-orthogonal multi-carrier modulation scheme applied to the S-IoT is designed and implemented for inter-ship communication. The proposed scheme has the characteristics of large communication bandwidth, wide communication coverage and high communication reliability, and can overcome the interference caused by non-orthogonality of subcarriers to achieve the target BER. The basic design and principle of the scheme are given in this paper, and the availability and effectiveness of the proposed scheme are verified by simulation. The next step is to combine non-orthogonal multi-carrier modulation technology with channel coding to further improve the overall system performance.

References

1. Vermesan, O., Friess, P., Guillemin, P., Gusmeroli, S., Sundmaeker, H.: Internet of things strategic research roadmap. Inf. Secur. Technol. **29**(16), 300–304 (2014)
2. Yue, H., Guo, L., Li, R., Asaeda, H., Fang, Y.: DataClouds: enabling community-based data-centric services over the internet of things. IEEE Internet Things J. **1**(5), 472–482 (2014)
3. Minoli, D., Sohraby, K., Occhiogrosso, B.: IoT considerations, requirements, and architectures for smart buildings-energy optimization and next-generation building management systems. IEEE Internet Things J. **4**, 269–283 (2017)
4. Xu, K., Qu, Y., Yang, K.: A tutorial on the internet of things: from a heterogeneous network integration perspective. IEEE Netw. **30**(2), 102–108 (2016)
5. Chong, R.S.E.: Systems engineering of a terabit elliptic orbit satellite and phased array ground station for IoT connectivity and consumer internet access. IEEE Access **4**, 9941–9957 (2016)
6. Maalel, N., Natalizio, E., Bouabdallah, A., Roux, P., Kellil, M.: Reliability for emergency applications in internet of things. In: IEEE International Conference on Distributed Computing in Sensor Systems, vol. 19, no. 14, pp. 361–366 (2013)
7. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., Ayyash, M.: Internet of things: a survey on enabling technologies, protocols, and applications. IEEE Commun. Surv. Tuts. **17**(4), 2347–2376 (2015)
8. Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M.: Internet of things (IoT): a vision, architectural elements, and future directions. Future Gener. Comput. Syst. **29**(7), 1645–1660 (2013)
9. Technical characteristics for a VHF data exchange system in the VHF maritime mobile band (Recommendation ITU-R M.2092-0). International Telecommunications Union (2014)
10. Terrestrial Trunked Radio (TETRA); Voice Plus Data (V+D); Part2: Air Interface (AI). ETSI (2007)
11. Lei, C., Sun, S.: The demodulation of non-orthogonal multiple modulation based all phase. In: IEEE International Conference on Circuits and Systems for Communications, pp. 235–239 (2008)
12. He, M., Liang, D., Cao, Q.: A modulation with higher bandwidth efficiency than OFDM. In: International Conference on Signal Processing Systems, vol. 1, pp. V1-393–V1-397 (2010)
13. Pandharipande, A.: Principles of OFDM. IEEE Potentials **21**(2), 16–19 (2002)



A Novel GPS Spoofing Algorithm Based on Modifying Navigation Message

Qiong Yang(✉), Yi Zhang, and Chengkai Tang

School of Electronics Engineering, Northwestern Polytechnical University, Xi'an 710072, China
20l4nickyoung@mail.nwpu.edu.cn

Abstract. Satellite navigation spoofing technology induces the victim receivers to capture and track the forged navigation signal, so that the receivers work out the spoofed position and time. This paper presents a novel GPS spoofing method based on modifying navigation message, modify the handover word, spoof the calculated pseudo ranges and the satellite positions, then spoof the positioning result of receiver. The influence of the spoofing method on GPS receiver have been simulated, the results show that the proposed method has the approximate linear relationship with the eastern and northern spoofed distance, but has nonlinear relationship with the up spoofed distance. It has great significance to airspace supervision and defense.

Keywords: GPS · Navigation message modification · Spoofing
Handover word (HOW)

1 Introduction

With the rapid development of the global navigation satellite system (GNSS), many military and key civilian areas increasingly rely on satellite navigation systems. However, the progress of the jamming technology makes the security of the satellite navigation system challenged. GPS jamming technology can be divided into suppression jamming and spoofing. Spoofing has the characteristics of low power and high concealment, which is the main method of jamming. Iran used GPS spoofing to capture the US RQ170 unmanned reconnaissance aircraft in 2011, capture Sentry drone in 2012 and trap two US navy patrol boats in January 2016 [1]. Wang use multiple directional receiving antennas to control the time delay of each satellite signal accurately and success spoof receiver time. However, the delay of the kind of spoofing algorithm cannot exceed 15 μs; otherwise it is easy to be detected by the abnormal clock difference alarm [2]. Reference [3] completes the step by step guided spoofing of aircraft trajectory and calculating the delay and Doppler variation of the satellite signal by radar. But these spoofing modes require complex algorithm and expensive devices. In reference [4], the integrated navigation system is proposed to spoof the incoming GPS guidance weapon, yet the harshness of transmitter distribution limits engineering implementation of such spoofing patterns. In addition, with the development of spoofing detection and anti-spoofing technology, the traditional time delay control algorithm is invalid in clock frequency drift anomaly detection [5], power anomaly

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 46–53, 2018.
https://doi.org/10.1007/978-3-319-78130-3_6

detection [6], anti-spoofing based on tracking loop filtering [7], anti-spoofing based on antenna array [8, 9]. This paper proposes a GPS spoofing algorithm based on modified navigation messages. By investigating the positioning principle of GPS receiver. Modification the Hand over Word (HOW) in the navigation messages every 6 s, makes the receiver get the wrong Time of Week (TOW) Counter, then get the false pseudo range and the satellite positions. The simulation results show the effect of spoofing signal on the localization of GPS software receiver and the relation between the modified value HOW and the spoofed distance.

2 Principle of Receiver

The pseudorange single point positioning method of GPS receiver is obtained by receiving the C/A code from the satellites as ranging signals, measure the distance between the receiver and the satellites, then the three dimensional coordinates of the receiver are solved. It needs at least 4 satellite signals to solve the 4 unknown parameters which are three dimensional coordinates and clock error.

$$\begin{cases} \rho_1 = \sqrt{(x-x_1)^2+(y-y_1)^2+(z-z_1)^2}+b \\ \rho_2 = \sqrt{(x-x_2)^2+(y-y_2)^2+(z-z_2)^2}+b \\ \rho_3 = \sqrt{(x-x_3)^2+(y-y_3)^2+(z-z_3)^2}+b \\ \rho_4 = \sqrt{(x-x_4)^2+(y-y_4)^2+(z-z_4)^2}+b \end{cases} \tag{1}$$

$\rho_1, \rho_2, \rho_3, \rho_4$ donate the pseudorange of four satellite respectively, b is receiver clock error equivalent distance. (x, y, z) is the coordinates of the receiver in the WGS-84 coordinate system, (x_i, y_i, z_i) is the coordinates of the i-th satellite in the WGS-84 coordinate system. The pseudo ranges and satellite coordinates are closely related to $t^{(S)}$ which is the launch time of satellite signal.

$$t^{(S)} = TOW + 0.020(30w + b) + 0.001(c + CP/1023) \tag{2}$$

TOW is shot for the Time of Week Counter of navigation message, w donates the w-th received navigation message data code word of the receiver and each word contains 30 bits. In the current word, the receiver gets b bits of message. The length of each bit is 20 ms; the receiver has received c entire period C/A navigation codes whose period is one millisecond. CP donates the code phase measurement. The above parameters constitute the launch time of satellite signal.

3 Proposed Spoofing Algorithm

The civilian L1 signal transmitted from GPS satellites is described in this equation

$$S_{L1ca}(t) = A_c C_i(t) D_i(t) \sin(\omega_{L1}(t) + \varphi_{L1}) \tag{3}$$

where A_c is amplitude of C/A code, C_i is i-th PRN C/A code, D_i is i-th PRN navigation message, ω_{L1} is the angular frequency of L1 signal, φ_{L1} is signal phase.

It is well known that the signal structure, modulation type, navigation message format, carrier frequency and C/A code of GPS are all public. The navigation data is predictable and stable, GPS navigation data is maintained predictability within two hours, the ephemeris is updated every 12.5 min, meanwhile the receiver can obtain ephemeris data in one minute. This leads to the GPS signal to be easily spoofed.

According to the navigation message format, Handover Word is the second word of every sub-frame and is updated every 6 s. HOW is the first 17 bits of TOW, if the HOW is counterfeited, TOW will be cheated accordingly. Thus, the launch time of the satellite signal is modified. So the satellite's pseudo range and position are spoofed and then the receiver gets the wrong positioning results and time.

$$S^s_{L1ca}(t) = A^s_c C_i(t) D^s_i(t) \sin(\omega_{L1}(t) + \varphi_{L1}) \tag{4}$$

Although the parity code in each word helps the receiver to check if the demodulated messages contain the error bits. The parity check codes in a word are generated by calculating the first 24 bits of the current word and the last 2 bits of the previous frame according to formula 5.

$$\begin{cases} D_i = D_{\overline{30}} \oplus d_i, i = 1,2,...,24 \\ D_{25} = D_{\overline{29}} \oplus d_1 \oplus d_2 \oplus d_3 \oplus d_5 \oplus d_6 \oplus d_{10} \oplus d_{11} \oplus d_{12} \oplus d_{13} \oplus d_{14} \oplus d_{17} \oplus d_{18} \oplus d_{20} \oplus d_{23} \\ D_{26} = D_{\overline{30}} \oplus d_2 \oplus d_3 \oplus d_4 \oplus d_6 \oplus d_7 \oplus d_{11} \oplus d_{12} \oplus d_{13} \oplus d_{14} \oplus d_{15} \oplus d_{18} \oplus d_{19} \oplus d_{21} \oplus d_{24} \\ D_{27} = D_{\overline{29}} \oplus d_1 \oplus d_3 \oplus d_4 \oplus d_5 \oplus d_7 \oplus d_8 \oplus d_{12} \oplus d_{13} \oplus d_{14} \oplus d_{15} \oplus d_{16} \oplus d_{19} \oplus d_{20} \oplus d_{22} \\ D_{28} = D_{\overline{30}} \oplus d_2 \oplus d_4 \oplus d_5 \oplus d_6 \oplus d_8 \oplus d_9 \oplus d_{13} \oplus d_{14} \oplus d_{15} \oplus d_{16} \oplus d_{17} \oplus d_{20} \oplus d_{21} \oplus d_{23} \\ D_{29} = D_{\overline{30}} \oplus d_1 \oplus d_3 \oplus d_5 \oplus d_6 \oplus d_7 \oplus d_9 \oplus d_{10} \oplus d_{14} \oplus d_{15} \oplus d_{16} \oplus d_{17} \oplus d_{18} \oplus d_{21} \oplus d_{22} \oplus d_{24} \\ D_{30} = D_{\overline{29}} \oplus d_3 \oplus d_5 \oplus d_6 \oplus d_8 \oplus d_9 \oplus d_{10} \oplus d_{11} \oplus d_{13} \oplus d_{15} \oplus d_{19} \oplus d_{22} \oplus d_{23} \oplus d_{24} \end{cases} \tag{5}$$

Where d_1, d_2 … d_{24} are 24 raw data bits, D_1, D_2 … D_{24} are 24 real data bits transmitted by satellites, D_{25}, D_{26} … D_{30} are 6 bits of real parity check codes, $D_{\overline{29}}$, $D_{\overline{30}}$ are the last two parity check codes of previous word.

Since power of received GPS signal is low on the surface of the earth, a slightly higher power of spoofing signal can invade the receiver and it's hard to be detected. Figure 1 shows the GPS spoofing system. Spoofing signal is generated in three steps:

Step 1: Acquiring and saving the authentic GPS signals,estimating the C/A code, carrier Doppler frequency shift, navigation messages of satellites.

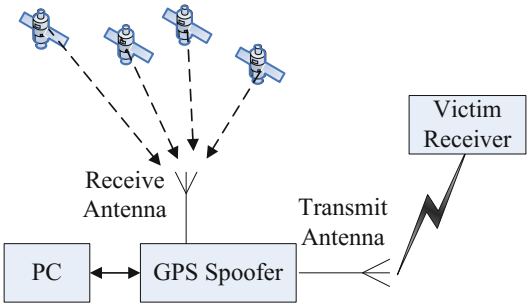


Fig. 1. GPS spoofing system

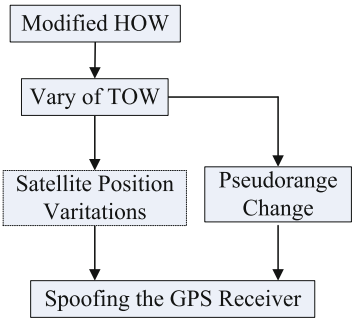


Fig. 2. Effects of GPS spoofing signal

Step 2: Modifying the Handover word of navigation message sub-frame every 6 s, while modifying the 6 bits parity check codes of HOW and other message keep intact.

Step 3: Amplifying and transmitting the counterfeit signal toward the target receiver, the spoofing signal power is higher than normal GPS signal by about 3 dB, so as to easily invade the victim receiver' acquisition and tracking loop (Fig. 1).

4 Simulations

To verify the proposed GPS spoofing algorithm, this paper counterfeits signal based on the real GPS signal and analyzes the influence of spoofing signal on GPS software receiver with Matlab R2009a. The receiver outputs the positioning results every 0.5 s, the time of week counter is modified at 12 s and the modified value change from −30 s to 30 s. Since the TOW appears once every 6 s, the modified value is a multiple of 6. The position errors due to spoofing are reported in details.

4.1 Scene1

The location of GPS Receiver A is (394494, 4991271, 129), the 4 strongest GPS signals captured by receiver A are PRN22, PRN3, PRN19, PRN15, Table 1 and Fig. 1 show the modified HOW value versus the spoofing distance of GPS receiver in the East, North, Up coordinates.

The results of Table 1 show that the spoofed error varies from 6385 m to about 32751 m with the modified value of HOW. The east and north spoofed error change linearly, the eastern spoofed rate is about 483 m/s and the northern spoofed rate is about −341 m/s. however, the relation between up spoofed error and modified HOW is

Table 1. Spoofed distance of GPS receiver A

HOW modified value/s	ΔE/m	ΔN/m	ΔU/m	Position error/m
−30	−15761	13549	−25312	32751
−24	−11848	8930	9808	17785
−18	−8714	6260	14312	17887
−12	−5790	4113	10515	12688
−6	−2896	2054	5312	6389
6	2907	−2051	−5303	6385
12	5818	−4103	−10646	12807
18	8738	−6162	−15716	19008
24	11731	−8378	−17304	22521
30	15101	−4540	261	19042

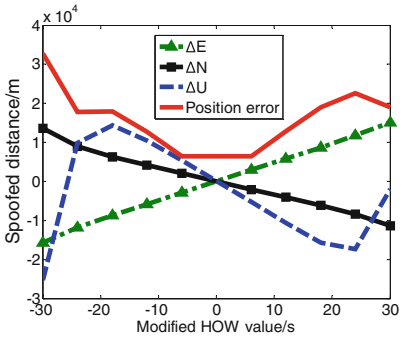


Fig. 3. Spoofing Error of receiver A

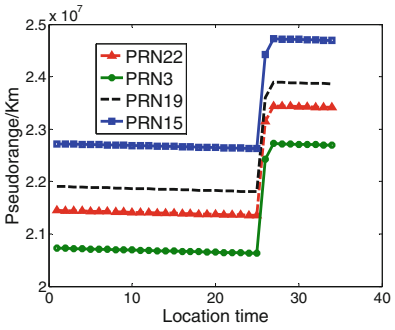


Fig. 4. Psedorange of receiver A

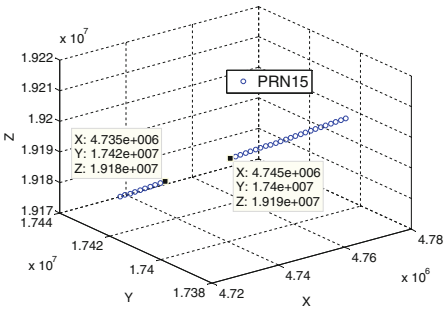


Fig. 5. Satellite PRN15 position

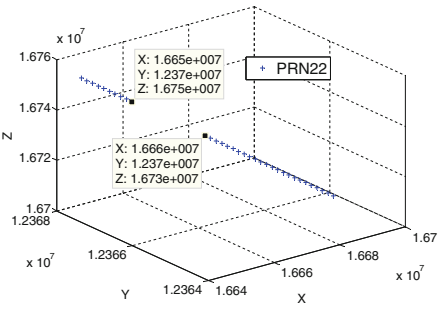


Fig. 6. Satellite PRN22 position

nonlinear. Figure 4 shows the variation of satellite pseudorange. The pseudorange of PRN22, PRN3, PRN19, PRN15 have the same amount jump after twelfth seconds. According to formula 1, the same amount change of pseudorange does not affect the position results. Figures 5 and 4 are respective calculated positions of satellite 15 and

Table 2. GPS Receiver B spoofed distance.

HOW modified value/s	ΔE/m	ΔN/m	ΔU/m	Position error/m
−30	−15761	13549	−25312	32751
−24	−11848	8930	9808	17785
−18	−8714	6260	14312	17887
−12	−5790	4113	10515	12688
−6	−2896	2054	5312	6389
6	2907	−2051	−5303	6385
12	5818	−4103	−10646	12807
18	8738	−6162	−15716	19008
24	11731	−8378	−17304	22521
30	15101	−4540	261	19042

satellite 22 under the condition that the modified HOW value is 6 s. The position of PRN15 jumps from (4744615, 17404895, and 19193961) to (4734899, 17417852, 19184444) at the twelfth second. The position of PRN22 jumps from (16664851, 12366252, and 16733087) to (16651186, 12366994, 16746261) (Figs. 5 and 6).

4.2 Scene2

The location of GPS Receiver B is (394494, 4991271, 129), the 4 strongest GPS signals captured by receiver B are PRN21, PRN22, PRN15, PRN18, Table 2 and Fig. 7 show the modified HOW value versus the spoofing distance of GPS receiver in the East, North, Up coordinates.

The results of Table 2 show that the spoofed error varies from 1604 m to about 4128 m with the modified value of HOW. The east and north spoofed error change linearly, the eastward spoofed rate is about 216 m/s and the northern spoofed rate is about −155 m/s. however, the relation between up spoofed error and modified HOW is nonlinear. Figure 8 shows the variation of satellite pseudorange. The pseudorange of PRN21, PRN22, PRN15, PRN18 have the same amount jump after twelfth seconds.

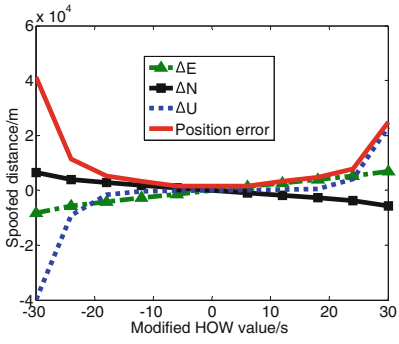


Fig. 7. Spoofing error of receiver B

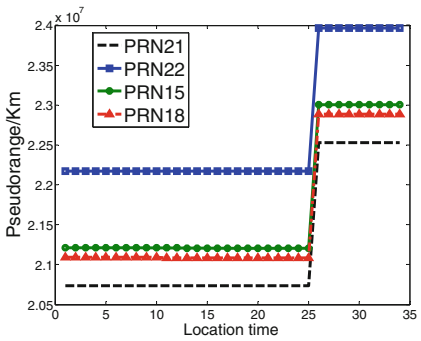


Fig. 8. Pseudorange of receiver B

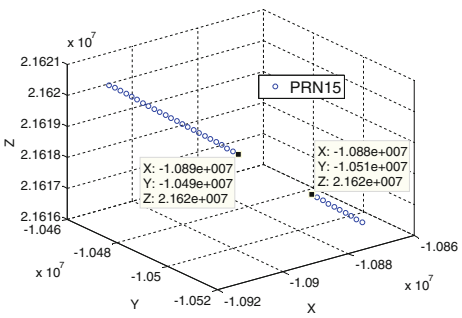


Fig. 9. Satellite PRN15 position

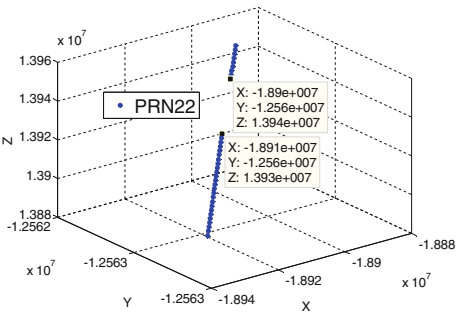


Fig. 10. Satellite PRN22 position

However, there is no effect on the position results. Figures 9 and 10 are respective calculated positions of satellite 15 and satellite 22 under the condition that the modified HOW value is 6 s. The position of PRN15 jumps from (−10894067, −10494541, 21618652) to (−10882420, −10508480, 21617551) at the twelfth second. The position of PRN22 jumps from (−18911908, −12562454, 13924308) to (−18900482, −12562260, 13940171).

The above two simulation experiments show that the proposed GPS spoofing algorithm based on modifying Handover word can spoof the position results effectively. The influence of modified HOW values to east and north spoofing error have an approximately linear relationship, but the receivers at different locations have different eastern and northern spoofing rate. The relationship between the modified value and up spoofing error is nonlinear. The reason for the successful spoofing GPS receiver is that the modified HOW spoofs the satellite position calculated by the receivers.

5 Conclusion

This paper proposes a GPS spoofing algorithm based on modifying the Handover word of navigation message and analyzes the effect on the position results of GPS software receiver. The simulation results show that the modified HOW can spoof the east and north position of GPS receiver linearly, but spoof the up direction nonlinearly. The handover word in navigation message is unpredictable for the receiver, a small number modified value of HOW (±30 s) is not easily detected by the receiver and this method can spoof the receiver thousands of meters. Therefore, the algorithm has the advantages of strong concealment and high spoofing efficiency. This paper has certain application value to the research of anti-spoofing technology of satellite navigation receiver.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grants 61202394.

References

1. Mark, L.P., Todd, E.H.: Attackers can spoof navigation signals without our knowledge. here's how to fight back GPS lies. IEEE Spectr. **53**, 26–32 (2016)
2. Huang, L., Gong, H., Zhu, X.W., Wang, F.X.: Research of re-radiating spoofing technique to GNSS timing receiver. J. Natl. Univ. Defense Technol. **35**, 93–96 (2013)
3. Zhang, H.S., Gao, G.G., Kou, L.: Deceptive jamming technology of GPS based on the track induction method. J. Projectiles Rockets Missiles Guidance **33**, 149–152 (2013)
4. Tippenhauer, N.O., Popper, C., Rasmussen, K.B., Capkun, S.: On the requirements for successful GPS spoofing attacks. In: 18th ACM Conference on Computer and Communications Security, Chicago, pp. 75–79 (2011)
5. Hu, Y.F., Cao, K.J., Bian, S.F.: GNSS spoofing detection algorithm based on clock frequency drift monitoring. Syst. Eng. Electron. **37**, 1629–1632 (2015)
6. Todd, E.H.: Detection strategy for cryptographic GNSS anti-spoofing. IEEE Trans. Aerosp. Electron. Syst. **49**, 1073–1090 (2013)
7. Mosavi, M.R., Nasrpooya, Z., Moazedi, M.: Advanced anti-spoofing methods in tracking loop. J. Navig. **69**, 883–904 (2016)
8. Magiera, J., Katulski, R.: Applicability of null-steering for spoofing mitigation in civilian GPS. In: Vehicular Technology Conference, pp. 1–5 (2014)
9. Magiera, J., Katulski, R.: Detection and mitigation of GPS spoofing based on antenna array processing. J. Appl. Res. Technol. **13**, 45–57 (2015)



Implementation of Narrowband Interference Suppression Technology for DSSS Systems in Time Domain and Frequency Domain

Xiongfei Li, Xiangyang Meng, and Pan Liu(✉)

Academy of Space Electronic Information Technology,
No. 504, East Chang'an Road, Chang'an District, Xi'an City,
Shaanxi Province, China
lxf_li@163.com, meng.xiangyangl988@163.com,
bumbleliu@hotmail.com

Abstract. Direct sequence spread spectrum (DSSS) system needs some anti-jam algorithms for communication when strong interference exists. Both Normalized least mean square (NLMS) algorithm in time domain and anti-jam algorithm based on FFT in frequency domain are widely used for the excellent performance and small hardware costs. This paper implemented the two algorithms on FPGA and researched their system performances and hardware costs, finally some useful guides for later system design were given.

Keywords: NLMS · FFT · Narrowband interferences suppression
DSSS

1 Introduction

Considering its strong anti-jamming capacity, excellent confidentiality and multiple access convenience, direct sequence spreading (DSS) has been widely applied in anti-interference communication systems. Meanwhile in practice interference cancelling technologies are generally combined with DSS to filter out the intense interference which surpasses the DSS system's interference tolerance of the receiver within its bandwidth.

Varieties of narrow band interference (NBI) suppression algorithms are studied in DSS systems, including time domain adaptive filtering algorithm, transform domain processing and NBI cancelling algorithm in time or frequency domain analysis. For the tradeoff between the suppression effects and complexity, normalized least mean square (NLMS) algorithms in [1–3] based on Wiener filter and frequency domain interference suppression algorithms are commonly adopted.

Standard LMS algorithm realization on DSPs or FPGAs are discussed in [4–6], without detailed module design schemes. The realization of frequency domain NBI suppression algorithms are studied in [7, 8]. [7] focused on the butterfly computation structure of Fast frequency transform (FFT) during hardware realization, while the related FFT IP core is available in FPGAs. [8] introduced a scheme for interference threshold decision. The main scheme is based on a model, and the sum of energies of a

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 54–63, 2018.
https://doi.org/10.1007/978-3-319-78130-3_7

frequency after FFT is a random variable observing the exponential distribution in the model. Interferences can be eliminated after being recognized by detection theory. Because repetitive data processes are required for algorithms in [8], low efficiency can't be avoided.

The NLMS algorithm based on Wiener filtering and the narrow band interference suppression algorithm in transform domain are implemented on the XILINX FPGA chip separately. From practical prospective, the detailed design for each module and the whole source consumption are given in this paper entirely, as well as the comparison of anti-jamming capability between the two algorithms in DSS system with the existence of single-tone interference NBI separately.

2 NBI Suppression in Time Domain

The basic idea of NBI suppression for DSS system in time domain is to deduct the estimated interference from the received signal, and the signal is estimated the interference by taking the advantage of difference between the DSS signal and the NBI signal. DSS signal has the characteristics like flatness in spectrum, low correlation between simples and unpredictability in time domain while the NBI signal shows relatively high correlation from the samples and it's predictable in time domain.

A precise estimate of NBI is achievable by Werner filtering because of the optimum resolution under the linear mean square error (LMSE) criteria. The Werner resolution is often acquired by recursive algorithm due to the involved signal estimation and matric inversion hinder the practical implementation.

The procedure of the NLMS algorithm is as follows:

(a) Calculate the corresponding output of the linear filter:

$$y(n) = w^T(n)x(n). \tag{1}$$

(b) Calculate the error between the output and desired signal:

$$e(n) = y(n) - d(n). \tag{2}$$

(c) Adjust the filter parameters according the error based on the equation below:

$$w(n+1) = w(n) + \frac{u}{E\left\{\|x(n)\|^2\right\}} e(n)\,X(n). \tag{3}$$

Here $x(n)$ is the received signal at time n. $w^T(n)$ is the transpose of the linear filter $w(n)$ and $y(n)$ is the output at the same time. $d(n)$ stands for the desired signal while $e(n)$ corresponds to the error signal. u is the adjusting step and the mean power of the input is shown as $E\left\{\|x(n)\|^2\right\}$.

The structure of NLMS implemented on FPGA is shown in Fig. 1. Four modules including adaptive filter, error calculation and coefficient update and signal power

detection are used to implement the three procedures introduced before. Detailed designs for the four modules on FPGA are given in the last paper.

- Adaptive filter
  The adaptive filter design is based on central tapped delay line filter (transversal filter). Interference suppression is accomplished by adjusting the coefficients of the filter. Werner resolution attained from such a kind of filter is conjugate symmetric and has less source consumption compared with ordinary filters in the same order. Meanwhile the multiplier number can be further reduced by introducing the time division multiplexing in computation.

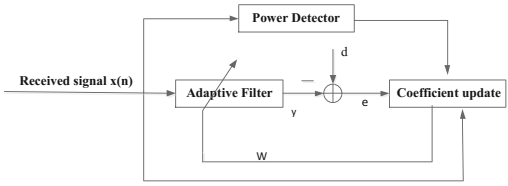


Fig. 1. The structure of NLMS implemented on FPGA

In practice, one chip is sampled more than once, and strong correlations are among those samples of one chip. The detected interference will be mixed with massive DSS signals if all samples are directly applied to estimate and predict the interference. Signal damage is inevitable when interference mixed with DSS signal is applied to interference suppression. The higher the signal to noise ratio is, the damage of the server is stronger. A simple way to decrease the damage is to delay T samples among adjacent taps when the sample rate is T times as the chip rate. The structure of 12th order central tapped delay line filter is shown as Fig. 2.

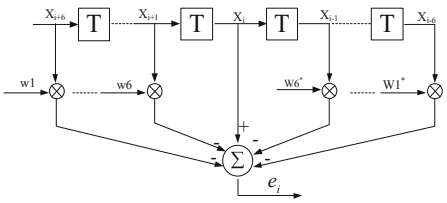


Fig. 2. Central tapped delay line filter

According to the structure, Eq. 1 is easily rewritten as follow:

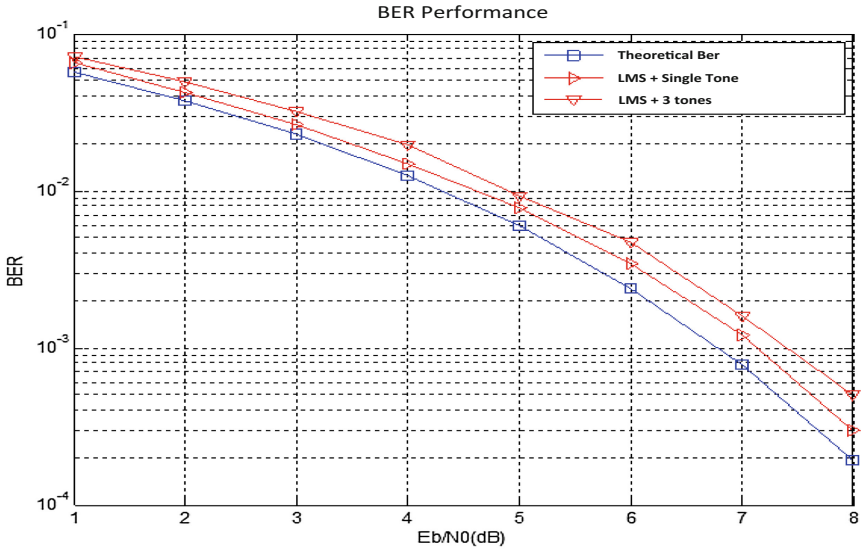
$$\begin{aligned} &conj(\vec{w}).*\vec{x}.' \\ &= (w_{r1} - j*w_{i1})(x_{r1} + j*x_{i1}) + \cdots + (w_{r1} + j*w_{i1})(x_{r12} + j*x_{i12}) \\ &= w_{r1}(x_{r1} + x_{r12}) + w_{i1}(x_{i1} - x_{i12}) + \cdots + j*[w_{r1}(x_{i1} + x_{i12}) + w_{i1}(x_{r12} - x_{r1})] \end{aligned} \quad (4)$$

- Power detection
  The main function of this module is to calculate the averaged power of input, and export the quantized reciprocal of power, i.e. $1\Big/E\left\{\|x(n)\|^2\right\}$ as its output. The implementation on FPGA is to pre-store the quantized reciprocal of power in a ROM, and access the ROM with an address of the average absolute value of the input $|x(n)|_{ave}$. Simplified way of calculating the absolute value of $x(n) = I + jQ$ is elaborated in [9], and adopted in this paper.
- Coefficient update module
  Only the first half of the coefficients needs to be updated each time due to the conjugate symmetry characteristic of the coefficients themselves. The coefficients of the filter are converged to Werner resolution with lower resource consumption and faster convergence speed by this updating way. The update step u is chosen as $1/2^N$ for convenience.

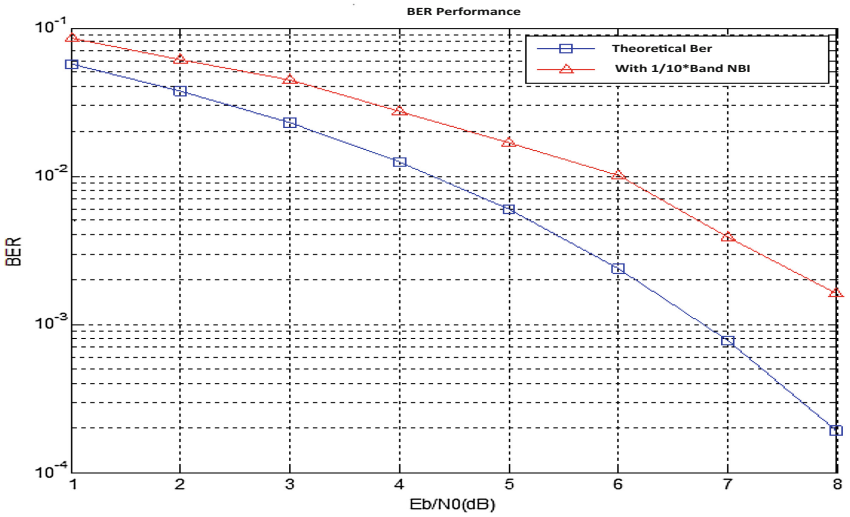
2.1 Simulation and Results Analysis

The design of NLMS algorithm is achieved on the FPGA from Xilinx based on the implementation introduction above and simulation results are also given below.

Figure 3 shows the NBI suppression performance of NLMS. The interference-to-signal ratio (ISR) is set as 40 dB, and the sample rate to chip rate ratio is 4 in simulation. The suppression performances to different interferences are given separately. The single-tone and multi-tone interference suppression performance is shown in Fig. 3(a), a 0.3 dB performance loss can be attained at the bit error rate (BER) of 10^{-3} with existing the single-tone interference, while the performance loss reaches 0.6 dB with multi-tone interference (consists of 5 single-tone interference). The NBI suppression performance is given in Fig. 3(b), the performance loss is 1.5 dB at the same bit error rate when the interference bandwidth is one tenth of the signal bandwidth. The noise only curves in both pictures are the theoretical BER performance versus SNR without interference.



(a) NLMS performance with NBI interference



(b) NLMS performance with NBI interference

Fig. 3. NLMS performance in different interferences

Table 1 below exhibits the resource consumption of NLMS algorithm on a FPGA from Xilinx.

Table 1. The resource consumption of NLMS algorithm

	Used	Available	Utilization
Slices	1970	24576	8%
RAM	4	320	1%
DSP48s	36	512	7%

3 NBI Suppression in Frequency Domain (FFT-Based NBIS)

The significant difference between the spectrum of NBI and the DSS signal is that the power spectrum density (PSD) of the NBI has an obvious peak in the frequency band and the interference falls in when the PSD of the DSS signal is relatively flat in its frequency band. By setting a threshold to minimize the interference in frequency domain, the interference can be efficiently suppressed after restoring signal from frequency domain through inverse fast Fourier transform (IFFT).

The flow diagram of NBI suppression in frequency domain is shown in Fig. 4. Data block window is applied to avoid spectrum leakage. An effective way is deployed by delaying a data block a half of its duration time and constructing a same NBI suppression branch to reduce the energy loss of the data block caused by the data block window. Both FFT and IFFT are realized by the IP cores on the FPGA. Only one FFT and one IFFT IP core is required by using high-rate clock and deploying time-division multiplexing introduced above, and the resource consumption can be lower.

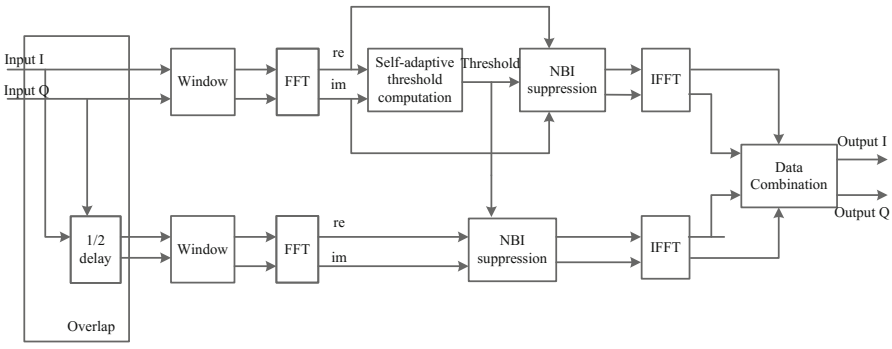


Fig. 4. FFT-based NBIS structure

- Data block window
  Effective side lobe suppression and interference energy leakage decrease are achievable by adding a data block window in time domain before FFT. The width

of the main lobe, the peak of the side lobe and the process loss are the key factors taken into consideration to determine the best window when all available windows are in the same order. In this paper, a hamming window is applied after comparison. It is shown as below.

$$w(n) = 0.54 - 0.46\cos\left(2\pi\frac{n}{N}\right), 0 \le n \le N. \tag{5}$$

- Adaptive threshold algorithm
  The threshold is essential to the performance of interference suppression. Due to its simple and real-time process with considerable performance, adaptive threshold filter algorithm is adopted in this paper. The procedure of recursive threshold estimate is as follows:

$$H(t) = 2\mu(t-1) \tag{6}$$

$$N(k) = \begin{cases} |X(k)|, 0 < |X(k)| \le H(t) \\ 0, \text{others} \end{cases} \tag{7}$$

$$S(k) = \begin{cases} 1, 0 < |X(k)| \le H(t) \\ 0, \text{others} \end{cases} \tag{8}$$

$$\mu(t) = A\mu(t-1) + B\sum_{k=0}^{N-1} N(k) / \sum_{k=0}^{N-1} S(k) \tag{9}$$

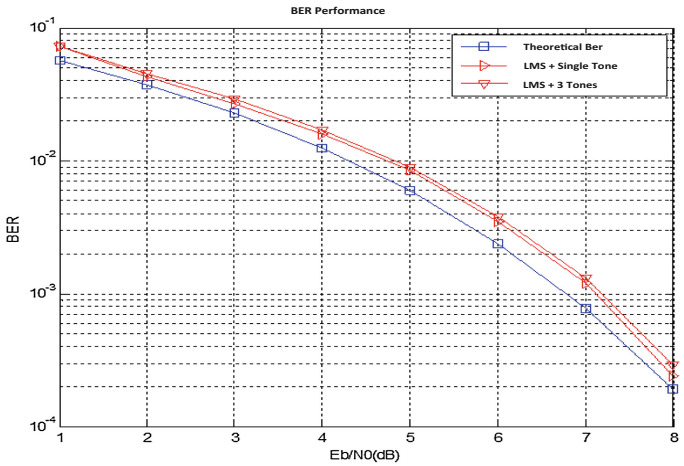
Here $\mu(t-1)$ is the mean amplitude of the $(t-1)^{\text{th}}$ data block, i.e.,$\frac{1}{N}\sum_{k=0}^{N-1}|X(k)|$,while the threshold for the $(t)^{\text{th}}$ data block is $H(t)$, A and B are two recursive coefficients.

- Interference suppression
  Segment compression is applied to suppress the interference here, the detailed method is shown in Eq. 10, where $\overline{X}(k)$ is the frequency component after suppression. The interference amplitude can be clipped by a number which is the power of 2. In practice, it is easily to realize by right shifting the quantized amplitude in binary form without using any complicated multipliers.

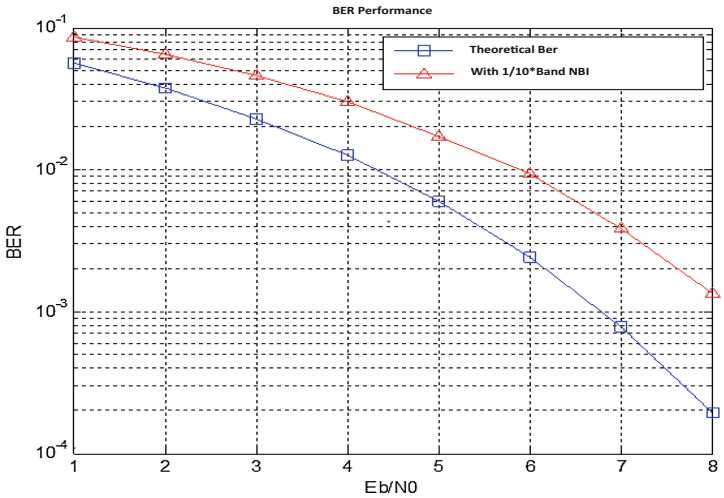
$$\overline{X}(k) = \begin{cases} X(k) & 0 \le |X(k)| \le H(t) \\ \frac{1}{8}X(k) & H(t) \le |X(k)| \le 2H(t) \\ \vdots & \vdots \end{cases} \tag{10}$$

3.1 Simulation and Result Analysis

The NBI interference suppression in frequency domain introduced in Sect. 3 is implemented on a FPGA from Xilinx, and related simulation results under the same condition as Sect. 2.3 are shown in Fig. 5. From the results in Fig. 5(a), a 0.5 dB performance loss can be attained at the bit error rate (BER) of 10^{-3} when existing either the single-tone interference or the multi-tone interference. Figure 5(b) shows the



(a) FFT-based NBIS performance with tones-interference



(b) FFT-based NBIS performance with NBI interference

Fig. 5. FFT-based NBIS performance in different interferences

performance loss is about 1.5 dB at the bit error rate (BER) of 10^{-3} with NSI occupying one tenth bandwidth of the DSS signal.

Table 2 exhibits the resource consumption of NBI suppression in frequency domain on the same FPGA as the Sect. 2.4.

Table 2. Detailed resource consumption

	Used	Available	Utilization
Slices	3952	24576	16%
RAM	25	320	8%
DSP48s	20	512	4%

4 Conclusion

The NLMS NBI suppression algorithm based on adaptive filter in time domain and the NBI suppression based on FFT in frequency domain are both implemented on the same FPGA from XILINX. Detailed module design, the performance under different interferences and resource consumption are given in this paper.

The NLMS algorithm suffers extra 0.3 dB performance loss when there is multi-tone interference, as well as from Fig. 3(a) and Fig. 5(a) the frequency domain NBI suppression is not sensitive to the multi-tone interference. A close performance can be achieved by both algorithms with encountering the NBI, the performance loss is about 1.5 dB at the BER of 10^{-3}.

By comparing Table 1 with Table 2, it can be concluded that NLMS algorithm consumes more multiplier resource, it is about twice as much as the frequency domain NBI suppression algorithm requires, and the required slice of NLMS algorithm is only about a half of the frequency domain NBI suppression algorithm needs. In practical scenarios of the DSS receiver, slice resource is the factor which is more weighted, so NLMS is more desirable in this way.

We suffer a close performance loss under the condition of single-tone interference and NBI by using NLMS algorithm and frequency NBI suppression algorithm. However a recursive time is important to NLMS algorithm which makes it unfit for the burst interference. Proper algorithms should be chosen based on the real application scenarios.

Acknowledgments. This paper is supported by the Pre-research Fund of General Equipment Department (No. 9140A21060115HT05001).

References

1. Feintuch, P.L.: An adaptive recursive LMS filter. Proc. IEEE **64**(11), 1622–1624 (1976)
2. Chen, X., Guo, W., Zhen, Y.: Frequency domain interference suppression in a DSSS system. In: IEEE International Conference on Circuits and Systems and West Sino Expositions, pp. 247–251 (2002)

3. Di Pietro, R.C.: An FFT based technique for suppressing narrow-band interference in PN spread spectrum communications systems. In: 1989 International Conference on Acoustics, Speech, and Signal Processing, ICASSP-89, pp. 1360–1363. IEEE, May 1989
4. Yi, Y., Woods, R., Ting, L.K., Cowan, C.F.N.: High speed FPGA-based implementations of delayed-LMS filters. J. VLSI Sig. Process. **39**(1), 113–131 (2005)
5. Ting, L.K., Woods, R., Cowan, C.F.: Virtex FPGA implementation of a pipelined adaptive LMS predictor for electronic support measures receivers. IEEE Trans. Very Large Scale Integr. Syst. **13**(1), 86–95 (2005)
6. Allred, D., Krishnan, V., Huang, W., Anderson, D.: Implementation of an LMS adaptive filter on an FPGA employing multiplexed multiplier architecture. In: 2004 Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, vol. 1, pp. 918–921. IEEE, November 2003
7. Kumpumaki, T.J., Isohookana, M.A., Juntti, J.K.: Narrow-band interference rejection using transform domain signal processing in a hybrid DS/FH spread-spectrum system. In: Proceedings of MILCOM 1997, vol. 1, pp. 89–93. IEEE, November 1997
8. Zhang, C.H., Lu, S.J., Zhang, E.Y.: Windowed DFT-based narrow-band interference suppression for DSSS communication system. J. PLA Univ. Sci. Technol. (Nat. Sci.) **4**, p. 002
9. Gang, X.: The theory of GPS and receiver design. Beijing Electronic Industry Press (2009) 谢钢. GPS原理与接收机设计. 北京电子工业出版社 (2009).



Spinal Codes Based Efficient Rateless Transmission Scheme for Massive MIMO System with Imperfect CSI

Liangliang Wang[1], Yin Sun[2], Xuming Lu[1], Xiang Chen[1,3](✉), and Hongzhou Tan[1,3]

[1] School of Electronic and Information, Sun Yat-sen University, Guangzhou, China
chenxiang@mail.sysu.edu.cn
[2] Department of ECE, The Ohio State University, Columbus, OH, USA
[3] Joint SYSU-CMU Shunde International Joint Research Institute, Shunde, China

Abstract. The potential performance gains promised by massive multi-input and multi-output (MIMO) rely heavily on the access to accurate channel state information (CSI), which are difficult to obtain in practice when channel coherence time is short and the number of mobile users is huge. Therefore, a critical question is how to make the system perform well with imperfect CSI. Rateless codes with adaptive code-rates can free the system from the accurate CSI requirements and guarantee it to approach the maximum achievable rate (MAR) as well as improve the achieved-rate over that based on the fixed-rate codes. In this paper, a recently proposed family of rateless codes, called spinal codes, will be involved in massive MIMO transmission scheme, which in general approaches to the MAR with sufficiently large encoding block-size. In addition, multi-level puncturing and dynamic block-size allocation (MPDBA) techniques are proposed, where the block-sizes are determined by user MAR to curb the average retransmission delay for successfully decoding the messages, which further enhancing the system retransmission efficiency. Multi-level puncturing, which is MAR dependent, limits the gap between the system MAR and the related achieved-rate. Theoretical analysis are provided to demonstrate that spinal codes with MPDBA can guarantee the system retransmission efficiency as well as achieved-rate. Numerical simulation results are presented to demonstrate these benefits.

Keywords: Massive MIMO · Imperfect CSI · Rateless codes
Spinal codes · Multi-level puncturing · Block-size optimization

1 Introduction

Massive multi-input and multi-output (MIMO), achieving high spectral efficiency and low power consumption, has been widely regarded as an promising technique for 5G wireless communication systems. However, its benefits rely

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 64–74, 2018.
https://doi.org/10.1007/978-3-319-78130-3_8

heavily on the accuracy of the channel state information (CSI) that are available for multi-user procoding. Unfortunately, the collection of accurate CSI is costly because of the short channel coherence time and the huge number of mobile users. Therefore, a critical question is how to improve the throughput performance of massive MIMO systems under imperfect CSI. Traditional fixed-rate codes, such as LDPC [1] and Turbo codes [2], may suffer from the significant throughput loss resulted from the rate-mismatching under inaccurate CSI. Therefore, developing a resilient transmission scheme for massive MIMO in the presence of imperfect CSI is of great importance.

In fact, for multi-user cellular systems, where the base station is not equipped with a large number of antennas, rateless codes have been proved to perform well under inaccurate CSI in [3,4]. Therefore, rateless codes based transmission scheme for massive MIMO is considered in this paper. When CSI are not available at transmitter, rateless codes with adaptive code-rates performs well, some related works about ratless codes, such as spinal codes, strider codes and raptor codes, have been presented in [5–7], respectively. In [5], the authors proved that spinal codes outperform LDPC codes as well as strider and raptor codes in fading channels with inaccurate CSI. Therefore, spinal codes are integrated into resilient transmission scheme, where spinal encoders encode users' original messages into multiple infinite streams of symbols, which are transmitted continuously through numbers of (re)transmissions, called passes, until the senders receive the acknowledge (ACK) indicating the successful decoding, from the receivers. Thus, the system will benefit from the multi-retransmission diversity and achieves a robust performance. Once the encoding block-sizes for spinal codes are allowed to be sufficiently large, users will gradually approach their MAR with maximum-likelihood (ML) decoding algorithm [8].

Zero forcing (ZF) precoding technique is considered in our work to mitigate multiuser interferences. When the number of antennas at the base station is huge, non-linear precoding techniques, such as dirty-paper-coding (DPC) [9], vector perturbation (VP) [10] and lattice-aided [11], have limited throughput gain over ZF [12]. Therefore, under imperfect CSI, ZF and spinal codes based scheme for massive MIMO can achieve the same or even better performance than that based on the non-linear precoding methods with fixed-rate codes as well as reduce the precoding implementation complexity.

However, in order to guarantee the system retransmission efficiency (measured in bits per symbol per second), reducing the retransmission delay, which is equivalent to reduce the pass number, should be considered before spinal codes are integrated into massive MIMO. From [8], it is known that the pass number is proportional to the block-size while inversely proportional to the maximum achievable rate (MAR). Therefore, once we initialize the encoders for all users with a large enough static block-size for MAR-approaching purpose, the users with lower MARs, locating in the cell-edge or trapping in a deep fading channel, will significantly enlarge the average pass number for decoding, which further degrades the system retransmission efficiency performance. Because of this, we developed a multi-level puncturing and dynamic block-size allocation (MPDBA)

scheme, where the MARs are obtained as priori knowledge to determine the block-sizes dynamically, which can reduce the pass number as well as retransmission delay. Different puncturing method is implemented for different MARs, which is proved to guarantee the system retransmission efficiency and achieved-rate. The numerical simulation results show that spinal codes with MPDBA can make massive MIMO with imperfect CSI work reliably and retransmission efficiently.

The rest parts of this paper are organized as follows. Section 2 presents the system model. The efficient transmission scheme for the system with spinal codes is proposed in Sect. 3. Section 4 shows some simulation results to illustrate the benefits of proposed scheme, then conclusions are drawn in Sect. 5.

2 System Model

Throughout this paper, upper case boldface letters are used to denote matrices. $(\bullet)^H$ and $(\bullet)^T$ represent the Hermitian transpose and transpose, respectively. $E[\bullet]$ denotes the expectation operator. Tr$(\bullet)$ stands for trace and $\lceil \bullet \rceil$ is the round up to integer operator.

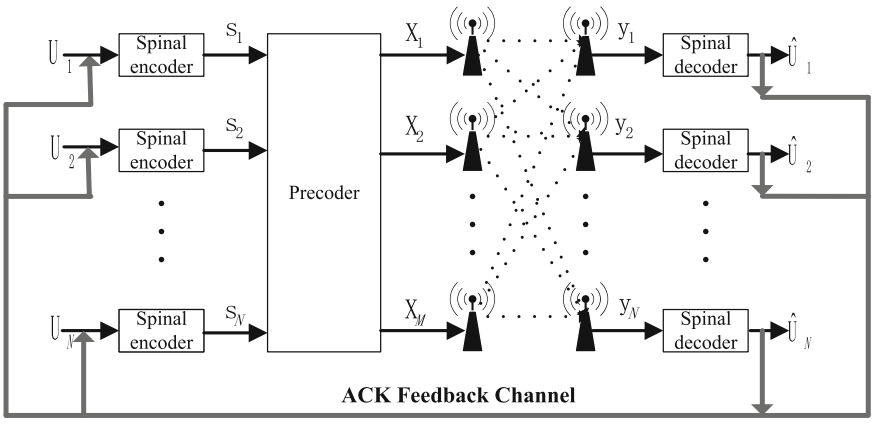


Fig. 1. Spinal codes based massive MIMO systems.

A $N \times M$ downlink massive MIMO system is shown in Fig. 1, where the single-cell system has a base station with M antennas and N single-antenna users, $M \gg N$. For user n, $n = 1, 2, \ldots, N$, spinal encoder [5], illustrated in Fig. 2, divides U_n, denoted by message I and $I = 1, 2, \ldots$, into B blocks with size $K_n(I)$ bits, hash function uses each block to generate a sequence of spine values, then the modulated symbols are yielded by a mapper and pseudo random number generator (RNG) function, which uses spine values as its seeds. After encoding, ZF maps the modulated symbols to the precoded symbols, which are transmitted over fading channels. If decodes successfully, the receiver will send

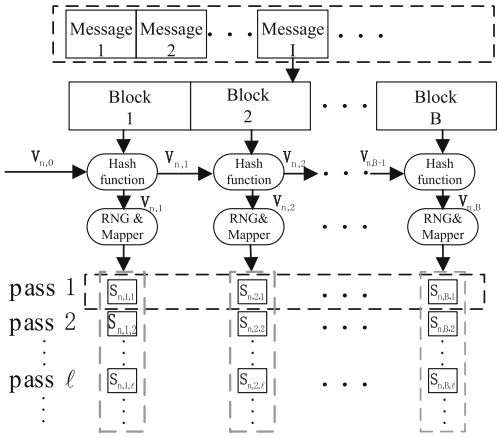


Fig. 2. Structure of spinal encoder.

the ACK to the corresponding sender through the feedback channel, otherwise, the sender will generate more redundant symbols to transmit in the following passes until receiving the ACK.

Given $L_n(I) \geq 1$ is the pass number for user n to successfully decode the message I. At lth pass, for $l = 1, 2, \ldots, L_n(I)$, all users' modulated symbols are mapped to a $N \times B$ matrix of symbols $\mathbf{S}_l = (\mathbf{S}_{1l}, \mathbf{S}_{2l}, \ldots, \mathbf{S}_{Bl})$. $\mathbf{S}_{il} = (S_{1il}, S_{2il}, \ldots, S_{Nil})^T$, for $i = 1, 2, \ldots, B$, follows Gaussian distribution and $\mathrm{Tr}\left(\mathbf{S}_{il}\mathbf{S}_{il}^H\right) = P_l$, P_l is total transmit power. Assuming CSI are prior knowledge obtained by channel estimation methods at transmitter, and downlink MIMO channel remains ergodic and stationary during B symbols periods, $\mathbf{S}_l$ is then precoded to transmit signal $\mathbf{X}_l \in \mathbb{C}^{M \times B}$ by

$$\mathbf{X}_l = \beta_l \hat{\mathbf{W}}_l \mathbf{S}_l, \tag{1}$$

where $\hat{\mathbf{W}}_l = \hat{\mathbf{H}}_l^H \left(\hat{\mathbf{H}}_l \hat{\mathbf{H}}_l^H\right)^{-1}$ denotes the ZF precoding matrix for lth pass, $\hat{\mathbf{H}}_l \in \mathbb{C}^{N \times M}$ denotes the estimated channel matrix, and $\beta_l = \sqrt{N/Tr\left(\hat{\mathbf{W}}_l \hat{\mathbf{W}}_l^H\right)}$ is power control factor used to normalize $\hat{\mathbf{W}}_l$ such that $\mathrm{Tr}\left(\mathbf{X}_{il}\mathbf{X}_{il}^H\right) = P_l$.

$\mathbf{X}_l$ is then transmitted through a MIMO fading channel, the received signal matrix $\mathbf{Y}_l \in \mathbb{C}^{N \times B}$ is given by

$$\mathbf{Y}_l = \beta_l^{-1} \mathbf{H}_l \mathbf{X}_l + \beta_l^{-1} \mathbf{N}_l, \tag{2}$$

where $\mathbf{H}_l = \mathbf{D}_l^{1/2} \mathbf{G}_l$ denotes the MIMO channel matrix, $\mathbf{G}_l \in \mathbb{C}^{N \times M}$ and $\mathbf{D}_l \in \mathbb{R}^{N \times N}$ represent the complex small-scale and large-scale fading coefficients matrix, respectively. Assuming that $\mathbf{G}_l$ has zero mean and unit variance independent and identically distributed (i.i.d.) complex Gaussian entries, large-scale fading coefficients, denoted as $\mathbf{D}_l = diag\,(d_{1l}, d_{2l}, \ldots, d_{Nl})$, are same for different

antennas but user-dependent. $\mathbf{N}_l \in \mathbb{C}^{N\times B}$ denotes the noise matrix with i.i.d. complex Gaussian random variables with zero mean and unit variance. Then the average transmit SNR is P_l/N.

If perfect CSI are available at transmitter, then $\hat{\mathbf{H}}_l = \mathbf{H}_l$. However, imperfect CSI always arise in any practical estimation schemes. According to the channel estimation model described in [13], $\hat{\mathbf{H}}_l$ can be given by

$$\hat{\mathbf{H}}_l = \mathbf{H}_l + \delta_e \Delta\mathbf{H}_l, \tag{3}$$

where $\Delta\mathbf{H}_l \in \mathbb{C}^{N\times M}$ denotes the channel estimation error matrix with zero mean and unit variance i.i.d. complex Gaussian random variables uncorrelated with that of $\mathbf{H}_l$, δ_e is a non-zero parameter to measure the quality of channel estimation and is appropriately chosen depending on the channel dynamics. Since the focus of this paper is to study the performance of spinal codes, δ_e is assumed to be known at transmitter.

Based on (3), the ZF precoding matrix can be expressed as

$$\hat{\mathbf{W}}_l = \beta_l \left(\mathbf{H}_l + \delta_e\Delta\mathbf{H}_l\right)^H \left[\left(\mathbf{H}_l + \delta_e\Delta\mathbf{H}_l\right)\left(\mathbf{H}_l + \delta_e\Delta\mathbf{H}_l\right)^H\right]^{-1}, \tag{4}$$

substituting (4) into (2), the *ith* symbol vector at *lth* pass is given by

$$\mathbf{Y}_{il} = \mathbf{S}_{il} + \beta_l^{-1}\mathbf{N}_{il} - \delta_e\Delta\mathbf{H}_l\hat{\mathbf{H}}_l^H\left(\hat{\mathbf{H}}_l\hat{\mathbf{H}}_l^H\right)^{-1}\mathbf{S}_{il}, \tag{5}$$

where $\delta_e\Delta\mathbf{H}_l\hat{\mathbf{H}}_l^H\left(\hat{\mathbf{H}}_l\hat{\mathbf{H}}_l^H\right)^{-1}\mathbf{S}_{il}$ is the additional interference item brought by the channel estimation error.

Once $l = L_n(I)$, spinal decoder for user n collects enough mutual information to retrieve the message I successfully [5]. Then the user current data rate (measured in bits per symbol) is achieved by

$$R_n(I) = \frac{K_n(I)}{L_n(I)}, \tag{6}$$

based on (6), the average achieved-rate of spinal codes based massive MIMO is given by $R_{achieved} = \lim\limits_{I\to\infty}\frac{1}{I}\sum_{I=1}^{\infty}\sum_{n=1}^{N}R_n(I)$.

Based on (5), assuming that P_{nl} is the average signal power for the desired symbols $S_{nil} \in \mathbf{S}_{il}$, we have $\sum_{n=1}^{N}P_{nl} = \mathrm{Tr}\left(\mathbf{S}_{il}\mathbf{S}_{il}^H\right) = P_l$. Denoting

$$\hat{\mathbf{N}}_{il} = \beta_l^{-1}\mathbf{N}_{il} - \delta_e\Delta\mathbf{H}_l\hat{\mathbf{H}}_l^H\left(\hat{\mathbf{H}}_l\hat{\mathbf{H}}_l^H\right)^{-1}\mathbf{S}_{il} \tag{7}$$

as the received noise, and the covariance for $\hat{\mathbf{N}}_{il}$ can be proved to be

$$E\left[\hat{\mathbf{N}}_{il}\hat{\mathbf{N}}_{il}^H\right] = \beta_l^{-2}\boldsymbol{I}_N + \boldsymbol{\lambda}_l^{-1}P_l\delta_e^2, \tag{8}$$

$\boldsymbol{\lambda_l} = diag\left(\lambda_{1l},\ldots,\lambda_{Nl}\right)$ is the singular value of $\hat{\mathbf{H}}_l\hat{\mathbf{H}}_l^H$, and is approximated by

$$\boldsymbol{\lambda_l} \approx M\left(\mathbf{D}_l + \delta_e^2\mathbf{I}_N\right). \tag{9}$$

Based on (8) and (9), the average MAR for user n is given by

$$R_n^{upper} = E\left[log_2\left(1 + \frac{M\beta_l^2 P_{nl}}{M + \beta_l^2 P_l \delta_e^2 \left(d_{nl} + \delta_e^2\right)^{-1}}\right)\right]. \tag{10}$$

Thus, the upper bound of ergodic achievable rate for massive MIMO with ZF under imperfect CSI is expressed as $R_{upper} = \sum_{n=1}^{N} R_n^{upper}$.

Lemma 1. *If $K_n(I) \geq R_n^{upper}$, then there exists $(L_1(I), L_2(I), \ldots, L_N(I))$ for users to achieve $R_n^{upper} - R_n(I) = O\left(\frac{(R_n^{upper})^2}{K_n(I)}\right)$.*

The proof of this lemma is similar to that in [8], and hence is omitted here. This indicates that when block-size $K_n(I) = K_{\max}$, where $K_{\max}$ denotes the sufficiently large block-size in this paper, the gap between $R_n(I)$ and R_n^{upper} can be arbitrarily small such that each user can approach R_n^{upper} with $L_n(I)$, and spinal codes based massive MIMO approaches R_{upper} as well.

From [8], $L_n(I)$ can be given by

$$L_n(I) = \left\lceil \frac{K_n(I)}{R_n^{upper}} \right\rceil + O(1). \tag{11}$$

When the small large-scale fading coefficients d_{nl} deteriorate R_n^{upper} in (10), based on (11), $K_n(I) = K_{\max}$ will enlarge $L_n(I)$ significantly. If each pass occupies the constant time T (measured in seconds), then the larger $L_n(I)$ makes the average retransmission delay spent on decoding the message I be costly, and the average system retransmission efficiency, defined as

$$r = \lim_{I \to \infty} \frac{1}{I} \sum_{I=1}^{\infty} \sum_{n=1}^{N} \frac{R_n(I)}{T L_n(I)}, \tag{12}$$

will be limited as well.

The greedy idea is an effective way to achieve a preferable system retransmission efficiency across different SNRs. That is, we try to make sure that $\frac{R_n(I)}{TL_n(I)}$ for each message I is better, which will further achieve a satisfied value for r. When $R_n(I)$ is ideal, to decrease $L_n(I)$ is an effective way to enhance $\frac{R_n(I)}{TL_n(I)}$. Therefore, based on (11), $K_n(I)$ can be dynamically determined by R_n^{upper} to reduce $L_n(I)$.

3 Multi-level Puncturing and Dynamic Block-Size Allocation Scheme

In this section, a multi-level puncturing and dynamic block-size allocation scheme is proposed. For each user, the dynamic block-size allocation problem is formulated as

$$K_n^*(I) = \underset{R_n^{upper} \leq K_n(I) < K_{max}}{\arg\min} L_n(I), \tag{13}$$

where $K_n^*(I)$ is the desired block-size for message I.

However, we can't obtain the solution for (13) by solving (11) directly, which includes an unexpected constant item O(1). Instead, we use the cumulative distribution function (CDF) of pass numbers $L_n(I)$ to determine $K_n^*(I)$. Figure 3 presents the CDF of pass numbers corresponding to the different block-sizes and SNRs, from which we obtain that under different SNRs, smaller block-sizes guarantee the system to decode successfully with less pass number and high probability. Therefore, we choose

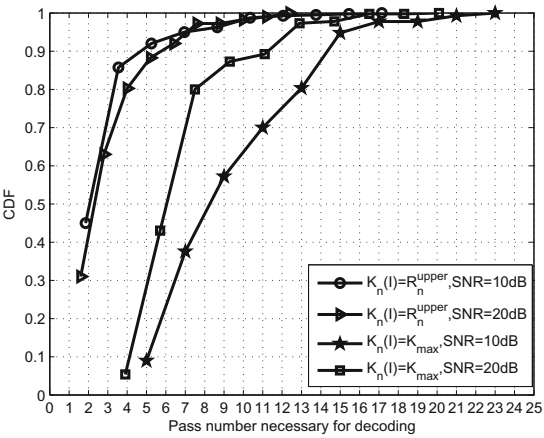


Fig. 3. CDF curves of pass number needed for successfully decoding under different SNRs, $K_{\max} > K_n^{upper}$, $\delta_e = 0.1$.

$$K_n^*(I) = \lceil R_n^{upper} \rceil \tag{14}$$

for each spinal encoder to encoding the message I.

When $K_n(I) = K_n^*(I)$, the gap between the MAR and achieved-rate is given by

$$R_n^{upper} - R_n(I) = \theta\left(R_n^{upper}\right), \tag{15}$$

which indicates that the dynamic block-size scheme will cause an unavoidable achieved-rate performance loss at higher R_n^{upper}. To reduce the loss as much as possible, a multi-level puncturing method is implemented for spinal codes, where the puncturing patten is given in [8]. With proper link-layer scheduling schemes for the punctured codes in massive MIMO, the user n current rate in (6) is rewritten as

$$R_n(I) = \frac{K_n(I) L_{sub,n}(I)}{L_{sub,n}(I) + L_n(I) - 1}, \tag{16}$$

where $L_{sub,n}(I) = 1, 2, \ldots, B$ is the sub-pass number to transmit symbols in a pass.

Theorem 1. *Consider that the symbols of MPDBA based spinal codes with parameter $K_n^*(I)$, P_l are transmitted over a MIMO channel with imperfect CSI, then the system can achieve $R_n^{upper} - R_n(I) = \theta\left(\frac{R_n^{upper}}{L_{sub,n}(I)}\right)$ with $r = \theta\left(\frac{(R_n^{upper})^2}{L_{sub,n}(I)}\right)$.*

This theorem indicates that for the transmission scheme in massive MIMO with MPDBA based spinal codes, we can choose a proper $L_{sub,n}(I)$ (i.e. lower $L_{sub,n}(I)$ for lower R_n^{upper} and higher $L_{sub,n}(I)$ for higher R_n^{upper}) for different R_n^{upper} such that the gap between the achieved-rate and the MAR is limited as well as a better retransmission efficiency is guaranteed under different $L_{sub,n}(I)$.

4 Simulation Results and Analysis

In this section, some numerical simulation results are illustrated to verify the performance of MPDBA based spinal codes scheme for a 4×64 MIMO system, $K_{\max} = 20$.

Figure 4 compares the average achieved-rates of massive MIMO based on spinal codes and LDPC codes, respectively. LDPC codes are from 802.16e, decoded with a powerful decoder. We can obtain that spinal codes outperform LDPC codes across all SNRs. Moreover, the simple decoder structure of spinal codes also avoids the de-mapping complexity caused by different constellations mapping operations of LDPC codes.

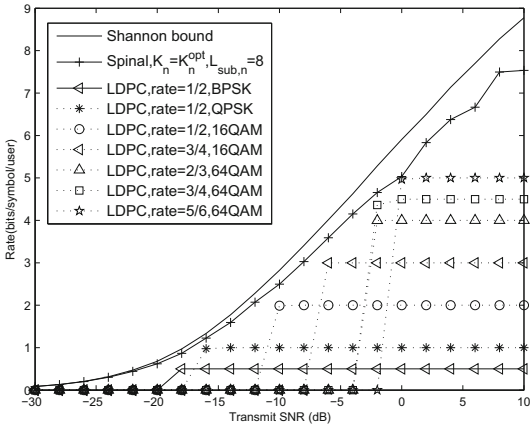


Fig. 4. Average rates achieved by spinal codes and LDPC codes based massive MIMO, $\delta_e = 0.1$.

In Fig. 5, we compare the average achieved-rates of spinal codes based massive MIMO with $K_n(I) = K_{\max}$ and $K_n(I) = K_n^*(I)$, respectively. Massive MIMO can approach its MAR gradually by enlarging the block-size, which verifies the

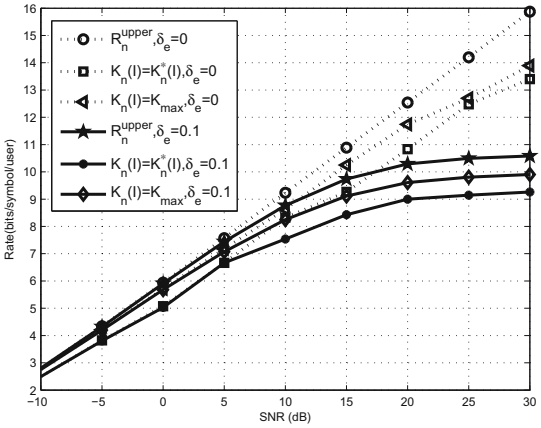


Fig. 5. Average rates achieved by spinal codes based massive MIMO, $K_{\max} > K_n^*$, $L_{sub,n}(I) = 8$.

conclusion in Lemma 1. We also obtain that the achieved-rate performance of spinal codes with $K_n(I) = K_n^*(I)$ follows conclusions in Theorem 1, that is, larger R_n^{upper} will cause bigger achieved-rate performance loss, especially for lower $L_{sub,n}(I)$.

Figure 6 shows that we can enlarge $L_{sub,n}(I)$ to minimize the gap between the MARs and the rates achieved by spinal codes with MPDBA. For lower R_n^{upper}, smaller $L_{sub,n}(I)$ can achieve a satisfying achieved-rates result while a larger $L_{sub,n}(I)$ is necessary for the MAR approaching-purpose for higher R_n^{upper}.

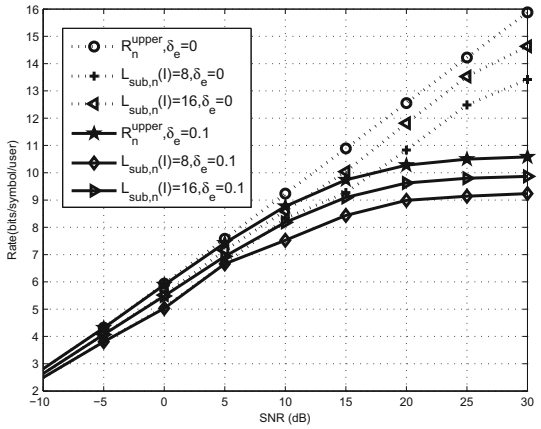


Fig. 6. The effects of $L_{sub,n}(I)$ on average rates achieved by spinal codes based massive MIMO.

Figure 7 shows that the application of MPDBA for spinal encoders will enhance the system retransmission efficiency. Different proper $L_{sub,n}(I)$ for the MPDBA based spinal codes can achieve a preferable retransmission efficiency, which also indicates that once the system achieves a better retransmission efficiency, there will be more proper $L_{sub,n}(I)$ to choose from to obtain better achieved-rates performance illustrated in Fig. 6. The numerical results in Figs. 6 and 7 verify the Theorem 1, that is, spinal codes with MPDBA can guarantee the system reliability as well as retransmission efficiency.

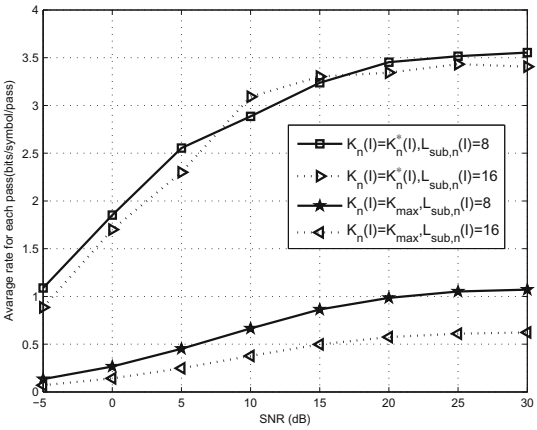


Fig. 7. Average system retransmission efficiency, $K_{\max} > K_n^*(I)$, $\delta_e = 0.1$.

5 Conclusions

The application of spinal codes in massive MIMO with imperfect CSI has great benefits. Before spinal codes are integrated into the system, an efficient transmission scheme for massive MIMO with spinal codes is proposed, where the maximum achievable rate (MAR) is used to determine the block-size dynamically to reduce the retransmission delay spent on decoding the messages, and multi-level puncturing scheme is considered for different MARs to limit the gap between the MAR and achieved-rate. Some theoretical analysis are provided to prove that the multi-level puncturing and dynamic block-size allocation (MPDBA) based spinal codes can guarantee the system achieved-rate as well as retransmission efficiency. Numerical simulation results illustrate that spinal codes with MPDBA can improve the achieved-rate efficiently over that based on the fixed-rate codes as well as enhance the system retransmission efficiency with limited achieved-rate performance degradation. Therefore, the proposed method can guarantee spinal codes based massive MIMO system to be reliable and practical.

References

1. Gallager, R.G.: Low-Density Parity-Check Codes. Springer International Publishing, Cham (1960)
2. Berrou, C., Glavieux, A., Thitimajshima, P.: Near shannon limit errorcorrecting coding and decoding: turbo-codes. 1. In: IEEE International Conference on Communications, vol. 2, pp. 1064–1070 (1993)
3. Sun, Y., Koksal, C.E., Kim, K.H., Shroff, N.B.: Scheduling of multicast and unicast services under limited feedback by using rateless codes. In: IEEE INFOCOM 2014, Toronto, pp. 1671–1679 (2014)
4. Sun, Y., Koksal, C.E., Lee, S.J., Shroff, N.B.: Network control without CSI using rateless codes for downlink cellular systems. In: IEEE INFOCOM 2013, pp. 1016–1024 (2013)
5. Perry, J., Iannucci, P.A., Fleming, K.E., Balakrishnan, H., Shah, D.: Spinal codes. ACM SIGCOMM Comput. Commun. Rev. **42**(42), 49–60 (2012)
6. Gudipati, A., Katti, S.: Strider: automatic rate adaptation and collision handling. ACM SIGCOMM Comput. Commun. Rev. **41**(4), 158–169 (2011)
7. Etesami, O., Shokrollahi, A.: Raptor codes on binary memoryless symmetric channels. IEEE Trans. Inf. Theory **52**(5), 2033–2051 (2006)
8. Balakrishnan, H., Iannucci, P., Perry, J., Shah, D.: De-randomizing shannon: the design and analysis of a capacity-achieving rateless code. arXiv preprint arXiv:1206.0418 (2012)
9. Costa, M.: Writing on dirty paper. IEEE Trans. Inf. Theory **29**(3), 439–441 (1983)
10. Hochwald, B.M., Peel, C.B., Swindlehurst, A.L.: A vector-perturbation technique for near-capacity multiantenna multiuser communication-Part II: perturbation. IEEE Trans. Commun. **53**(3), 537–544 (2005)
11. Windpassinger, C., Fischer, R.F.H., Huber, J.B.: Lattice-reduction-aided broadcast precoding. IEEE Trans. Commun. **52**(12), 2057–2060 (2004)
12. Rusek, F., Persson, D., Lau, B.K., Larsson, E.G., Marzetta, T.L., Edfors, O., Tufvessonet, F.: Scaling up MIMO: opportunities and challenges with very large arrays. IEEE Sig. Process. Mag. **30**(1), 40–60 (2012)
13. Wang, C., Au, E.K.S., Murch, R.D., Mow, W.H., Cheng, R.S., Lau, V.: On the performance of the mimo zero-forcing receiver in the presence of channel estimation error. IEEE Trans. Wirel. Commun. **6**(3), 805–810 (2007)



Low-Cost Topology Design for Wireless Ad Hoc Networks with Connectivity Constraint

Mengmeng Xu[1,2(✉)], Qinghai Yang[2], and Hengzhou Xu[1]

[1] School of Network Engineering, Zhoukou Normal University, Zhoukou 466001, China
mmxu@zknu.edu.cn

[2] School of Telecommunications Engineering, Xidian University, Xi'an 710071, China

Abstract. In this paper, we investigate the topology design to balance the network cost and connectivity, which are two important metrics to guide the design of network topology in ad hoc networks. To this end, we focus on how to construct the network topology, which has the minimum network cost subject to a connectivity constraint. A new topology metric for each link, called connectivity-cost-ratio, is defined as the ratio of the connectivity contribution to the link cost. Then, a heuristic algorithm, low-cost topology design with connectivity constraint (LCTDCC), is proposed to switch off the links, which have small value of connectivity-cost-ratio, i.e., large cost but small contributions on connectivity. Some cost functions are also devised for various network applications. Simulation results validate the efficiency of our proposed algorithm.

Keywords: Wireless ad hoc network · Topology design
Network cost · Algebraic connectivity

1 Introduction

Future wireless ad hoc networks are envisioned to support a wide variety of applications, which have different requirements on network configuration. Network cost (such as energy, bandwidth) is the most crucial factor for network topology design, particularly in the rigorous environments where the network resource is limited. In general, a topology with more redundant links means it suffers more network cost. On the other hand, many applications, such as search and rescue, target tracking, desire a robust network topology, which has the ability of withstanding a sudden link failure. Obviously, for a topology with small number of maintainable links, the network cost is low, however, the robustness (or connectivity) is bad. Therefore, low-cost topology design with connectivity constraint is an interesting problem due to the conflicting objectives.

Various topology design approaches have been proposed to construct minimally connected topologies, in which the minimum possible energy is consumed,

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 75–83, 2018.
https://doi.org/10.1007/978-3-319-78130-3_9

such as localized minimum spanning tree [1,2], and some game-based topology control algorithms [3–6]. In [7,8], the authors defined some new link weights for topology design to improve the network's lifetime. Topology design methods developed to minimize the network cost (link price, congestion) were investigated in [9,10]. However, these approaches have overmuch emphasis on network cost, but neglect the importance of network connectivity. Algebraic connectivity, which defined as the second smallest eigenvalue of the graph Laplacian from spectral graph theory, has been proved to demonstrate the robustness of network connectivity [11]. For this reason, it was used in [12] for robust routing design, in [13] for constructing reliable backbone network, and in [14] to keep the connectivity and save the energy for the Internet.

The focus of our work is to design a network topology, which has the minimum network cost subject to a connectivity constraint. A new topology metric, called connectivity-cost-ratio, is defined, which is the function of algebraic connectivity and link cost. This metric allows us to switch off the links which have large cost but small contributions on network connectivity. The low-cost topology design with connectivity constraint (LCTDCC) algorithm is proposed to reduce the network cost and keep the network connectivity. We also devise some cost functions for various network applications.

The rest of the paper is organized as follows. Section 2 presents the system model and the problem statement for topology design. We define a new topology metric of connectivity-cost-ratio and propose the LCTDCC algorithm in Sect. 3. Some cost functions are devised in Sect. 4. We evaluate the performance of our algorithm through computer simulations in Sect. 5, and conclude the paper in Sect. 6.

2 System Model and Problem Statement

2.1 System Model

In a network $\mathcal{G}(\mathcal{V},\mathcal{E})$, a set of n nodes are uniformly distributed in a planar field, where $\mathcal{V}=\{1,2,\cdots,n\}$ denotes the set of nodes and $\mathcal{E}$ is the set of links. Inevitably, the link maintenance between two nodes i and j incurs cost, such as power/energy, bandwidth and delay. Let each wireless link e_{ij} have a cost c_{ij} of constituting a network topology. The *network cost* C in the network $\mathcal{G}$ is defined as the sum of all link cost, i.e., $C=\sum_{e_{ij}\in\mathcal{E}} c_{ij}$. We will discuss several possible cost functions for different applications in Sect. 4. In general, smaller c_{ij} means that the link e_{ij} is more suitable of being in the topology.

In this work, algebraic connectivity defined in the following is exploited as the metric of network connectivity instead of the conventional measures – node/edge connectivity. Firstly, we denote the adjacency matrix of $\mathcal{G}$ with $A=(a_{ij})$. The matrix A is a symmetric $n\times n$ binary matrix, whose entry a_{ij} is either one or zero, depending on whether the link between nodes i and j is "on" or "off". Let $\mathcal{N}_i$ be the set of nodes that can directly communicate with node i. We define the degree matrix of $\mathcal{G}$ as $D=\mathrm{diag}(\{|\mathcal{N}_1|,|\mathcal{N}_2|,\cdots,|\mathcal{N}_n|\})$. Obviously, we have $|\mathcal{N}_i|=\sum_{j=1}^{n} a_{ij}$. The Laplacian matrix of network graph $\mathcal{G}$ is then defined as the difference between D and A, i.e., $L=D-A$. Let the eigenvalues of L

denote by $\mu_i(\mathcal{G}), i = 1, \cdots, n$, ordered in an increasing sense. As the matrix L is a symmetric positive semi-definite matrix, its smallest eigenvalue is zero, i.e., $\mu_1(\mathcal{G}) = 0$. The quantity $\mu_2(\mathcal{G})$ is known as the *algebraic connectivity* of $\mathcal{G}$. To demonstrate that $\mu_2(\mathcal{G})$ is a good measure for network connectivity, we summarize several important properties as follows [11]:

Property 1. The value of algebraic connectivity $\mu_2(\mathcal{G})$ is strictly greater than 0 if and only if $\mathcal{G}$ is a connected graph.

Property 2. For $\mathcal{G}_1(\mathcal{V}, \mathcal{E}_1)$ and $\mathcal{G}_2(\mathcal{V}, \mathcal{E}_2)$, if $\mathcal{G}_1$ is obtained from $\mathcal{G}_2$ by deleting some links, we have $\mu_2(\mathcal{G}_1) \leq \mu_2(\mathcal{G}_2)$.

Property 3. The algebraic connectivity $\mu_2(\mathcal{G})$ satisfies the following inequality

$$\mu_2(\mathcal{G}) \leq v(\mathcal{G}) \leq \delta(\mathcal{G}) \leq \min_i |\mathcal{N}_i|, \tag{1}$$

and

$$\mu_2(\mathcal{G}) \geq 2\delta(\mathcal{G})(1 - \cos(\pi/n)), \tag{2}$$

where $v(\mathcal{G})$ and $\delta(\mathcal{G})$ are node connectivity and edge connectivity of the network, respectively.

According to the above properties, if a graph adds more links, its algebraic connectivity $\mu_2(\mathcal{G})$ would increase. Let $\mathcal{G}^{-e}$ denote the network graph that is achieved from $\mathcal{G}$ by deleting the link $e \in \mathcal{E}$. Clearly, we have the variation $\Theta_e := \mu_2(\mathcal{G}) - \mu_2(\mathcal{G}^{-e})$ is great than zero. This implies that Θ_e measures the *connectivity contribution* of the link e in graph $\mathcal{G}$. Additionally, the algebraic connectivity has a continuous value and a close relationship with the conventional connectivity metrics. In conclusion, the algebraic connectivity could serve as a good metric to measure the network connectivity.

2.2 Problem Statement

The objective of our work is to construct the network topology, which has the minimum network cost subject to a connectivity constraint. That is, for an initial topology graph $\mathcal{G}_{ini}(\mathcal{V}, \mathcal{E}_{ini})$, find a sub-graph $\mathcal{G}_{obj}(\mathcal{V}, \mathcal{E}_{obj})$ such that the topology $\mathcal{G}_{obj}$ has the minimum network cost (i.e., $\min \sum_{e_{ij} \in \mathcal{E}_{obj}} c_{ij}$) but its algebraic connectivity is greater than a_{th} (i.e., $\mu_2(\mathcal{G}_{obj}) \geq \mu_{th}$). To this end, a new topology metric for each link is firstly defined, called connectivity-cost-ratio, to measure the connectivity contribution per unit cost. Then, the LCTDCC (Low-Cost Topology Design with Connectivity Constraint) heuristic algorithm is proposed and analyzed.

3 LCTDCC Algorithm

As we know, removing the links with large cost is a basic way to reduce the network cost. But the blind removals of these links do not help to construct the low-cost topology under the given connectivity constraint. Actually, choosing

the links which have small connectivity contributions of could allow more link removals. Therefore, a link, which incurs large cost, but only contributes a small value of connectivity, should be deleted firstly. To capture this view, a new topology metric for each link e_{ij} called *connectivity-cost-ratio* is defined as

$$\eta(\mathcal{G}, e_{ij}) = \frac{\Theta_{e_{ij}}^{\beta}}{c_{ij}}, \tag{3}$$

where $\Theta_{e_{ij}} = \mu_2(\mathcal{G}) - \mu_2(\mathcal{G}^{-e_{ij}})$ is the connectivity contribution of link $e_{ij} \in \mathcal{E}$ in topology $\mathcal{G}$. The exponent β in (3) is a control factor.

In Algorithm 1, we shown the procedure of the proposed LCTDCC heuristic algorithm. The link removals are proceeded in an iterative manner. The inner **For** loop from line 3 to 10 in the algorithm associates to each link the value of connectivity-cost-ratio according to (3). If the removal of one link e_{ij} results in the connectivity below the given constraint μ_{th}, we set the connectivity-cost-ratio of this link as infinity, i.e., this link could not be switched off. The link with the minimum value of connectivity-cost-ratio is chosen in line 11 and then removed from line 12 to 16. The algorithm is terminated if no link could be selected to switch off.

LCTDCC algorithm is an innovative approach that jointly reduces the network cost in the wireless network and keeps the network connectivity. With the removals of links one by one, the network connectivity will be dropped from $\mu_2(\mathcal{G}_{ini})$ to 0 gradually. Since the algorithm in each iteration selects a unique link

Algorithm 1. LCTDCC heuristic Algorithm

Input: $\mathcal{G}_{ini}$; c_{ij}; μ_{th};
Output: $\mathcal{G}_{obj}$;
1: $\mathcal{G}(\mathcal{V}, \mathcal{E}) = \mathcal{G}_{ini}(\mathcal{V}, \mathcal{E}_{ini})$;
2: **for** $m = 1$ to $|\mathcal{E}_{ini}|$ **do**
3: **for** each link $e_{ij} \in \mathcal{E}$ **do**
4: evaluate $\mu_2(\mathcal{G}^{-e_{ij}})$;
5: **if** $\mu_2(\mathcal{G}^{-e_{ij}}) \geq \mu_{th}$ **then**
6: compute the connectivity-cost-ratio $\eta(\mathcal{G}, e_{ij})$;
7: **else**
8: $\eta(\mathcal{G}, e_{ij}) = \infty$;
9: **end if**
10: **end for**
11: choose the link e_{wv} which has the minimum value of connectivity-cost-ratio;
12: **if** $\eta(\mathcal{G}, e_{wv}) \neq \infty$ **then**
13: $\mathcal{E} = \mathcal{E} - e_{wv}$;
14: **else**
15: **break**;
16: **end if**
17: $m = m + 1$;
18: **end for**
19: **return** $\mathcal{G}(\mathcal{V}, \mathcal{E})$

to delete, our heuristic algorithm could converge to a unique stable topology with the connectivity constraint μ_{th}. As the algebraic connectivity having a computational complexity equal to $O(n^2)$, the overall algorithm complexity is equal to $O(|\mathcal{E}_{ini}|^2 \cdot n^2)$. The network topology derived by our LCTDCC algorithm has the ability of withstanding sudden link failures due to the connectivity constraint, which could meet the reliability requirements of some actual applications.

4 Discussion of Cost Function

Our heuristic algorithm can be used in ad hoc networks to derive a cost-minimized topology with connectivity constraint. The cost function in this work is general, which could be defined according to various practical applications. In the following, we will discuss some possible cost functions that maybe used in ad hoc networks.

Power Consumption. Most of the topology design approaches have been proposed to construct energy-efficiently connected topologies. Two nodes i and j could communicate with each other, if both of their transmit power p_i and p_j are larger than the minimum transmit power $p_{ij}^{\min}$ required to support a connection between node i to j, i.e., $\min\{p_i, p_j\} \geq p_{ij}^{\min}$. The minimum power consumption $p_{ij}^{\min}$ on link e_{ij} in the free space propagation model is usually characterized as

$$p_{ij}^{\min} = \frac{p^{th}(4\pi)^2}{G_i G_j \lambda^2} d_{ij}^{\tau} \tag{4}$$

where p_{th} is the signal capture threshold, G_i (or G_j) is the antenna gain of node i (or j), λ is the wave length, d_{ij} is the distance between nodes i and j, and τ is the path loss factor typically in the range $2 \leq \tau \leq 4$. Obviously, an energy-efficient topology with the connectivity constraint could be derived by our LCTDCC algorithm, if the power consumption on each link is used as the cost function.

Packet Error Probability. In a realistic communication scenario, the packets exchanged among the nodes may be received with errors, because of channel fading or noise. The symbol error probability ρ_s over an additive white Gaussian noise is usually written in the form $\rho_s(\gamma) = c_1 Q(\sqrt{c_2 \gamma})$ and the corresponding packet error probability is given by

$$PER = 1 - (1 - \rho_s(\gamma))^M, \tag{5}$$

where $Q(\cdot)$ is the Gaussian Q function, c_1, c_2 are two constants, γ is the instantaneous signal-to-noise ratio following Gaussian distribution, and M is the number of symbols in the packet. We can define the link cost with respect to the packet error probability as

$$c_{ij} = \mathbf{E}\left[\frac{1}{2}\left(PER_{i\to j} + PER_{j\to i}\right)\right], \tag{6}$$

where $\mathbf{E}[\cdot]$ denote the expect operation, $i \rightarrow j$ means the sender i and the receiver j. Similarly, the generated topology may have the best performance for correct packet transmission under the connectivity constraint, if the packet error probability is considered as the cost function.

Unwillingness for Packet Forwarding. As the network resource (e.g., battery energy, bandwidth) is limited, a node will get increasingly reluctant to use its precious resource. Assume that the current available resource of node i is R_i. Let an increasing function $f_i(x)$ be the pricing function indicating the price when x units of resource have been used. Then a specific unwillingness function for node i can be conceived by

$$U_i(R_i, r_i) = \int_{R_i^0 - R_i}^{R_i^0 - R_i + r_i} f_i(x)\mathrm{d}x, \tag{7}$$

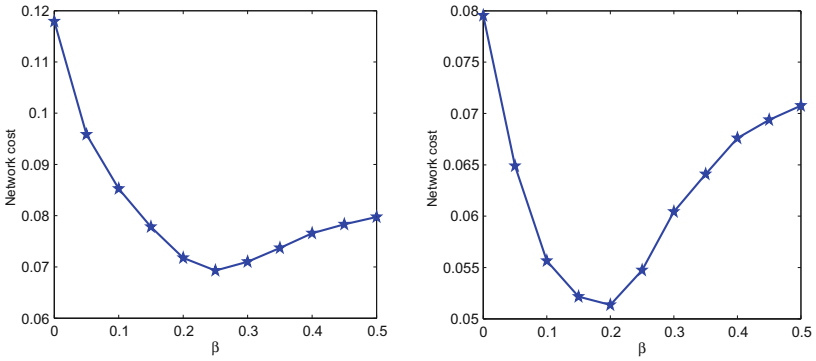
where R_i^0 denotes the initial resource for node i, $r_i \leq R_i$ is the resource requirement for packet forwarding. This specific unwillingness function is an increasing function of the resource requirement r_i and a decreasing function of its available resource R_i. The cost on link e_{ij} in this scenario is defined as the average unwillingness of nodes i and j, i.e., $c_{ij} = (U_i + U_j)/2$. Obviously, if we periodically configure the topology according to the LCTDCC algorithm based on the unwillingness, dynamic topologies may be derived for the efficient resource usage.

More different factors could be considered as the cost function in our work, such as the traffic load, node degree, signalling overhead. We can also use the weight combination of them to integrate various factors.

5 Simulation Results

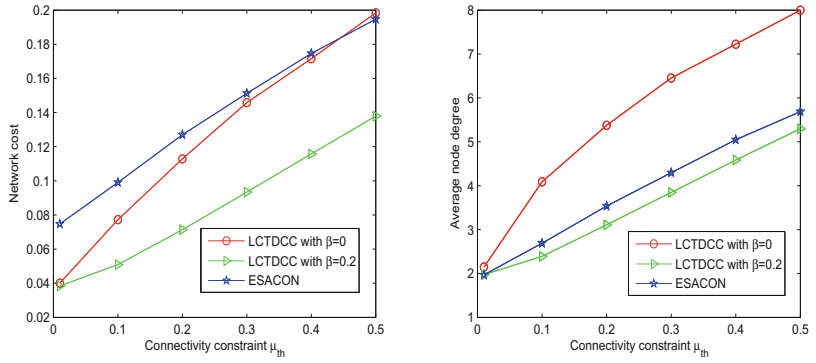
In this section, computer simulations are provided to illustrate the proposed algorithms. 60 nodes are randomly distributed in a 500×500 m^2 region in the simulation. The initial topology is generated by all nodes transmitting with the coverage range $d_{\max} = 150$ m. Two nodes are able to communicate directly via a wireless link, if they are within the range of each other. As a link with long distance generally means a heavy power consumption, an excessive interference or a large packet error probability, the link cost c_{ij} on link e_{ij} is assumed as $c_{ij} = Kd_{ij}^2$, where K is a constant. The value of K is set as 10^{-5}, such that the cost function and the connectivity contribution in (3) have similar order of magnitudes.

Firstly, we investigate the impact of weight parameter β in (3) on the network cost with connectivity constraints $\mu_{th} = 0.2$ and 0.1. For different values of β, the corresponding network costs of the generated topologies by our algorithm are shown in Fig. 1. Obviously, there exist two lowest points in the curves, that is, a pure consideration of the link cost (i.e., $\beta = 0$) is not enough to get the desired topology which has the minimum possible network cost with the connectivity



(a) $\mu_{th} = 0.2$ (b) $\mu_{th} = 0.1$

Fig. 1. Network cost vs. weight parameter β



(a) Network cost (b) Average node degree

Fig. 2. Network performance vs. connectivity constraint μ_{th}

constraint. This is because plenty of small-cost links are reserved, contributing small network connectivity but a large proportion of the network cost. Therefore, by capturing the network cost and the connectivity contribution jointly, the derived network topologies may have better performance in network cost.

Next, we run our algorithm by varying the connectivity constraint a_{th} from 0 to 0.5 to get the corresponding network performance, i.e., network cost and average node degree (i.e., average number of neighbors). LCTDCC is implemented with weight parameter $\beta = 0.2$, which seems a better choice as shown in Fig. 1, and $\beta = 0$, i.e., only based on the link cost. We also compare the proposed LCTDCC algorithm with ESACON in [14] which has the similar objective with this work in topology design. The variations of network cost and average node degree with connectivity constraint μ_{th} are shown in Fig. 2(a) and (b), respectively. As the threshold μ_{th} increases, more links should be reserved, resulting in more network cost and larger node degree. Obviously, the topology with a

specific connectivity constraint generated by LCTDCC with $\beta = 0.2$ has the least network cost and node degree compared against LCTDCC with $\beta = 0$ and ESACON.

6 Conclusions

The inherent tradeoff between network cost and connectivity is an essential requirement in wireless ad hoc networks. Therefore, the focus of our work is on the topology design to compromise between reducing the network cost and improving the network connectivity. A new topology metric of connectivity-cost-ratio for each link was defined to guide our topology design. We proposed the LCTDCC algorithm in a wireless ad hoc network to reduce the network cost with a connectivity constraint. The cost function in this work is general, which could be defined by various practical applications.

Acknowledgements. This research was supported by NSF China (61471287), Foundation of Key Scientific Research Projects of Henan Educational Committee (18B510022, 16A520105), and the Development Project of Henan Provincial Department of Science and Technology (172102310124).

References

1. Li, N., Hou, J.C., Sha, L.: Design and analysis of an MST-based topology control algorithm. IEEE Trans. Wirel. Commun. **4**, 1195–1206 (2005)
2. Xu, H., Huang, L., Qiao, C., Wang, X., Sun, Y.: Topology control with vMIMO communication in wireless sensor networks. IEEE Trans. Wirel. Commun. **12**, 6328–6339 (2013)
3. Komali, R.S., Mackenzie, A.B., Gilles, R.P.: Effect of selfish node behavior on efficient topology design. IEEE Trans. Mobile Comput. **7**, 1057–1070 (2008)
4. Zarifzadeh, S., Yazdani, N., Nayyeri, A.: Energy-efficient topology control in wireless ad hoc networks with selfish nodes. Comput. Netw. **56**, 902–914 (2012)
5. Waqas, A., Mahmood, H.: A game theoretical approach for topology control in wireless ad hoc networks with selfish nodes. Wirel. Pers. Commun. **96**, 249–263 (2017)
6. Xu, M., Yang, Q., Kwak, K.S.: Algebraic connectivity aided energy-efficient topology control in selfish ad hoc networks. Wirel. Netw. **23**, 1331–1341 (2017)
7. Shang, D., Zhang, B., Yao, Z., Li, C.: An energy efficient localized topology control algorithm for wirelss multihop networks. J. Commun. Netw. **16**, 371–377 (2014)
8. Chu, X., Sethu, H.: An energy balanced dynamic topology control algorithm for improved network lifetime. In: IEEE WiMob, Larnaca, pp. 556–561. IEEE Press (2014)
9. Huang, M., Chen, S., Zhu, Y., Wang, Y.: Topology control for time-evolving and predictable delay-tolerant networks. IEEE Trans. Comput. **62**, 2308–2311 (2013)
10. Nahir, A., Orda, A., Freund, A.: Topology design of communication networks: a game-theoretic perspective. IEEE/ACM Trans. Netw. **22**, 405–414 (2014)
11. Abreu, N.M.M.: Old and new results on algebraic connecitivty of graphs. Linear Algebra Appl. **423**, 53–73 (2007)

12. Pandana, C., Liu, K.J.: Robust connectivity-aware energy-efficient routing for wireless sensor networks. IEEE Trans. Wirel. Commun. **7**, 3904–3916 (2008)
13. Zhang, Z., Wang, X., Xin, Q.: A new performance metric for construction of robust and efficient wireless backbone network. IEEE Trans. Comput. **61**, 1495–1506 (2012)
14. Cuomo, F., Abbangnale, A., Cianfrani, A., Polverini, M.: Keeping the connectivity and saving the energy in the internet. In: IEEE INFOCOM, Shanghai, pp. 319–324. IEEE Press (2011)



Waveform Design of UNB Modulation and Performance Analysis

Aoqin Duan[1,2(✉)], Qinghe Du[1,2], and Bo Sun[3]

[1] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China
duan2402729060@stu.xjtu.edu.cn, duqinghe@mail.xjtu.edu.cn
[2] National Simulation Education Center for Communications and InformationSystems, Xi'an, China
[3] ZTE Corporation, Xi'an, China
sun.bo1@zte.com.cn

Abstract. To improve the spectrum efficiency and alleviate scarcity, ultra-narrow band (UNB) technique is proposed with high bandwidth efficiency and low bit error rate (BER). In this paper, the waveform design schemes, which are sine amplitude modulation (AM-sine), dual-frequency amplitude modulation (DFAM) and constant modulus dual-frequency (CMDF) modulation for wireless communication networks employing UNB technique, are proposed. We investigate the Euclidean distance of binary modulated signals for UNB communication under the bandwidth restriction compared to very-minimum waveform difference keying (VWDK). We derive the simplicity to optimize our proposed schemes by introducing sine-like signal. By applying the concavity optimization theory, we get more smooth phase waveform than VWDK. We formulate the optimization problem to minimize the average BER subject to these conditions, one is the Euclidean distance between modulated signals and carrier signal, the other one is the signal energy. Simulation results are provided to show the superiority of our proposed schemes as compared to VWDK.

Keywords: AM-sine · DFAM · CMDF · Ultra narrow band

1 Introduction

In recent years, with the massive deployment of the Internet of Things (IoT) communication devices, spectrum resource [1] is in the state of being scarce. Therefore, it is worth to further research for the development of IoT communication [2]

Supported by the National Natural Science Foundation of China under the Grant No. 61461136001, the National Natural Science Foundation of China under the Grant No. 61671371, Science and Technology Program of Shaanxi Province under the Grant No. 2016KW-032, the ZTE Industry Academic Research Cooperation Fund and the undamental Research Funds for the Central Universities.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 84–94, 2018.
https://doi.org/10.1007/978-3-319-78130-3_10

to alleviate the spectrum shortage. On the one hand, we can improve communication performance by increasing network capacity, enhancing spectral efficiency or reducing the detection bit error rate (BER). On the other hand, pluralism and high integration design are also helpful to improve the network services performance. However, the existing traditional techniques, such as ultra-wide band (UWB) [4], which propose the conception of spectrum sharing, will must lead to the interference when communications. Cognitive Radio (CR) [3] dynamically allocated spectrum resource to improve spectrum utilization at the expense of the high complexity and high cost of equipments. However, these technologies still can't meet the requirements of the current communication businesses.

In order to attain high bandwidth efficiency, researchers use a variety of coding techniques to design the signal waveform modulation in communication system. Then, ultra-narrow band (UNB) [14], wherein, is proposed by Walker, a patent in [5]. Besides, he proposed a series of the other UNB modulation methods, such as enhanced Variable Phase Shift Keying (VPSK) [6], Very Minimum Shift Keying (VMSK) [7,11,13] and Pulse Position Phase Reversal Keying (3PRK) [8] and so on. There are mainly two research teams in the world that work on UNB modulation. One is led by Walker, an America engineer, in New Jersey, and the other team is led by Wu Lenan, a professor of Southeast University, Nanjing, China. In this paper, based on the Very-minimum Waveform Difference Keying (VWDK) which is proposed by Wu Lenan, we proposed our own waveform design schemes to analyze the narrow spectrum and low BER performance of the transmitted signals, compared to VWDK with the same power attenuation bandwidth constraint.

The rest of the paper is organized as follows. The system model is described in Sect. 2. In Sect. 3, we provided the optimization problem and made the convexity analysis. The proposed waveform design schemes of UNB communication are described in Sect. 4. In Sect. 5, simulation results are presented to evaluate the performance of the proposed schemes. The conclusions are drawn in Sect. 6.

2 System Model

2.1 Signal Transmission Model

UNB communication, a long-known technique, use the traditional model of carrier communication, which belongs to digital communication, but also has the inextricably linked with the analog communication.

As we all know, as to the simple binary digital communication system model, after transformation and encoding of the source signals, the signal transmitter is controlled by the binary digital (0 or 1) which is outputted every T seconds. During the symbol period T, send signal $s_1(t)$ (on behalf of the transfer of number 1); or send signal $s_0(t)$ (on behalf of the transfer of number 0). Then through the transmission channel, the received signals are attenuated or distorted that can be ignored with the channel noise $n(t)$. At the receiver, we can obtain the mixed received signal and the expression is recorded as: $y(t) = h(t)x(t) + n(t)$, where $h(t)$ represents the complex channel coefficient (or fading process) and

the transmitted signal $s_i(t),\ (i = 0, 1)$ is defined in $[0, T)$. The signal energy per bit can be expressed as

$$E_{si} = \int_0^T s_i^2(t)dt \tag{1}$$

where $x(t)$ represents the digital baseband transmitted signal with a continuous time and $n(t)$ an additive white Gaussian noise with zero mean, and whose power spectral density is $N_0/2$.

In our model, considering the width of the narrow band signal spectrum, we assume that our signal transmission is always in the flat fading channels, complex channel coefficient $h(t)$ is deemed to a constant value h. Without loss of generality, we assume that $h = 1$. The expression of $x(t)$ is expressed as follows:

$$x(t) = \sum_{n=-\infty}^{+\infty} a_n s_1(t - iT) + \sum_{n=-\infty}^{+\infty} (1 - a_n)s_0(t - iT) \tag{2}$$

where $\{a_n\}$ is the binary random sequence with equal probability. Meanwhile, $s_1(t)$ and $s_0(t)$ are the transmitted signals that respect to data 1 and data 0, respectively. T is the duration for each data bit. Then the concise form of the received signal is

$$y(t) = x(t) + n(t), \tag{3}$$

compared with $y(t) = h(t)x(t) + n(t)$.

2.2 Requirements of UNB

UNB technique uses the narrow physical bandwidth to transmit the data with a high data rate, which is proposed to improve the spectrum efficiency with a constant bit rate.

Theoretically, UNB modulation makes a greatly reduced of transmission bandwidth requirements with the narrower frequency bandwidth than other existing modulation technologies. However, how much the UNB signal bandwidth is defined that can be said that it meet the UNB transmission? And now there is still no uniform definition of UNB communication bandwidth. The only one of the commercial UNB system, proposed by SigFox in the C-UNB program, defines UNB in the uplink transmission bandwidth is 250 Hz and downlink is 1000 Hz. Besides, the UNB supporters hold that UNB signal transmission bandwidth is the combination between the spectrum width of carrier and the bandwidth of sideband signals with the power attenuate to 60 dB compared with the carrier frequency component, and the other components of the signal which beyond this bandwidth are considered as adjacent noise is ignored. And the existing technologies for high bandwidth efficiency of UNB modulation are all based on the Federal Communications Commission (FCC) standard (60 dB required by FCC).

3 Optimized Problem and Analysis for Sine-Like UNB Waveform Scheme

3.1 Optimization Problem Formulation

To maximize the Euclidean distance of the modulated signals $s_0(t)$ and $s_1(t)$, and minimize the demodulation error rate with the bandwidth and energy constraints. The parameters α, the difference between modulated signals and carrier sine waveform, adjusts the Euclidean distance of the modulated signals indirectly. The symbol energy is E. Through these assumption, we hope to get the modulated signal waveform with narrower frequency bandwidth and lower BER than VWDK. Therefore, we formulate the bandwidth-constrained Euclidean distance indirectly maximization problem by using the α value and the energy limits as follows:

$$\max \int_0^T (s_1(t) - s_0(t))^2 dt. \tag{4}$$

$$\text{s.t.:} \int_0^T \big(s_1(t) - \sin(\omega_0 t)\big)^2 dt \le \alpha \ , \int_0^T (s_0(t) - \sin(\omega_0 t))^2 dt \le \alpha; \tag{5}$$

$$\int_0^T s_1^2(t) dt \le E, \qquad \int_0^T s_0^2(t) dt \le E. \tag{6}$$

When data 0 is transmitted, modulated waveform $s_0(t)$ is selected and when 1 is transmitted, modulated waveform $s_1(t)$ is selected or vice versa. Where the carrier frequency is $\omega_0 = 1$. Equation (5) restricts the Euclidean distance between modulation signals and carrier signal, Eq. (6) is the energy limitation with the data. The signal also meet the requirement that DC component is zero during one symbol period. The equivalent optimization problem of the Eqs. (4), (5) and (6) can be described as follows:

$$\min \quad -\int_0^T (s_1(t) - s_0(t))^2 dt \tag{7}$$

$$\text{s.t.: Eqs. (5) and (6)} \tag{8}$$

3.2 Convexity Analysis

Based on the above problem, we use the Lagrangian approach to analyze the convexity of it under the bandwidth constraint which is given by Eq. (8). According to the definition [16] of a convex optimization problem is that its objective function is convex on its convex domain set. Then, we write the convex Lagrangian function as follows:

$$
\begin{aligned}
L\big(s_1(t), s_0(t); a, b, c, d\big) &= -\int_0^T \big(s_1(t) - s_0(t)\big)^2 dt \\
&+a\Big(\Big[\int_0^T \big(s_1(t) - \sin(\omega_0 t)\big)^2 dt\Big] - \alpha\Big) \\
&+b\Big(\Big[\int_0^T \big(s_0(t) - \sin(\omega_0 t)\big)^2 dt\Big] - \alpha\Big) \\
&+c\Big(\Big[\int_0^T s_1^2(t)\, dt\Big] - E\Big) + d\Big(\Big[\int_0^T s_0^2(t)\, dt\Big] - E\Big) \quad (9)
\end{aligned}
$$

where $a \geq 0, b \geq 0, c \geq 0, d \geq 0$ are the Lagrangian multipliers.

To analyze and work out the optimum solution, we take the first-order derivatives of $L\big(s_1(t), s_0(t); a, b, c, d\big)$ with respect to $s_1(t)$ and $s_0(t)$, respectively. As to the Lagrangian function, we have known that when the Lagrangian multipliers meet the conditions that $a + c - 1 > 0$ and

$$
\begin{vmatrix} \frac{\partial^2 L}{\partial s_1^2(t)} & \frac{\partial^2 L}{\partial s_1(t) s_0(t)} \\ \frac{\partial^2 L}{\partial s_0(t) s_1(t)} & \frac{\partial^2 L}{\partial s_0^2(t)} \end{vmatrix} = \begin{vmatrix} 2\big(a + c - 1\big) & 2 \\ 2 & 2\big(b + d - 1\big) \end{vmatrix} > 0, \quad (10)
$$

the Lagrangian function is convex. The unique solution of Lagrangian function is obtained at the point, the minimal point of the optimization problem, where the first-order derivatives of Lagrangian function is equal to zero. The expression for the unique solution is given by Eq. (11).

$$
\begin{cases}
s_1(t) = \dfrac{\big[b\big(a-1\big)+a\big(d-1\big)\big]\sin\big(\omega_0 t\big)}{\big(a+c-1\big)\big(b+d-1\big)-1}, & t \in \big[0, T\big); \\
s_0(t) = \dfrac{\big[a\big(b-1\big)+b\big(c-1\big)\big]\sin\big(\omega_0 t\big)}{\big(a+c-1\big)\big(b+d-1\big)-1}, & t \in \big[0, T\big).
\end{cases} \quad (11)
$$

4 Sine-Like Dual-Frequency Signal Design

Based on the VWDK [9,10], we consider the following waveform design schemes with high detection performance of UNB modulation. Then we make a combination of the traditional binary digital communication theory, the convex optimization problem with any points that satisfy the Karush-Kuhn-Tucker (KKT) conditions, and high order polynomial to smooth our proposed signal waveform. It should be meet the requirement that the sideband power, which out-of-main lobe have low power level (60 dB required by FCC). We named our proposed design schemes as sine Amplitude Modulation (AM-sine), Dual-frequency Amplitude Modulation (DFAM), and Constant Modulus Double-frequency (CMDF) Modulation.

4.1 AM-sine Waveform

By solving the above optimization problem, we get the expression in Eq. (11) named as AM-sine waveform and written as $s_1^{AM}(t), s_0^{AM}(t)$.

We assume that the binary digital signal's normalized power $E = 1$, then we can adjust the Euclidean distance between the modulation signals $s_1^{AM}(t), s_0^{AM}(t)$ and the carrier, which is restricted to α. By applying the gradient descent method with fixed step size m, we select the initial value for Lagrange multipliers, i.e., $\boldsymbol{\lambda}_0 = (a, b, c, d) = (12, 3, 2, 1)$, to make the one-dimensional search along the negative gradient direction. The expression can be written as follows, $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + m\nabla L(\boldsymbol{\lambda}_k), k = 0, 1, 2, \ldots, K-1$. K is the total number of iterations. Finally, we attained the optimal convergence values of Lagrangian multipliers, i.e., $\boldsymbol{\lambda}^* = (a^*, b^*, c^*, d^*)$. The convergence curves are shown in Fig. 1, while the proposed AM-sine waveform is shown in Fig. 2(a).

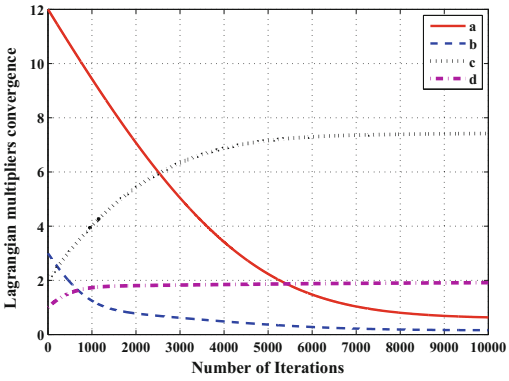


Fig. 1. Optimized Lagrange convergence curves

4.2 DFAM Waveform

According to AM-sine waveform, what the problem that we cannot avoid is the time-selective fading in the actual application scenarios. If we want to get greater difference of the signals than AM-sine waveform in time domain. It will inevitably lead to a loss of energy efficiency, while the waveform exhibits a certain degree of difference in time domain. Therefore, we give out the following research which is named DFAM modulation and recorded as $s_1^{DFAM}(t), s_0^{DFAM}(t)$.

First, the secondary high frequency section $\sin^2(t)$ is introduced into the AM-sine signal.

Second, we use the cubic polynomial function to smooth the dual-frequency sine waveform. The signal expressions are given by

$$s_1^{DFAM}(t) = \begin{cases} \frac{\tilde{a}}{\sin(\tau)}\sin^2(t), & t \in [0, t_L), \\ \sum_{i=0}^{3} a_{i1} t^i, & t \in [t_L, t_R), \\ \tilde{a}\sin(t), & t \in [t_R, T - t_R), \\ \sum_{i=0}^{3} b_{i1} t^i, & t \in [T - t_R, T - t_L), \\ -\frac{\tilde{a}}{\sin(\tau)}\sin^2(t), & t \in [T - t_L, T); \end{cases} \tag{12}$$

and

$$s_0^{DFAM}(t) = \begin{cases} \frac{\tilde{b}}{\sin(\tau)}\sin^2(t), & t \in [0, t_L), \\ \sum_{i=0}^{3} a_{i0}t^i, & t \in [t_L, t_R), \\ \tilde{b}\sin(t), & t \in [t_R, T - t_R), \\ \sum_{i=0}^{3} b_{i0}t^i, & t \in [T - t_R, T - t_L), \\ -\frac{\tilde{b}}{\sin(\tau)}\sin^2(t), & t \in [T - t_L, T). \end{cases} \tag{13}$$

where $\tilde{a} = \frac{\left(b^*+d^*-1\right)a^*-b^*}{\left(a^*+c^*-1\right)\left(b^*+d^*-1\right)-1}$, $\tilde{b} = \frac{\left(a^*+c^*-1\right)b^*-d^*}{\left(a^*+c^*-1\right)\left(b^*+d^*-1\right)-1}$, $0 \le t_L < \tau < t_R \le \frac{T}{4}$, τ is defined as the time when polynomial is used to smooth the dual-frequency signal. The proposed DFAM waveform is shown in Fig. 2(b). The following expressions are used to make sure the polynomial coefficients a_{ij}, b_{ij} $(i = 0, 1, 2, 3;\ j = 0, 1)$.

(1) $\frac{\tilde{a}}{\sin(\tau)}\sin^2(t)$ and $\sum_{i=0}^{3} a_{i1}t^i$ are tangent at $t = t_L$, meanwhile, $\tilde{a}\sin(t)$ and $\sum_{i=0}^{3} a_{i1}t^i$ are tangent at $t = t_R$. Then we have the following equation that

$$\begin{bmatrix} 1 & t_L & t_L^2 & t_L^3 \\ 1 & t_R & t_R^2 & t_R^3 \\ 0 & 1 & 2t_L & 3t_L^2 \\ 0 & 1 & 2t_R & 3t_R^2 \end{bmatrix} \begin{bmatrix} a_{01} \\ a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} = \begin{bmatrix} \frac{\tilde{a}}{\sin(\tau)}\sin^2(t_L) \\ \tilde{a}\sin(t_R) \\ \frac{2\tilde{a}}{\sin(\tau)}\sin(t_L)\cos(t_L) \\ \tilde{a}\cos(t_R) \end{bmatrix}. \tag{14}$$

(2) $\tilde{a}\sin(t)$ and $\sum_{i=0}^{3} b_{i1}t^i$ are tangent at $t = T - t_R$, meanwhile, $-\frac{\tilde{a}}{\sin(\tau)}\sin^2(t)$ and $\sum_{i=0}^{3} b_{i1}t^i$ are tangent at $t = t_R$. Then we have the following equation that

$$\begin{bmatrix} 1 & T-t_R & (T-t_R)^2 & (T-t_R)^3 \\ 1 & T-t_L & (T-t_L)^2 & (T-t_L)^3 \\ 0 & 1 & 2(T-t_R) & 3(T-t_R)^2 \\ 0 & 1 & 2(T-t_L) & 3(T-t_L)^2 \end{bmatrix} \begin{bmatrix} b_{01} \\ b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} = \begin{bmatrix} \tilde{a}\sin(T-t_R) \\ -\frac{\tilde{a}}{\sin(\tau)}\sin^2(T-t_L) \\ \tilde{a}\cos(T-t_R) \\ -\frac{2\tilde{a}}{\sin(\tau)}\sin(T-t_L)\cos(T-t_L) \end{bmatrix}. \tag{15}$$

Using Eqs. (14) and (15) to calculate the polynomial coefficients a_{i1}, b_{i1} correspond to bit 1. Similarity, as to bit 0, we can do to get the expression coefficients a_{i0}, b_{i0}. By adjusting the values of polynomial smoothing time t_L, τ, and t_R, we can obtain the modulation signal with different time domain divisions.

Third, the method to remove the other high-order discrete linear spectrums generated by the periodical component of the waveform is $\tilde{s}_1(t) = \frac{s_1(t)-s_0(t)}{2} + A_1\sin(\omega_0 t)$, $\tilde{s}_0(t) = \frac{s_0(t)-s_1(t)}{2} + A_1\sin(\omega_0 t)$. Therefore, the modulation waveform turn into as Eq. (16).

$$\begin{cases} s_1^{DFAM}(t): & s_1^{DFAM}(t) - \left(\frac{s_1^{DFAM}(t)+s_0^{DFAM}(t)}{2}\right) + A_1\sin(\omega_0 t); \\ s_0^{DFAM}(t): & s_0^{DFAM}(t) - \left(\frac{s_1^{DFAM}(t)+s_0^{DFAM}(t)}{2}\right) + A_1\sin(\omega_0 t). \end{cases} \tag{16}$$

where $\omega_0 = 1$, $A_n = \frac{2}{T}\int_0^T \left(\frac{s_1(t)+s_0(t)}{2}\right)\sin(nt)dt$ is the amplitude of Nth-degree harmonic.

4.3 CMDF Waveform

Combined with AM-sine and DFAM, the further consideration to avoid the problem of spectrum leakage brought by amplitude fluctuation, we process the DFAM signal into the constant modulus signal, named as CMDF modulation and recorded as $s_1^{CMDF}(t)$, $s_0^{CMDF}(t)$.

The transition process from DFAM to CMDF as follows.

(1) Remove the carrier frequency from $s_1^{DFAM}(t)$ and $s_0^{DFAM}(t)$, respectively. Then we get the modulation signals $\tilde{s}_1(t)$, $\tilde{s}_0(t)$ without carrier component.

$$\begin{cases} \tilde{s}_1(t) = s_1^{DFAM}(t) - \frac{2\sin(\omega_0 t)}{T} \int_0^T s_1^{DFAM}(t) \sin(\omega_0 t)\, dt; \\ \tilde{s}_0(t) = s_0^{DFAM}(t) - \frac{2\sin(\omega_0 t)}{T} \int_0^T s_1^{DFAM}(t) \sin(\omega_0 t)\, dt. \end{cases} \tag{17}$$

(2) Remove other high-order harmonics from $\tilde{s}_1(t)$ and $\tilde{s}_0(t)$, we attain the signals named as $\tilde{s}_1^{con}(t)$, $\tilde{s}_0^{con}(t)$, which are only contain the continuous spectrum.

$$\begin{cases} \tilde{s}_1^{con}(t) = \tilde{s}_1(t) - \left(\frac{\tilde{s}_1(t)+\tilde{s}_0(t)}{2}\right); \\ \tilde{s}_0^{con}(t) = \tilde{s}_0(t) - \left(\frac{\tilde{s}_1(t)+\tilde{s}_0(t)}{2}\right). \end{cases} \tag{18}$$

(3) At last, the carrier component is added to the signals which are attained at step 2). The carrier frequency is used to pinpoint the timing information. Therefore, the final waveform expression can be written as follows:

$$\begin{cases} s_1^{CMDF}(t) = \tilde{s}_1^{con}(t) + \frac{1}{T} 2A_1 \sin(\omega_0 t) \int_0^T \left(\frac{\tilde{s}_1(t)+\tilde{s}_0(t)}{2}\right) \sin(\omega_0 t) dt; \\ s_0^{CMDF}(t) = \tilde{s}_0^{con}(t) + \frac{1}{T} 2A_1 \sin(\omega_0 t) \int_0^T \left(\frac{\tilde{s}_1(t)+\tilde{s}_0(t)}{2}\right) \sin(\omega_0 t) dt. \end{cases} \tag{19}$$

5 Simulation Evaluations

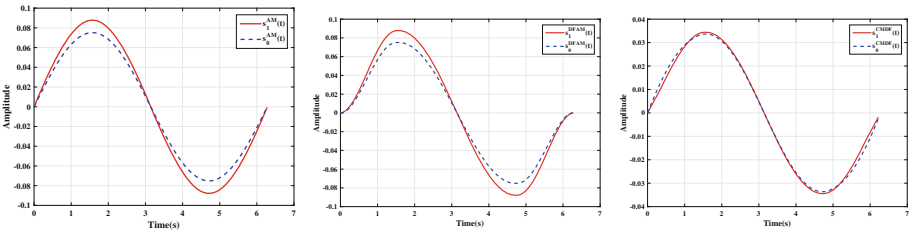
In this part, we mainly present the simulation results and the performance analysis of our proposed schemes with the existing technology–VWDK.

First, we give the waveforms of our proposed waveform schemes, which are corresponding to AM-sine, DFAM and CMDF. Then, the simulation experiments to evaluate the BER performance between these waveforms and VWDK is established in the following assumptions: binary information sequence $\{a_n\}$ appeared to a equiprobability, without loss of generality, we adopt the correlation reception with $n(t)$ an additive white Gaussian noise with zero mean, and whose power spectral density is $N_0/2$ in the AWGN channel. The BER simulalion uses BPSK, and is implemented in MATLAB, the simulation results are shown in following figures.

According to Eq. (9), the fitting parameters for different Lagrangian multipliers are provided in Fig. 1. The optimal values are shown in Eq. (20).

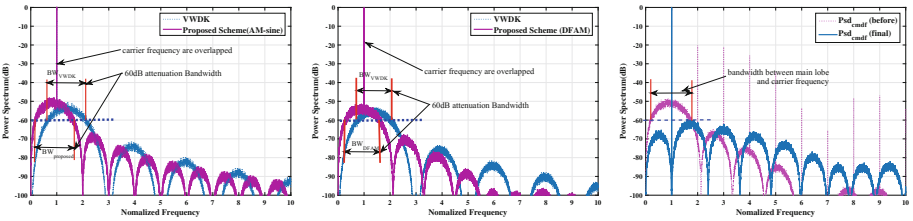
$$a^* = 0.6988, b^* = 0.1686, c^* = 7.4089, d^* = 1.9075. \tag{20}$$

Figure 2(a) shows the AM-sine modulated waveforms with different amplitudes. Use it to transfer the UNB signal, the variation of the amplitude is beneficial to capture and synchronization the transmitted signals. Spectrum is shown in Fig. 3(a) along with the equal 60 dB attenuation bandwidth with VWDK. From the Fig. 4(a), we can see that the performance gain of AM-sine waveform will remain 2.5 dB compared to VWDK with the same bandwidth constraint. The performance improvement indicates the advantage of AM-sine scheme.



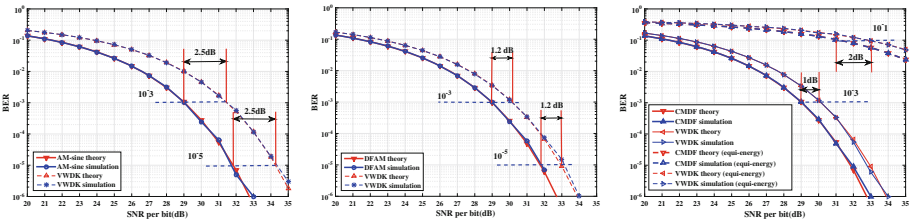
(a) (b) (c)

Fig. 2. Our proposed waveforms. (a) AM-sine waveform (Eq. (11)). (b) DFAM waveform (Eqs. (12) and (13)). (c) CMDF waveform (Eq. (19)).



(a) (b) (c)

Fig. 3. Spectrums of our proposed schemes and VWDK. (a) Spectrums of AM-sine and VWDK. (b) Spectrums of DFAM and VWDK. (c) Spectrums of CMDF.



(a) (b) (c)

Fig. 4. BER evaluations and comparisons of our proposed schemes and VWDK. (a) BER comparisons of AM-sine and VWDK. (b) BER comparisons of DFAM and VWDK. (c) BER comparisons of CMDF and VWDK.

Figures 2(b) and 3(b) shows the DFAM waveform and its spectrum, respectively. Figure 2(b) embodies a DFAM plot that isn't constrained by sinusoidal waveforms. Compared to the AM-sine signal and VWDK waveform, we get a bigger difference in the time domain, but at the expense of a certain energy efficiency. Also, the BER of 10^{-3} can be obtained the gain about 1.2 dB with the same frequency bandwidth compared to VWDK is shown in Fig. 4(b).

Figure 2(c) depicts the CMDF waveform during one bit period. In order to avoid the spectrum leakage caused by amplitude fluctuations, we proposed the CMDF waveform. The side-band spectrums shall be attenuated at least 60 dB relative to the carrier, which is required by FCC and shown in Fig. 3(c). It is visible that there is a detection performance gain. The range of gain is about 1–2 dB after the signal constant modulus processing.

6 Conclusions

In this paper, we made a combination of binary digital communication and the optimization theory under the enlightenment of VWDK, and then the sine-like waveform design schemes were put forward for UNB communication, which were introduced by the sine signal indirectly. Using the new design ideas, our proposed schemes can effectively enhance the SNR about 1–2.5 dB, compared to VWDK with the same bandwidth constraint (60 dB required by FCC).

References

1. Wu, L.: The progress of ultra narrow band communication. Prog. Nat. Sci. **17**(11), 1467–1473 (2007)
2. http://www.imt-2020.org.cn/zh/
3. Chakravarthy, V.D., et al.: Cognitive radio-adaptive waveforms having spectral sharing capability. In: IEEE Communication Society/WCNC 2005, vol. 2, no. 13–17, pp. 724–729, March 2005
4. Sablatash, M.: Adaptive and cognitive UWB radio and a vision for flexible future spectrum management. In: Information Theory, pp. 41–44 (2007)
5. Walker, H.R.: High speed binary data communication system. U.S. Patent, 004742532, 03 May 1988
6. Walker, H.R.: High speed data communication system using phase shift key coding. U.S. Patent, 005185765, 09 February 1993
7. Walker, H.R.: Digital modulation employing single sideband with suppressed carrier. U.S. Patent, 005930303. 27 July 1999
8. Walker, H.R.: Ultra narrow band modulation. In: International Conference on Computing, Communications and Control Technologies, CCCT 2004, Austin TX (2004)
9. Si, H., Li, X., Wu, L.: PSD analysis of very-minimum waveform difference keying signals. In: Michael, J.S., Hsingwei, C. (eds.) International Conference on CCCT(6), pp. 142–144. University of Texas at Austin, Austin (2004)
10. Li, X., Si, H., Wu, L.: On spectrum structure and optimization of VWDK modulated waveforms. In: Michael, J.S., Hsingwei, C. (eds.) International Conference on CCCT(6), pp. 149–152. University of Texas at Austin, Austin (2004)

11. Zhang, S.: Bandwidth efficiency analysis of UNB modulation schemes. In: International Conference on Wireless Communication, pp. 1–4 (2008)
12. Zhai, X., Lin, H., He, H., et al.: Typical UNB modulation methods and their spectrums. In: IET International Conference on Wireless Mobile and Computing, pp. 286–283 (2009)
13. Sayhood, K.H., Wu, L.: Raise bandwidth efficiency with sine-wave modulation VMSK. Microw. RF Mag. **4**, 79–84 (2001)
14. http://www.vmsk.org (2009)
15. Zhang, S.: opEBPSK based UNB demodulation by using special filters. In: International Conference on Intelligent Computation Technology and Automation, vol. 2, pp. 632–635 (2011)
16. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)

System-Level Simulation Platform for Device-to-Device Communications in 5G Networks

Zhihao Ding[1,2(✉)], Pinyi Ren[1,2], and Qinghe Du[1,2]

[1] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China
dingzhihao@stu.xjtu.edu.cn, {Pyren, duqinghe}@mail.xjtu.edu.cn
[2] National Communications and Information Systems Virtual Simulation Experiment Teaching Center, Xi'an, China

Abstract. Device-to-Device (D2D) Communication enables User Equipment (UE) to establish direct link between each other to transmit data, which can improve both the efficiency of spectrum and throughput, has become a key technology in the Fifth Generation (5G) system. However, there still exist many challenges to ensure the D2D communication and the 5G standard hasn't been released. In order to promote the development of D2D technology and provide a simulation tool for the research of 5G D2D communication, in this work, we present the design of system-level simulation platform for 5G D2D communication according to the guidelines of METIS Project. We introduce several key technologies in the simulation platform, including resource allocation, interference computation, power control, etc. Finally, we give the simulation results and summarize that our 5G D2D communication system-level simulation platform is able to offer rich interface for testing and evaluation of related technologies.

Keywords: 5G · Device-to-Device (D2D) · Simulation platform

1 Introduction

In the past few years, the development of communication technology has brought great convenience to people. The current Fourth Generation (4G) network provides fast uplink and downlink data transfer rates which supports a wide variety of applications. The trends of future explosive growth of data traffic, the rapid popularization of intelligent terminals, more network access points and a new round of changes in the mobile industry will go beyond the capacity of 4G mobile communication system by

This paper is supported by the National Natural Science Foundation of China under the Grant No. 61461136001, the National Science and Technology Major Project of China under Grant No. 2016ZX03001016-005 and the Key Research and Development Program of Shannxi Province under Grant No. 2017ZDXM-GY-012.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 95–105, 2018.
https://doi.org/10.1007/978-3-319-78130-3_11

the year 2020 [1]. Meanwhile, the evolution of 5G oriented wireless communication technology becomes more urgent and clear which has come to the foreground in recent years. Hence, organizations in different countries have started new technology research in 5G system to satisfy the harsh demands.

Compared with cellular communication, D2D communication has captured wide attention of researches because of unique advantages in improving the utilization of spectrum resources, expanding network coverage and reducing the power consumption of User Equipment [2]. Therefore, D2D communication is considered as an important complement to not only the 5G cellular communication, but also an important part of the Internet of things (IoT). D2D communication can be divided into two categories according to the ways of using spectrum resources, it is referred to as inband D2D and outband D2D [3, 4]. The inband D2D means that D2D users share the same spectrum resource as cellular users which known as multiplexing mode. Inband D2D is the best way to improve resource utilization, but brings serious channel interference between D2D users and cellular users especially when D2D users are close to cellular users. Inband D2D can be further divided into underlay D2D and overlay D2D on the basis of occupancy of cellular spectrum. Underlay D2D occupies the whole cellular spectrum as cellular network, while, overlay D2D just occupy a portion of the cellular spectrum and there is no overlap between each other. In addition, D2D communication works on unlicensed spectrum called outband D2D and there will be no interference between D2D and cellular communication.

Recently, lots of studies on D2D communication have been released. A secure data sharing protocol merges the advantage of cryptography in D2D communication is proposed in [5]. To improve the service in D2D communication, a power allocation scheme and a cellular spectrum sharing plan are proposed in [6]. Furthermore, the method to reduce delay in multi-hop D2D communications by using social network features and communication domain constraints is discussed in [7]. Interference reduced scheme between D2D links and cellular network is investigated in [8].

In this paper, we discuss the design and implementation of D2D communication system-level simulation platform. In general, we consider two types of D2D links, direct D2D link with underlay mode and two hop D2D link with overlay mode, as shown in Fig. 1. The direct D2D link build a direct link between two D2D devices while two hop D2D link requires a D2D relay to forward the information sent by the base station, for instance, a D2D user might download video or other information from base station while the long distance between each other leads to the poor communication quality. In this case, it will get better by a D2D user worked as relay to forward the information between the receive user and base station. Last but not least, our whole original intentions and goals are to provide a system-level simulation tool for D2D research.

The rest of this paper is arranged as follows. In Sect. 2, the overall architecture of the platform is briefly introduced. In Sect. 3, we will detailly discuss some of the modules and technologies in the simulation platform. In Sect. 4, the simulation results are shown under the typical parameters. Finally, a conclusion is given in Sect. 5.

2 System Description and System Architecture

2.1 System Description for the D2D Networks

D2D communication is a new technology of direct communication between devices under the control of the system by multiplexing cellular resources. In this work, we mainly set up two types of D2D links because these link types represent the typical D2D communication mode. Figure 1 shows the link type in this D2D simulation platform. Direct D2D communication operates on the cellular resources with underlay mode such as the link between UE1 and UE2. There are two hop links with overlay mode such as the link from base station to UE 4. Two different links are simulated to prove the ability that D2D communication can enhance network capacity.

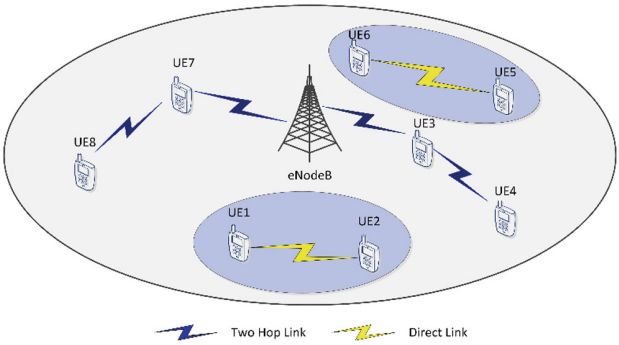


Fig. 1. System description for D2D networks

2.2 Architecture of the Simulation Platform

We consider a modular design method to build the whole simulation system and the platform is divided into several modules, including engine module, network initialization module, protocol stack module, network node module, statistical calculation module, etc. As a simplification, we only design four layers of the OSI seven-layer model including application layer, network layer, MAC layer and physical layer. Each module is connected via interface and functioned under the control of event-driven engine. The whole system architecture and special interactions among modules are shown in Fig. 2. We have designed a 5G system level simulation platform dedicated to Cellular and WIFI heterogeneous networks in our previous work [9]. In this latest work, we use some modules in our previous work and the modules circled in Fig. 2 are newly designed and will be illustrated in the following section.

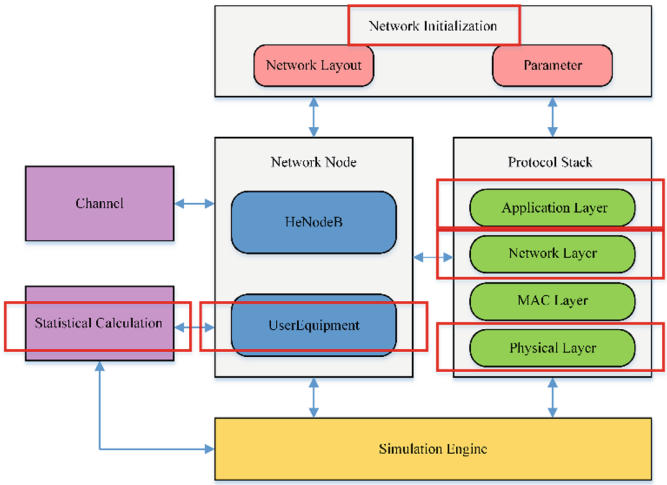


Fig. 2. The architecture of the D2D simulation platform

3 Implementation of Modules in D2D Simulation Platform

In this section, several modules and their functions of D2D communication system-level simulation platform are illustrated in detail.

3.1 Network Initialization

The main function of network initialization is to generate D2D users which can be classified as D2D transmit UE (source UE), D2D receive UE (destination UE) and D2D relay UE. For D2D communication, we first generate N source UEs with random distribution and each UE must select the way of wireless link, direct D2D link type or two hop D2D link type. If UE set a direct D2D link type, we generate a corresponding destination UE and ensure pairing for each UE while the distance between source UE and destination UE is uncertain because the coordinate of each UE is generated randomly. Besides, two hop D2D link type means source UE can only receive the message from the base station by relay UE. In order to forward the message from the base station to destination UE, we distribute some relay UE randomly in the simulation scenario.

3.2 Application Layer

Application layer provides various applications and we generate D2D traffic in application layer. A foreseen traffic model is defined by the METIS project to cope with the requirements of the year 2020 for D2D communication [10]. The main traffic model parameters for D2D mentioned in [10] are packet size with 20 Mbytes and inter packet delay with 20 ms.

As described in Table 1. Firstly, we initialize the data size and data arrive interval. Secondly, we generate the start time and end time of the traffic stream and create a empty list for storing data streams. Thirdly, we generate the streams until arrive the end time and each stream has a ID.

Table 1. D2D traffic generated algorithm

D2D Traffic Generated Algorithm	
1.	Initialize data size d=200000 and traffic arrive interval i=20000;
2.	Create a list of data streams and randomly generate the data arrive time and end time;
3.	**while** arrive time < end time
4.	Initialize j=0;
5.	Generate a data stream with id=j;
6.	Insert data stream to the data list;
7.	Update data arrive time according to the arrive interval;
8.	j=j+1;
9.	**End**;

3.3 Physical Layer

The resource allocation in physical layer is carried by base station in each slot time. There are two situations when referred to resource allocation, overlay D2D and underlay D2D. We will discuss two cases separately.

- Underlay D2D: D2D users and cellular users share same resource blocks(RBs) in underlay D2D. For this D2D simulation platform, D2D UEs use RBs allocated by base station with Round-Robin scheduling algorithm [11]. In direct D2D link, source UE is set as underlay D2D type and we apply a random algorithm to multiplex the whole spectrum of base station. Figure 3(a) shows the resource allocation in underlay D2D, each RB consists of 12 OFDM symbols, we can see that RB3 and RB4 are used by D2D and cellular users at the same time.
- Overlay D2D: Overlay D2D mode is applied in two hop D2D link when relay UE forward packets to destination UE. Relay UEs operate on spectrum resources orthogonal to cellular users in this situation and the Round-Robin scheduling algorithm is applied as well. If a RB is allocated to a cellular user, other D2D users can't use the same RB at the same time slot. As is shown in Fig. 3(b), RB1, RB3 and RB4 are used by cellular users while D2D users operate on RB2 and RB5. D2D and cellular users use different RBs in this mode.

We set up an open function interface for resource allocation in this platform which can apply self-defined resource allocation algorithm so we just deploy the most basic algorithm in the platform.

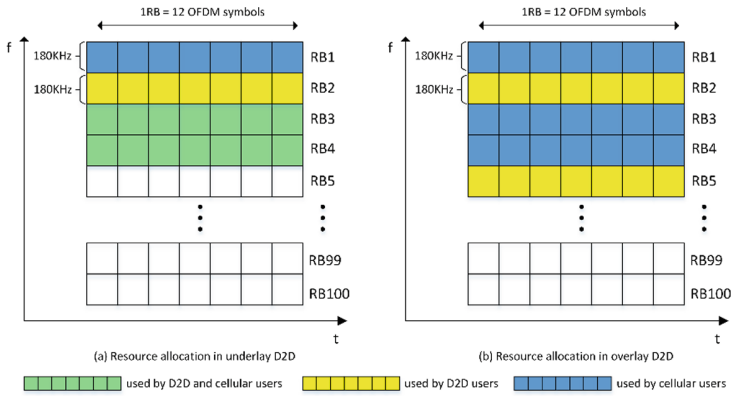


Fig. 3. Resource allocation scheme in D2D simulation platform

3.4 Power Control

A LTE uplink power control scheme is applied in this platform called open loop power control [12–14]. The D2D transmit power P is calculated as below:

$$P = \min\{P_{\max}, P_{noise} + SNR + \alpha \cdot PL\} \tag{1}$$

where $P_{\max}$ represents the maximum D2D transmit power, P_{noise} denotes the noise power level, SNR is a targeted received power level in relation to the noise floor between 0 dB and 30 dB, PL is the path loss between source UE and destination UE and α is a compensation factor defined in the LTE standard which is set to {0.4, 0.5, 0.6, 0.7, 0.8, 0.9}. We select $\alpha = 0.7$ in this power control scheme because it gives a good trade-off which has been discussed in [14]. In Fig. 4 the variance of the effective SINR is decreased than that without power control and the system performance gets better as well.

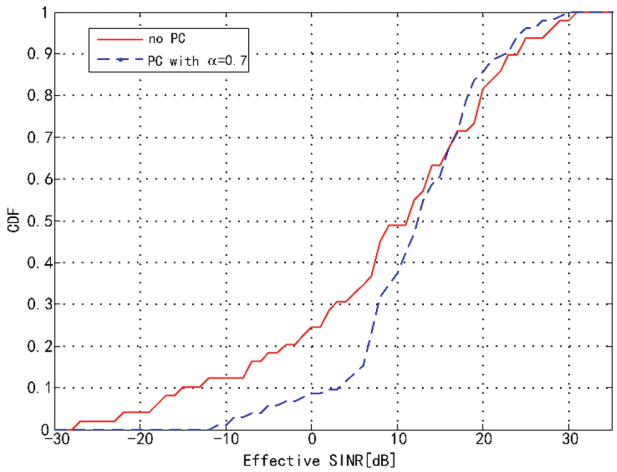


Fig. 4. CDF of effective SINR

3.5 Network Layer

The function of network layer is to select the appropriate route and an open function interface for routing is installed in this platform. We employ Dijkstra shortest path algorithm for routing [15]. The routing process can be summarized in two steps. First, a sorting procedure is carried out according to the power loss measured from base station to destination UE. Then, the worst link has the priority to choose relay with Dijkstra algorithm which can improve the system throughput at utmost.

Users can use the interface to design their own routing algorithm for performance testing.

3.6 Interference and SINR Measure

The fact that D2D communication reuses cellular frequency resources will cause serious cross interference. Communication between D2D pairs will interfere cellular communication, meanwhile, cellular communication interferes D2D communication. To compute interference and SINR, a family of methods known as MMSE receiver is used here during SINR measure [16, 17]. First, the cross interference of underlay D2D type can be depicted by Eq. 2. The first item represents the interference from other D2D link and the rest represents the interference from base station.

$$P_{cross} = \sum_{k=1}^{N_K} LH_k W_k P_k [LH_k W_k P_k]^H + \sum_{j=1}^{N_J} LH_j W_j P_j [LH_j W_j P_j]^H \tag{2}$$

where L represents the linear MMSE operator, H represents channel matrix between source and destination, W represents the precoding matrix, P represents the power allocation matrix, N_K is the number of D2D link and N_J is the number of base station. If D2D communication operates on overlay type, interference exists only between D2D links and is given in Eq. 3.

$$P_{cross} = \sum_{k=1}^{N_K} LH_k W_k P_k [LH_k W_k P_k]^H \tag{3}$$

Consequently, SINR in the i-th stream of the n-th sub-carrier is given in Eq. 4.

$$SINR_i = \frac{diag[DD^H]_i}{diag[\sigma^2 LL^H + I_{self} I_{self}^H + P_{cross}]_i} \tag{4}$$

where D represents the desired signal. The numerator of the equation represents the power of the desired signal, the first item of the denominator represents the power of noise, the second item of the denominator represents interference between streams and the last item of the denominator represents the cross interference in underlay type or the interference from other D2D link in overlay type.

4 Simulation Results

4.1 Simulation Scenario

METIS has described the identified scenario that is concerned with the connectivity required at any place and at any time by humans in dense urban environments, so we use the typical dense urban information society defined by METIS as shown in Fig. 5 [18, 19]. In this scenario, there are four buildings and each building has six floors. In addition, we deploy four HeNodeBs along the road. We deploy some UEs worked as D2D nodes randomly outside the road or inside the building. The scenario simulates the real city environment so the simulation results are close to the real city environment which provides reference value for us.

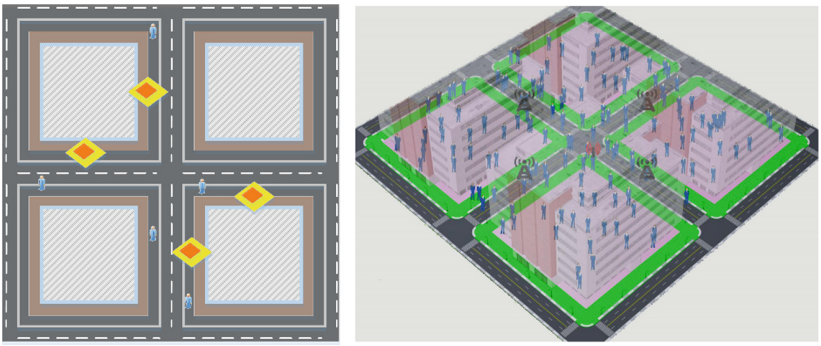


Fig. 5. Simulation scenario (a) plan graph of the scenario (b) 3D graph of the scenario

4.2 Simulation Parameters

We have taken the simulation parameters as shown in Table 2 [18, 19].

Table 2. Testing parameters

Parameters	Settings
HeNodeB Carrier Frequency	3.5 GHz
HeNodeB Bandwidth	20 MHz
D2D Carrier Frequency	3.5 GHz
D2D Bandwidth	20 MHz
HeNodeB/D2D Transmit Power	44 dBm/23 dBm
HeNodeB/D2D Antenna Height	3 m/1.5 m
HeNodeB number	4
D2D pairs number	variable
Penetration Loss [dB]	17 + 4*(n−1), n is number of floors
Channel Model	WINNER II B1
Traffic Model	D2D Traffic
Simulation Time	50TTI

4.3 Simulation Results

We use the above parameters for system level simulation test with different D2D pairs and basic scheme installed in this platform. The result of average throughput is shown in Fig. 6 and the corresponding CDF of average throughput is shown in Fig. 7. The average throughput is the sum of the average throughput of the direct D2D link and the two hop D2D link. The high average throughput of D2D communications in Fig. 6 declares D2D communication can lighten the burden of cellular network and improve spectral efficiency. It's obvious that D2D communication plays a pivotal role in solving peoples' needs in future 5G networks with the growing mobile terminals.

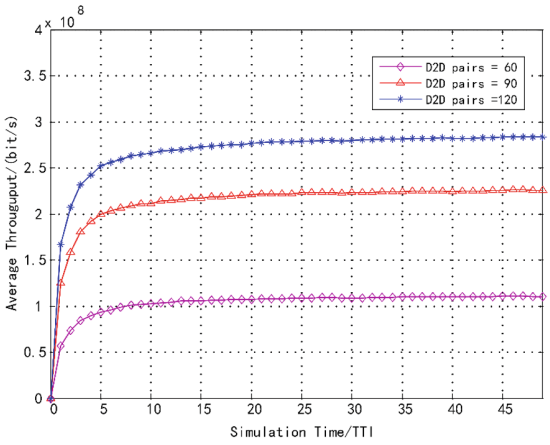


Fig. 6. D2D link average throughput

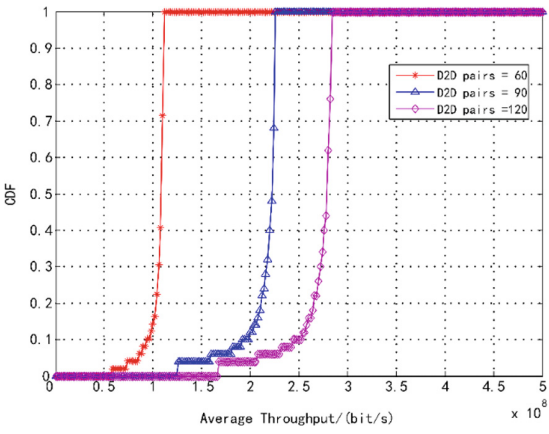


Fig. 7. CDF of D2D link average throughput

5 Conclusion

In this paper, we introduce the design and implement of D2D communication system-level simulation platform. This platform is designed under the guidance of METIS project and is oriented to future 5G network. In addition, we have emphasized on introducing several key technologies applied in this platform, including resource allocation, power control, interference and SINR measure and so on. Last but not least, we set open interface for academic research in D2D communication such as resource allocation interface, power control interface and routing interface.

References

1. IMT-2020 5G Promotion Group White Paper, 5G vision and requirements. http://www.imt-2020.org.cn/zh
2. Mishra, P., Pandey, S., Biswash, S.: Efficient resource management by exploiting D2D communication for 5G networks. IEEE Acess. **4**, 9910–9992 (2017)
3. Gandotra, P., Jha, R.K., Jain, S.: A survey on Device-to-Device (D2D) communication: architecture and security issues. J. Network Comput. Appl. **78**, 9–29 (2017)
4. Asadi, A., Wang, Q., Mancuso, V.: A survey on Device-to-Device communication in cellular networks. Commun. Surv. Tutor. **16**, 1801–1819 (2013)
5. Zhang, A., Chen, J., Hu, R.Q.: SeDS: Secure Data Sharing Strategy for D2D communication in LTE-advanced networks. IEEE Trans. Veh. Technol. **65**, 2659–2672 (2016)
6. Yin, R., Yu, G., Zhong, C., Zhang, Z.: Distributed resource allocation for D2D communication underlaying cellular networks. In: IEEE International Conference on Communications Workshops, pp. 138–143. IEEE Press, Budapest (2013)
7. Zhao, Y., Li, Y., Mao, H., Ge, N.: Social community aware long-range link establishment for multi-hop D2D communication networks. IEEE Trans. Veh. Technol. **65**, 9372–9385 (2016)
8. Babun, L., Yurekli, A.I., Guvenc, I.: Multi-hop and D2D communications for extending coverage in public safety scenarios. In: 40th IEEE Conference on Local Computer Networks, pp. 912–919. IEEE Press, Clearwater (2015)
9. Meng, L., Pinyi, R., Du, Q., Wenchao, O., Xinlei, X., Guobing, L.: Design of system-level simulation platform for 5G networks. In: IEEE International Conference on Communications in China, pp. 1–6. IEEE Press, Chengdu (2016)
10. Mobile and wireless communications Enablers for the Twenty-twenty Information Society. https://www.metis2020.com/
11. Arif, S., Ghaffar, N., Javed, A.: Implementation of alternating median based round robin scheduling algorithm. In: 16th IEEE International Conference on Computer and Information Technology, pp. 154–160. IEEE Press, Nadi (2016)
12. Wei, L., Papathanassiou, A., Li, Q.C., Wu, G.: System-level simulations for multi-hop D2D communications overlay LTE networks. In: International Conference on Computing. Networking and Communications, pp. 1–5. IEEE Press, Hawaii (2016)
13. Yang, Y., Quan, Q., Li, J., Wang, W.: System-level simulation platform with interference avoidance for Device-to-Device communication underlaying cellular system. In: 15th IEEE International Conference on Communication Technology, pp. 131–135. IEEE Press, Guilin (2013)

14. Simonsson, A., Furuskar, A.: Uplink power control in LTE - overview and performance, subtitle: principles and benefits of utilizing rather than compensating for SINR variations. In: 68th IEEE Vehicular Technology Conference, pp. 1–5. IEEE Press, Alberta (2008)
15. Broumi, S., Talea, M., Bakali, A., Smarandache, F.: Application of Dijkstra algorithm for solving interval valued neutrosophic shortest path problem. In: IEEE Symposium Series on Computational Intelligence, pp. 1–6. IEEE Press, Athens (2016)
16. Mehana, A.H., Nosratinia, A.: MMSE receive filtering for precoded MIMO systems. In: Asilomar Conference on Signals, Systems and Computers, pp. 2057–2061. IEEE Press, Pacific Grove (2013)
17. Bendlin, R., Joham, M., Nossek, J.A., Huang, Y.F.: Iterative MMSE transmit and receive filter design for frequency selective MU-MISO systems. In: IEEE International Conference on Communications, pp. 2739–2744. IEEE Press, Glasgow (2007)
18. Scenarios, requirements and KPIs for 5G mobile and wireless system. https://www.metis2020.com/
19. Initial channel models based on measurements. https://www.metis2020.com/



Joint Algorithm for MIMO-OFDM System Based on the Idea of Layered Iteration

Juan Wei, Shixuan Feng(✉), and Lihong Wang

School of Communication Engineering, Xidian University, Xi'an 710071, Shaanxi, People's Republic of China
1584623866@qq.com

Abstract. A new joint algorithm is proposed to resolve the contradiction between algorithm complexity and performance in multiple input multiple output-orthogonal frequency division multiplexing (MIMO-OFDM) system. Using the hierarchical structure of the MIMO-OFDM systems, channel estimation and signal detection are alternately performed iteratively. Firstly, to reduce the complexity of channel estimation algorithm, an improved minimum mean squared error (MMSE) channel estimation algorithm is proposed. Then, zero force-successive interference cancellation (ZF-SIC) and MMSE-SIC are used in the joint algorithm as signal detection algorithms. Simulation results show that the new algorithm can obtain good BER performance under single path QPSK and single path 16 QAM simulation conditions. Compared with the existing joint algorithms, the new algorithm can make full use of system's space resources and ensure the requirements of algorithm complexity and system performance.

Keywords: MIMO-OFDM · Channel estimation · Signal detection

1 Introduction

Multiple input multiple output-orthogonal frequency division multiplexing (MIMO-OFDM) has been universally acknowledged as a key technology of next generation wireless communication system by virtue of its advantages of high transmission rates, large capacity, high reliability, etc. Besides, it has been adopted in various international standards such as 3GPP-LTE, WiMAX, IMT-Advanced [1]. Joint channel estimation and signal detection algorithm has been the research trend in MIMO-OFDM systems owing to its good effects on simplifying the signal processing progress and improving the system integration. The research on joint channel estimation and signal detection algorithm for now can be mainly divided into the ideas based on QR decomposition [2] and spatial alternation [3].

On the premise of ensuring the system performance, the joint algorithm for MIMO-OFDM system based on the idea of layered iteration is proposed in this paper to reduce the complexity of the existing algorithms as far as possible. The channel estimation method is actually an improved MMSE channel estimation algorithm, which is based on Least Square (LS) estimation and MMSE estimation and utilizes discrete Fourier transform (DFT) and singular value decomposition (SVD). The signal detection algorithms are ZF and MMSE algorithms, and the successive interference cancellation (SIC) is also adopted to optimize the algorithms.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 106–113, 2018.
https://doi.org/10.1007/978-3-319-78130-3_12

2 Improved MMSE Channel Estimation Algorithms

There have been many studies on channel estimation of MIMO-OFDM systems [4–6]. In this paper, DFT technology and SVD transform are firstly used to reduce the complexity of the MMSE channel estimation algorithm [7].

MMSE estimation algorithm aims at minimizing the mean square error, and the expression in frequency domain can be written as

$$\hat{H}_{MMSE} = F\hat{h}_{MMSE} = FQ_{MMSE}F^{\mathrm{H}}X^{\mathrm{H}}Y \tag{1}$$

F contains the Fourier transform coefficients. Equation (1) can be further simplified as

$$\hat{H}_{MMSE} = R_{HH}\left[R_{HH} + \sigma_W^2\left(X^{\mathrm{H}}X\right)^{-1}\right]^{-1}\hat{H}_{LS} \tag{2}$$

Where R_{HH} denotes the channel autocorrelation matrix, $\hat{H}_{LS}$ represents the estimated value in frequency domain by LS algorithm, and it is given by

$$\tilde{H}_{LS} = X^{\dagger}Y = \left(X^{H}X\right)^{-1}X^{H}Y \tag{3}$$

In order to simplify the computation, $E\left\{\left(X^{\mathrm{H}}X\right)^{-1}\right\}$ is used to replace $\left(X^{\mathrm{H}}X\right)^{-1}$, which is called LMMSE estimation. So the estimated result based on LMMSE is

$$\tilde{H}_{LMMSE} = R_{HH}\left[R_{HH} + \frac{\beta}{SNR}I\right]^{-1}\tilde{H}_{LS} \tag{4}$$

Where $\beta = E\left\{|X_k|^2\right\} \cdot E\left\{|1/X_k|^2\right\}$.

The specific steps of improved MMSE estimation algorithm are as follows.

Step 1: By adopting Inverse DFT (IDFT), the channel estimation in frequency domain can be transformed into time domain.

$\hat{H}[k]$ denotes the channel estimation in frequency domain of the k^{th} subcarrier, and it can be obtained through Eqs. (3) and (4). Then adopting IDFT:

$$\mathrm{IDFT}\left\{\hat{H}[k]\right\} = h[n] + z[n] \triangleq \hat{h}[n], n = 0, 1, \cdots, N-1 \tag{5}$$

Where $z[n]$ denotes the noise component in time domain.

Step 2: Using linear transform in time domain, which means to do shortened operation on channel estimated value, then applying SVD transform on autocorrelation matrix of channel impulse response (CIR), which can reduce the matrix order.

Step 3: The channel estimated value in time domain is transformed back into the frequency domain by DFT, and the channel frequency response (CFR) can be obtained. This procedure can be seen as the separation of channel impulse response subspace and noise subspace.

By using DFT, $\hat{h}_{DFT}[n]$ is transformed to frequency domain, which means using DFT with N points to estimate the channel frequency response and realize interpolation:

$$\hat{H}_{DFT}[k] = \mathrm{DFT}\{\hat{h}_{DFT}[n]\} \tag{6}$$

The improved algorithm reduces the complexity without the bad effects on channel estimation performance. Besides, due to the merit that the channel energy is more concentrated in time domain, the bad effect on channel estimation which is caused by the noise beyond the maximum channel delay is eliminated by shortened operation, which can further improve the accuracy of channel estimation.

3 Signal Detection Algorithm

The best detection algorithm is ML detection, but the complexity grows with the number of transmitting antennas exponentially, which limits the application in reality. While the BER performance of ZF detection and MMSE detection are pretty good, and SIC technology can be adopted to form better performed algorithm, such as ZF-SIC and MMES-SIC [8].

QR decomposition is used in ZF algorithm. Assume that the channel is quasi-static, and $Nr \geq Nt$. The detection should begin at layer Nt according to SIC principle. $\breve{y}_{Nt}(k)$ can be given by

$$\breve{y}_{Nt}(k) = r_{Nt,Nt}s_{Nt}(k) + \breve{n}_{Nt}(k) \tag{7}$$

Where $r_{Nt,Nt}$ is the lower triangular matrix obtained through QR decomposition. In general, the statistical properties of soft decision in i^{th} layer, namely $\breve{y}_i(k)$ can be given by

$$\breve{y}_i(k) = r_{i,i}s_i(k) + \sum_{j=i+1}^{Nt} r_{j,i}\left(s_j(k) - \hat{s}_j(k)\right) + \breve{n}_i(k) \tag{8}$$

MMSE detection is based on the rule of minimum mean square error, and the Wiener solution $\boldsymbol{W}_{opt}$ can be given by

$$\boldsymbol{W}_{opt} = \boldsymbol{H}^{\mathrm{H}}\left[\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}} + \sigma^2\boldsymbol{I}_{Nt}\right]^{-1} \tag{9}$$

Where σ^2 is the variance of white Gaussian noise at receiver. $\boldsymbol{I}_{Nt}$ is unit matrix with $Nt \times Nt$ order. The decision value $\breve{\boldsymbol{Y}}$ can be obtained by $\boldsymbol{Y}$ right multiplying $\boldsymbol{W}_{opt}$, i.e.

$\breve{\boldsymbol{Y}} = \boldsymbol{Y}\boldsymbol{W}_{opt}$. The same as ZF detection, MMSE detection begins at the layer Nt, and $\breve{y}_{Nt}(k)$ can be given by

$$\breve{y}_{Nt}(k) = \sum_{j=1}^{Nr} w_{j,Nt}\left(s_{Nt}(k)h_{Nt,j} + \sum_{n=1}^{Nt-1} s_n(k)h_{n,j}\right) + \breve{n}_{Nt}(k) \tag{10}$$

Equation (10) shows that the signal at layer Nt will be interfered by other layers, but according to MMSE rule, the interference can be minimized. When detecting the layer $Nt-1$, the interference from layer Nt will be canceled. $\breve{y}_{Nt-1}(k)$ can be given by

$$\breve{y}_{Nt-1}(k) = \sum_{j=1}^{Nr} w_{j,Nt-1}\left[\underbrace{\sum_{n=1}^{Nt-2} s_n(k)h_{n,j}}_{\text{the first part}} + \underbrace{s_{Nt-1}(k)h_{Nt-1,j}}_{\text{the second part}} + \underbrace{(s_{Nt}(k) - \hat{s}_{Nt}(k))h_{Nt,j}}_{\text{the third part}}\right] + \breve{n}_{Nt-1}(k) \tag{11}$$

In Eq. (11), the first part in bracket represents the interference from undetected layers, the second part represents the detected signal in layer $Nt-1$, the third part represents the interference from the layer Nt that has already been detected. Thus, the interference from the layer Nt can be canceled before detecting the layer $Nt-1$, which refers to the update of the received matrix $\boldsymbol{Y}$. The update can be expressed as

$$\boldsymbol{Y}^{Nt-1} = \boldsymbol{Y}^{Nt} - \breve{\boldsymbol{s}}_{Nt}\boldsymbol{h}_{Nt} \tag{12}$$

Where $\breve{\boldsymbol{s}}_{Nt}\boldsymbol{h}_{Nt}$ is the interference from the layer Nt. Computing recursively and the signal detection of the whole system can be completed.

4 Joint Channel Estimation and Signal Detection Algorithm

The idea of the joint algorithm proposed in this paper is that channel estimation and signal detection are alternately performed iteratively. First, the initial estimated value can be obtained according to pilot signal, and the estimated value can be used for signal detection. Then updating the detected results, which can be used for the estimation of next layer, that is means hierarchical iteration.

According to the analysis above, there are 4 joint algorithms: LS-ZF-SIC, LMMSE-ZF-SIC, LS-MMSE-SIC, LMMSE-MMSE-SIC. The concrete steps are as follows.

Step 1: Initial estimation based on Eqs. (3) and (6), and the channel estimated value of the layer Nt can be obtained.

Step 2: Get the signal detected value in layer Nt through Eqs. (7) and (10). The channel estimation and signal detection in layer Nt are finished by now.

Step 3: Update the received matrix in layer Nt according to Eq. (12), then we can get the received matrix in layer $Nt-1$.

Step 4: Estimate the channel in layer $Nt-1$ according to the received matrix in layer $Nt-1$.

Step 5: Calculate the Eqs. (8) and (11) by using the channel estimates of layer $Nt - 1$, and the detected results of layer $Nt - 1$ can be obtained. Continue recycling, we can get joint algorithm for MIMO-OFDM system based on the idea of layered iteration, that is updating the received matrix according to the previous layer, and obtaining the channel estimated value by detected results.

The joint algorithm makes the best of the hierarchical structure of MIMO system. In the scheme, the estimated and detected results are all updated in each layer, which makes full use of the space resources in MIMO system. In the procedure, the complexity has been greatly reduced on the premise of ensuring the system performance. Besides, DFT and SVD are adopted in estimation and further reduce the complexity, SIC is used to improve the performance of the system in signal detection.

5 Simulation Results

The iterations of joint algorithm associates with the layer number of the MIMO system. Simulations in Figs. 1 and 2 assume a MIMO-OFDM system with V-BLAST layered structure and $Nt = Nr = 4$, so the iterations are 4. Single path QPSK and single path 16QAM are adopted. We use static Rayleigh fading channel whose coefficients keep the same in a subframe with 10 OFDM symbols, and the changes are independent between each subframe. The total number of subframe is 1000.

Figure 1 shows that under condition of single path QPSK modulation, the performance of MMSE detection is better than that of ZF detection, because the interference suppression in ZF algorithm is completely at the price of the noise enhancement. But the performance of LMMSE channel estimation is better than that of LS channel estimation, so the performance of LMMSE-ZF joint algorithm and LS-MMSE joint algorithm are approximate. LMMSE-MMSE joint algorithm performs best.

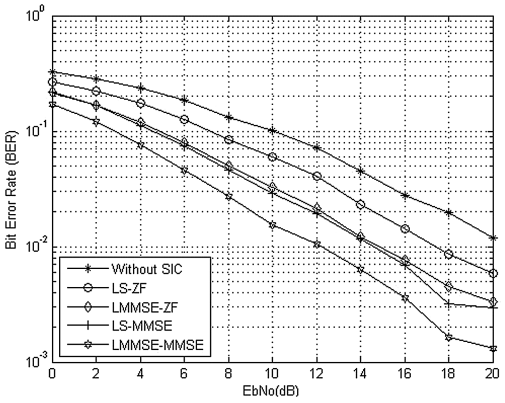


Fig. 1. Joint algorithm with single path QPSK

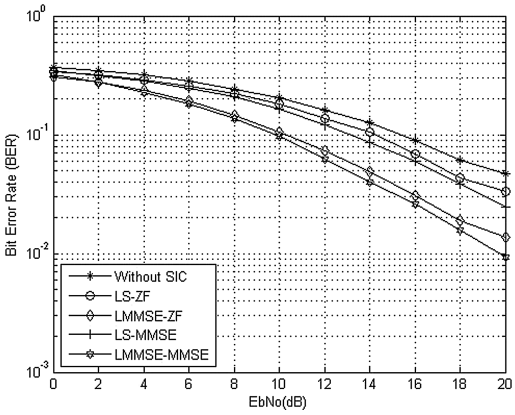


Fig. 2. Joint algorithm with single path 16-QAM

Figure 2 shows that under condition of 16-QAM modulation, the channel estimation has large effects on the system performance. The performance of LMMSE-ZF joint algorithm is close to that of LMMSE-MMSE joint algorithm, and the performance of LMMSE-ZF algorithm is better than that of LS-MMSE, which demonstrates that a better channel estimation algorithm could make up for performance reduction due to the signal detection algorithm to a certain extent.

Simulation conditions in Fig. 3 are as follows: single path QPSK modulation, the number of subcarrier is 1024, the subframe is formed by 10 OFDM symbols and 1 training symbol, space-time coding scheme is V-BLAST, $Nt = 4, Nr = 8$. The solid line in Fig. 3 represent the LS - ZF - SIC algorithm, and the dotted line represent existing algorithms.

Figure 3 shows that when adopting SAGE algorithm, the performance of joint algorithm proposed in this paper improves evidently. In general, ML algorithm is the best when the channel is ideal; SAGE algorithm is better than LS+ML and LMMSE +ML algorithms. When SNR = 8 dB, compared with LS+ML algorithm, the LMMSE +ML algorithm improves about 16 dB in BER performance; The performance of MAP-based SAGE+ML algorithm [9] and MAP-based SAGE+GS (Grover's Quantum Search) algorithm [10] improve about 15 dB than that of LMMSE+ML algorithm; The performance of LS-ZF-SIC joint algorithm improves about 15 dB over LS+ML algorithm, and is close to SAGE-based algorithms to some extent. Other joint algorithms proposed in this paper such as LMMSE-ZF-SIC, LS-MMSE-SIC and LMMSE-MMSE-SIC all have better performance than LS-ZF-SIC algorithm, which means the joint algorithm proposed in this paper can obtain better performance.

The search complexity of ML algorithm is $O(N)$, and GS algorithm is $O\left(\sqrt{N}\right)$. The joint algorithms proposed in this paper avoid the search operation of ML algorithm and GS algorithm, and the decision value of received signal can be obtained by W_{opt} directly. Besides, the complexity of LS and MMSE estimation algorithms are lower than that of MAP algorithm, so the total complexity can be further reduced, and it can be applied in engineering easily.

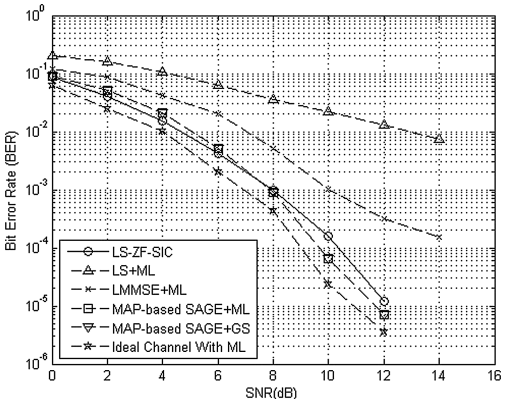


Fig. 3. Comparison of algorithms performance with 4×8 QPSK

6 Conclusion

In a MIMO-OFDM system, channel estimation and signal detection are jointed in the proposed algorithm, and channel estimation and signal detection are alternately performed iteratively. The improved MMSE channel estimation algorithm is adopted to reduce the computational complexity. ZF-SIC and MMSE-SIC are adopted for signal detection. So there are 4 joint algorithms: LS-ZF-SIC, LMMSE-ZF-SIC, LS-MMSE-SIC and LMMSE-MMSE-SIC. Simulation results show that in a MIMO-OFDM system with V-BLAST encoding structure, joint algorithms have better performance and are suitable for many kinds of modulation. Compared with the existing algorithms, the joint algorithm reduces the computational complexity on the premise of guarantee the system performance.

Acknowledgment. This study is supported by grants from the National Natural Science Foundation of China (Grants No. 51375385, 51675425), 111 Project (B08038).

References

1. Dixit, S., Katiyar, H.: Performance of OFDM in time selective multipath fading channel in 4G systems. In: 2015 Fifth International Conference on Communication Systems and Network Technologies (CSNT) (2015)
2. Yue, J., Kim, K.J., Gibson, J.D., et al.: Channel estimation and data detection for MIMO-OFDM systems. In: IEEE Global Telecommunications Conference, San Francisco, vol. 2, pp. 581–585 (2003)
3. Hedayati, M.K., Bakhshi, H., Cheraghi, M.: SAGE algorithm for semi-blind channel estimation and symbol detection for STBC MIMO OFDM systems. Wireless Pers. Commun. **71**(2), 1541–1555 (2013)
4. Yuan, Z., Zhang, C., Wang, Z.: An auxiliary variable-aided hybrid message passing approach to joint channel estimation and decoding for MIMO-OFDM. IEEE Sign. Process. Lett. **24**(1), 12–16 (2017)

5. Yu, J.L., Hong, D.Y.: A novel subspace channel estimation with fast convergence for ZP-OFDM systems. IEEE Trans. Wireless Commun. **10**(10), 3168–3173 (2011)
6. Shi, Q., Peng, C., Xu, W., Wang, Y.: Training signal design for MIMO channel estimation with correlated disturbance. In: Acoustics Speech and Signal Processing (ICASSP), pp. 5720–5724 (2014)
7. Moghaddam, S.S., Saremi, H.: A novel semi-blind channel estimation scheme for Rayleigh flat fading MIMO channels (joint LS estimation and ML detection). IETE J. Res. **56**(4), 193–201 (2010)
8. Hayakawa, R., Hayashi, K., Kaneko, M.: An overloaded MIMO signal detection scheme with slab decoding and lattice reduction. In: 2015 21st Asia-Pacific Conference on Communications (2015)
9. Shen, J., Wu, M.: A MAP-based SAGE channel estimation and data detection joint algorithm for MIMO system. In: IEEE 2nd International Conference on Software Engineering and Service Science, Beijing, pp. 82–85 (2011)
10. Shen, J., Wu, M., Luan, L.: Grover's quantum search based data detection algorithm for MIMO system. Int. J. Adv. Comput. Technol. **4**(1), 344–350 (2012)



Improved UWB Indoor Positioning Algorithms Based on BP Neural Network Model

Haojie Liu, Zhonghua Liang(✉), Danli Liu, and La'ning Ma

School of Information Engineering, Chang'an University,
Xi'an 710064, People's Republic of China
lzhxjd@hotmail.com

Abstract. Ultra-wideband (UWB) localization technique has been considered as a promising candidate for short-range positioning applications because of its advantages in terms of accuracy and penetrability. In this paper, backpropagation (BP) neural network is employed to improve the positioning accuracy of indoor environments. Theoretical analysis and simulation results show that the BP-based method outperforms other existing popular positioning algorithms with limited penalty in computational complexity. Therefore, it can be regarded as an alternative scheme for scenarios with the requirement of high position accuracy.

Keywords: Ultra-wideband (UWB) · Localization
BP neural network

1 Introduction

With the development of wireless communication technology, wireless localization technologies have been widely used in both military and civil applications. At present, the most popular positioning technology is mainly Global Position System (GPS). However, the GPS has two major disadvantages: (1) positioning accuracy can only achieve the accuracy of meter level; (2) the GPS signal cannot be received in completely closed indoor or underground environments, which makes GPS indoor positioning accuracy is low in non-line-of-sight environment. Ultra-wideband (UWB) as a potential short-range wireless transmission technology, with its unique advantages, is being increasingly applied to the field of wireless communications and positioning. And ultra-wideband (UWB) technology has fine time resolution and the positioning accuracy of centimeter-level [1].

Z. Liang—This work was supported in part by the National Natural Science Foundation of China under Grant No. 61271262, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2017JM6099, and in part by the Fundamental Research Funds for the Central University under Grant 310824171004.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 114–124, 2018.
https://doi.org/10.1007/978-3-319-78130-3_13

UWB signals occupies wide bandwidths and therefore some low-frequency components can penetrate the obstacle. This makes the UWB positioning feasible in a non-line of-sight (NLOS) environment. UWB indoor positioning methods usually utilize one or more of the following three categories of information to perform position estimation: angle of arrival (AOA), received signal strength indication (RSSI) and time of arrival (TOA). Recent research works show that a single indoor localization algorithm has been unable to meet the increasing needs of positioning accuracy. Therefore, hybrid positioning algorithms become more and more attractive [2].

The positioning equations are usually nonlinear equations based on TOA or time difference of arrival (TDOA) measurements, and therefore they need to be converted into linear equations for solutions. The least squares(LS) localization algorithm was presented in [3], but the solution are suboptimal. Taylor series expansion algorithm is an effective method to solve nonlinear equations with high accuracy and fast convergence. But its disadvantage is the need for a reliable initial position to guarantee the convergence. Chan algorithm only has the best estimation performance when the measurement error can be modeled as the zero-mean additive white Gaussian noise. In multipath environments, the performance of the Chan algorithm will decreases rapidly with the increase of TDOA measurement error. Previous works [10–12] demonstrated that the positioning accuracy can be improved via neural network based algorithms. Therefore, based on the discussions above, the backpropagation (BP) neural network is introduced in this paper to train the TOA values to improve the positioning accuracy.

The reminder of this paper is arranged as follows. Section 2 presents the UWB positioning system model, in which time-hopping (TH) pulse position modulation (PPM) signaling is employed. Section 3 introduces three popular localization algorithm and derives the proposed positioning method. Simulation results and the corresponding discussions are presented in Sect. 4. Finally, conclusion remark is given in Sect. 5.

2 PPM-TH-UWB Positioning System Model

In this section, we introduce the signal model, channel model and the receiving model, respectively.

2.1 Signal Model

In this paper, the PPM-TH-UWB signaling is used as the transmitting signal. The TH-PPM signal can be expressed as [4]

$$S(t) = \sum_{j=-\infty}^{+\infty} p(t - jT_s - C_j T_c - a_j \varepsilon), \tag{1}$$

where T_s is the pulse duration, C_j means the j-th symbol's TH codes, T_c represents the chip period, a_j denotes the binary sequence after repeated encoding, ε shows PPM displacement length.

2.2 IEEE 802.15.4a Channel Model

According to IEEE 802.15.4a standard, the generic UWB channel model can be described as [5]

$$h(t) = \sum_{k=0}^{K-1} \alpha_k \delta(t - \lambda_k), \tag{2}$$

where K is the total number of multi-path components, α_k and λ_k denote the complex amplitude and delay of the k-th multipath signals, respectively.

2.3 Receiver Model

There are two main signal reception schemes: energy receiver method and correlation receiver method [6–8]. The correlation receiver is employed in this paper. Where the time of peak value corresponding to the cross-correlation between the received signal and the template signal is the propagation delay of the signal, which is defined as

$$\tau = argmax_\tau \int_{T_d} r(t)s(t-\tau)d(t), \tag{3}$$

where T_d denotes the integral detection step.

3 Positioning Algorithms

3.1 LS Algorithm

The basic model of LS location algorithm is shown in Fig. 1 in 2-D space. D_1, D_2, D_3 are three reference base stations, c is the speed of light, and t_i denotes the unknown TOA parameter at base station i to be estimated. So the distance between reference base station i and unknown node is given by $d_i = ct_i$. Equations are written as [9]

$$\begin{Bmatrix} \sqrt{(x_1-x)^2+(y_1-y)^2} \\ \sqrt{(x_2-x)^2+(y_2-y)^2} \\ \vdots \\ \sqrt{(x_k-x)^2+(y_k-y)^2} \end{Bmatrix} = \begin{Bmatrix} d_1 \\ d_2 \\ \vdots \\ d_k \end{Bmatrix}, \tag{4}$$

where (x, y) are the unknown position coordinates, (x_k, y_k) are the known coordinates of reference base station k, and d_k is the distance from the unknown position to k-th base station. Equation (4) can be rewritten as

$$AI = b, \tag{5}$$

where

$$A = -2 \begin{bmatrix} X_1 - X_K & Y_1 - Y_K \\ X_2 - X_K & Y_2 - Y_K \\ \vdots & \vdots \\ X_{K-1} - X_K & Y_{K-1} - Y_K \end{bmatrix},$$

$$I = \begin{bmatrix} X & Y \end{bmatrix}^{\mathrm{T}},$$

$$b = \begin{bmatrix} D_1^2 - D_K^2 - X_1^2 + X_K^2 - Y_1^2 + Y_K^2 \\ D_2^2 - D_K^2 - X_2^2 + X_K^2 - Y_2^2 + Y_K^2 \\ \vdots \\ D_{K-1}^2 - D_K^2 - X_{K-1}^2 + X_K^2 - Y_{K-1}^2 + Y_K^2 \end{bmatrix}.$$

The target location can be obtained by

$$I = A^{-1}b. \tag{6}$$

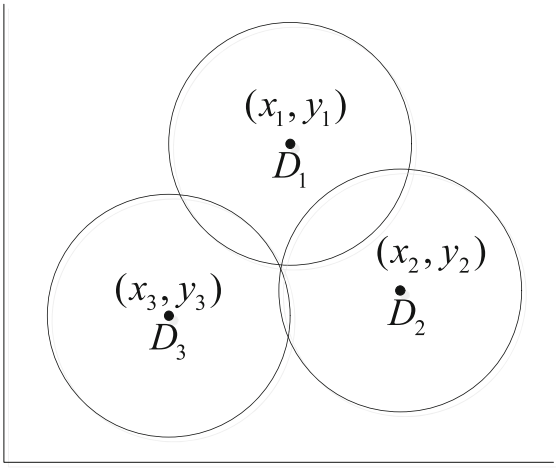


Fig. 1. LS algorithm schematics.

3.2 Chan Algorithm

Chan positioning algorithm is based on TDOA, and it is characterized by low computation complexity and high positioning accuracy in the case that the measurement error follows Gaussian distribution. Assuming there are m reference

base stations in the 2-D space, the distance from mobile station(MS) (x, y) to the i-th reference base station is r_i, (x_i, y_i) is the coordinate of the i-th reference base station, where $\Delta r_i = r_i - r_1$, $\Delta x_i = x_i - x_1$, $\Delta y_i = y_i - y_1$, $k_i = x_i^2 + y_i^2$. In line-of-sight (LOS) scenarios, Chan algorithm is presented in more detail as follows:

Step 1.
When MS is far away from reference base station, its positioning estimation is

$$Z_a^0 = \left[x \ y \ r_1^{'} \right]^{\mathrm{T}} \approx (G_a^T Q^{-1} G_a)^{-1} G_a^T Q^{-1} h, \tag{7}$$

where

$$G_a = - \begin{bmatrix} \Delta x_2 & \Delta y_2 & \Delta r_2 \\ \Delta x_3 & \Delta y_3 & \Delta r_3 \\ \vdots & \vdots & \vdots \\ \Delta x_m & \Delta y_m & \Delta r_m \end{bmatrix},$$

$$h = \frac{1}{2} \begin{bmatrix} \Delta r_2^2 - k_2 + k_1 \\ \Delta r_3^2 - k_3 + k_1 \\ \vdots \\ \Delta r_m^2 - k_m + k_1 \end{bmatrix}, n = \begin{bmatrix} \Delta n_2 \\ \Delta n_3 \\ \vdots \\ \Delta n_m \end{bmatrix},$$

and

$$Q = E(n n^{\mathrm{T}}) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

represents covariance matrix of the measurement errors with TDOA noise.

Step 2:
Weighted least squares (WLS) is performed to get the initial position coordinates as follows:

$$Z_a^1 = (G_a^T \mu^{-1} G_a)^{-1} G_a^T \mu^{-1} h, \tag{8}$$

where $\mu = BQB$, $B = diag(r_2^{'}, \cdots, r_m^{'})$, $r_i^{'} = Z_a^0(3) + \Delta r_i$.

Step 3:
The second WLS estimation is performed to get an improved result

$$Z_a^2 = (G_a^{'T} B^{'-1} G_a^T Q^{-1} G_a B^{'-1} G_a^{'})^{-1} \times (G_a^{'T} B^{'-1} G_a^T Q^{-1} G_a B^{'-1}) h^{'}, \tag{9}$$

where

$$B^{'} = \begin{bmatrix} Z_a^1(1) - x_1 & 0 & 0 \\ 0 & Z_a^1(2) - y_1 & 0 \\ 0 & 0 & Z_a^1(3) \end{bmatrix},$$

$$h^{'} = \begin{bmatrix} (Z_a^1(1) - x_1)^2 \\ (Z_a^1(2) - y_1)^2 \\ Z_a^1(3)^2 \end{bmatrix},$$

$$G_a^{'} = \begin{bmatrix} 1\ 0 \\ 0\ 1 \\ 1\ 1 \end{bmatrix}, Z_a^2 = \begin{bmatrix} (x - x_1)^2 \\ (y - y_1)^2 \end{bmatrix}.$$

In the above WLS operation, it is assumed that the measurement error of TDOA follows Gauss distribution, otherwise the location error will be increased significantly. Finally, the target position can be solved as

$$Z_p = \pm\sqrt{Z_a^2} + [x_1\ y_1]^T. \tag{10}$$

3.3 Taylor Algorithm

Each recursion of the Taylor series expansion based algorithm is performed to improve the estimation position of the mobile station (MS) based on measurement error of TDOA. Assume unknown location of the MS is (x, y), and (x_i, y_i) is the known coordinate of base station i. In the Taylor location algorithm, a set of nonlinear equations is linearized by initialing an estimate of actual position. According to the TOA estimate τ_i, we have

$$\begin{aligned} X^{'} &= (x_v, y_v) \\ &= \frac{\frac{1}{\tau_1}(x_1, y_1) + \frac{1}{\tau_2}(x_2, y_2) + \cdots + \frac{1}{\tau_n}(x_i, y_i)}{\frac{1}{\tau_1} + \cdots + \frac{1}{\tau_n}}. \end{aligned} \tag{11}$$

Define

$$\begin{aligned} f_i = &\sqrt{(x - x_{i+1})^2 + (y - y_{i+1})^2} - \\ &\sqrt{(x - x_1)^2 + (y - y_1)^2}, i = 1, 2, \cdots, N - 1, \end{aligned} \tag{12}$$

and let t_i be the TOA estimate at base station i. Thus

$$f_i(x, y) = d_{i+1,1} + \varepsilon_{i+1,1}, i = 1, 2, \cdots, N - 1, \tag{13}$$

where $d_{i+1,1} = c(t_{i+1} - t_1)$ and $\varepsilon_{i,1}$ are the corresponding range difference estimation error with covariance R. According to initial estimate X, then

$$x = x_v + \delta_x, y = y_v + \delta_y, \tag{14}$$

where δ_x and δ_y are the corresponding position errors. Expanding f_i and retaining the terms below second orders gives

$$f_{i,v} + a_{i,1}\delta_x + a_{i,2}\delta_y \approx d_{i+1,1} + \varepsilon_{i+1,1}, \quad i = 1, 2, \cdots, N - 1, \tag{15}$$

where

$$f_{i,v} = f_i(x_v, y_v), \tag{16}$$

$$a_{i,1} = \left.\frac{\partial f_i}{\partial x}\right|_{x_v, y_v} = \frac{x_1 - x_v}{\hat{d}_1} - \frac{x_{i+1} - x_v}{\hat{d_{i+1}}}, \tag{17}$$

$$a_{i,2} = \left.\frac{\partial f_i}{\partial y}\right|_{x_v, y_v} = \frac{y_1 - y_v}{\hat{d}_1} - \frac{y_{i+1} - y_v}{\hat{d_{i+1}}}, \tag{18}$$

$$\hat{d}_i = \sqrt{(x_v - x_i)^2 + (y_v - y_i)^2}. \tag{19}$$

Equation (15) can be rewritten as

$$A\delta = D + e, \tag{20}$$

where

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \\ \vdots & \vdots \\ a_{N-1,1} & a_{N-1,2} \end{bmatrix}, \delta = \begin{bmatrix} \delta_x \\ \delta_y \end{bmatrix},$$

$$D = \begin{bmatrix} \hat{d_{2,1}} - f_{1,v} \\ \hat{d_{3,1}} - f_{2,v} \\ \vdots \\ \hat{d_{N,1}} - f_{N-1,v} \end{bmatrix}, e = \begin{bmatrix} \varepsilon_{2,1} \\ \varepsilon_{3,1} \\ \vdots \\ \varepsilon_{N,1} \end{bmatrix}.$$

The weighted least-square estimator for (20) produces

$$\delta = [A^T R^{-1} A]^{-1} A^T R^{-1} D. \tag{21}$$

The process is repeated until the stopping criterion is satisfied, otherwise this estimate is defined as

$$X^{'} = X^{'} + \delta. \tag{22}$$

3.4 BP Neural Network Model

The inaccuracy of TOA estimation may lead to significant error in positioning. In this paper, BP neural network [10–12] is employed to mitigate the effect in location algorithms, BP neural network model is composed of two modules:

(1) *Data preparing module:* If the positions of reference base station are fixed, we randomly generated 1000 groups of the positions of unknown node for three positioning models in the range of 10 m × 10 m, then we will get 1000 groups of TOA.
(2) *Updating the weight module:* This module uses three-layer network to update the weights, the structure in Fig. 2 is adopted:

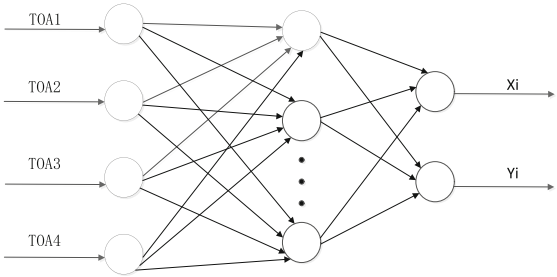


Fig. 2. BP neural network structure for training.

Input layer: The coordinate of predicted values as input values each model, among 80% were used for the training set, and 20% as the test set. In this paper, the number of neurons is 6.

Hidden layer: In the design of neural networks, the number of nodes has an impact on the performance of the neural network in the hidden layer. Generally speaking, the more complicated architecture of the network is, the more accurate the output result will be. Considering the training time of the network and accuracy of the compromise approach, in this paper, the number of neurons is 8. In order to approach any precision non-linear function, the transfer function of the hidden layer is set to Tanh-Sigmoid activation function.

Output layer: The output layer is composed 2 nodes corresponding to the positioning of the 2-D coordinates.

4 Simulation Results and Discussions

We use coordinate of predicted values and its true position to train BP neural network at first and then we generate a group of new position of unknown node to simulate. Among them, CM1 represents line-of-sight (LOS) indoor residential environments, CM2 denotes non-line of sight indoor residential environments.

Figure 3 presents the root mean square error (RMSE) of the existing three popular positioning algorithms with different SNR values in CM1. With the increase of SNR values, the positioning accuracy of several algorithms has been improved. When the SNR values is greater than 0 dB, the curve tends to be stable. It shows that SNR has little influence on the positioning accuracy when SNR is high. The positioning accuracy of the chan and Taylor algorithms can be less than 10 cm which meet the requirements of high-precision positioning.

Figure 4 shows the impact of SNR values on the RMSE performance with BP neural network in CM1 and CM2 channels, respectively. As the SNR increases, BP neural network has a better performance in CM1 channel compared with Fig. 3. Even though in CM2 channel with serious NLOS environment, RMSE value is basically stable in 18 cm.

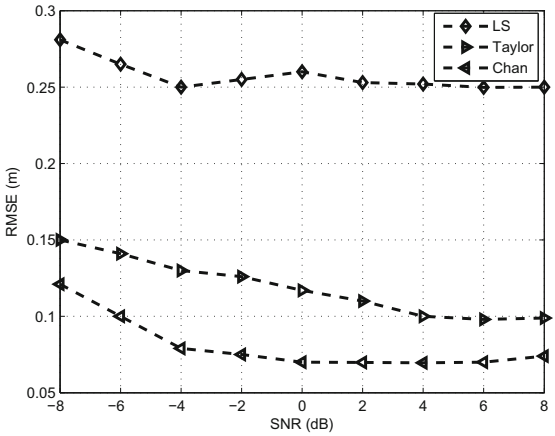


Fig. 3. RMSE vs SNR for positioning results with traditional algorithm.

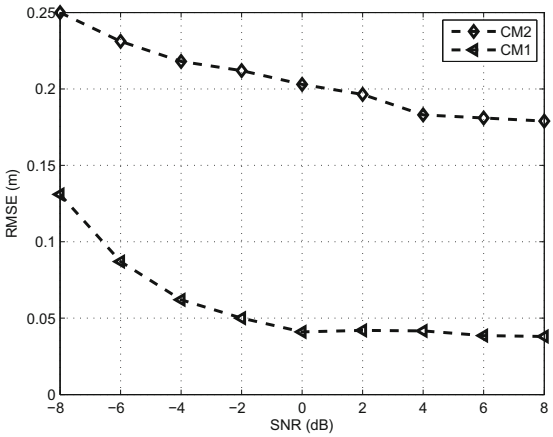


Fig. 4. Impact of SNR on the RMSE performance with BP neural network in CM1 and CM2 channels.

We present the RMSE of BP neural network with the different number of training in Fig. 5. We see that the RMSE will decrease gradually when the number of training increases, the network accuracy of positioning reaches the best at 800 epochs. If the training process is too long, it will be likely to make the model overfitting.

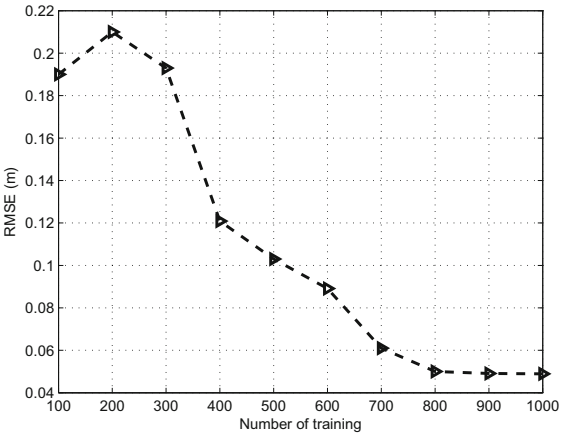


Fig. 5. RMSE of BP neural network with the different number of training.

5 Conclusions

In this paper, we investigated the existing three popular positioning algorithms in 2-D space. Both analysis and simulation results have shown that the BP neural network model can significantly reduce the RMSE of indoor environment. After sufficient training, the BP neural network can reach at sufficient accuracy of the positioning. Nevertheless, computational complexity should be considered in practical applications.

References

1. Li, H., Ren, B.: Wireless location for indoor based on UWB. In: Proceedings of Chinese Control Conference, pp. 6430–6433 (2015)
2. He, K., Zhang, Y., Zhu, Y., Xia, W., Jia, Z., Shen, L.: A hybrid indoor positioning system based on UWB and inertial navigation. In: Proceedings of IEEE International Conference on Wireless Communications and Signal Processing, pp. 1–5, October 2015
3. Kang, J., Lee, S., Kang, M.K., Park, Y.J., Kim, K.: Weighted-Beacon least square positioning method based on measurement. In: Proceedings of 2nd International Conference on Ubiquitous and Future Networks, pp. 55–59 (2010)
4. Wu, W., Wu, Z., Xie, W.: UWB PPM-TH and PAM-DS system with time reversal and its improved solution. In: Proceedings of IEEE International Conference on Information and Automation for Sustainability, pp. 332–336 (2012)
5. Molisch, A.F., Balakrishnan, K., Chong, C.-C., Emami, S., Fort, A., Karedal, J., Kunisch, J., Schantz, H., Schuster, U., Siwiak, K.: IEEE 802.15.4a channel model - final report, IEEE P802.15-04-0662-00-004a. Technical report, October 2004
6. Guvenc, I., Sahinoglu, Z.: Threshold selection for UWB TOA estimation based on kurtosis analysis. IEEE Commun. Lett. **9**, 1025–1027 (2005)
7. Gezici, S., Poor, H.V.: Position estimation via ultra-wide-band signals. Proc. IEEE **97**, 386–403 (2009)

8. Guvenc, I., Sahinoglu, Z.: Threshold-based TOA estimation for impulse radio UWB systems. In: Proceedings of IEEE International Conference on Ultra-wideband, pp. 420–425 (2005)
9. Wu, D., Bao, L., Li, R., Zeng, F.: Fast localization using robust UWB coding in wireless sensor networks. In: Proceedings of MSN' International Conference on Mobile Ad-Hoc and Sensor Networks, pp. 319–326, December 2009
10. Jie, D., Cui, X., Zhang, H., Wang, G.: A ultra-wideband location algorithm based on neural network. In: Proceedings of IEEE Wireless Communications, Networking and Mobile Computing, pp. 1–4 (2010)
11. Ergut, S., Rao, R.R., Dural, O., Sahinoglu, Z.: Localization via TDOA in a UWB sensor network using Neural Networks. In: Proceedings of IEEE International Conference on Communications, pp. 2398–2403, May 2008
12. Wang, X.: A UWB positioning scheme based on BP neural network in NLOS environment. In: Proceedings of IEEE International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1–3 (2011)



Performance of Transmitted Reference UWB Communication System with Time-Hopping Multiple Access

Danli Liu, Wenjing Wu, Zhonghua Liang(✉), Jinjin Liu, and Xiaojun Yang

School of Information Engineering, Chang'an University,
Xi'an 710064, People's Republic of China
lzhxjd@hotmail.com

Abstract. Transmitted reference (TR) system has become an attractive non-coherent ultra-wideband (UWB) communications due to its simple structure, no need for channel estimation, low synchronization requirements, and robustness. In this paper, time-hopping (TH) technology is introduced to the TR-UWB system, in which each frame within the transmitted signal for one symbol is divided into several non-overlapping time slots (or chips). In order to guarantee reliable communications and mitigate multi-access interference (MAI), it is necessary to employ TH sequence with good performance. Therefore, as two categories of classical TH sequences, m and Gold sequences are used to realize multiple access communications. Theoretical analysis and simulation results indicate that the bit error rate (BER) performance of the Gold sequence is better than the m sequence. For the same TH sequence, the BER performance degrades with the increase of number of users. Furthermore, the BER performance can be improved by increasing the number of elements of a TH sequence, at the cost of information transmission rate.

Keywords: Ultra-wideband (UWB) · Time-hopping (TH)
m sequence · Gold sequence

1 Introduction

Ultra-wideband (UWB) radio technology has attracted much attention in the academic and industrial communities, due to its low power consumption, high data rate and strong penetration [1,2]. The main difference between UWB technology and other wireless technologies are the extremely wide bandwidth occupied by transmitted signals composed of nanosecond ultra-narrow pulses [3].

Z. Liang—This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61271262 and 61473047, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2017JM6099, and in part by the Fundamental Research Funds for the Central University under Grant 310824171004.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 125–134, 2018.
https://doi.org/10.1007/978-3-319-78130-3_14

Many scholars have conducted a lot of research on the multiple access performance with transmitted reference (TR-) UWB system [4]. In the TR-UWB system, a pair of reference and modulated data pulses are used to transmit one data symbol [5]. The TR-UWB system usually uses a simple structure of delayed-multiply autocorrelation receiver (AcR), which can collect multi-path components without channel estimation, and requires lower synchronization [6]. The autocorrelation receiver (AcR) has been widely considered as a popular non-coherent receiver due to its simple implementation and robust performance in multi-path channels [7].

Time-hopping (TH) is a code division multiple access (CDMA) technology, in which each user can access the channel a unique TH code assigned [8]. The multi-access interference (MAI) can be mitigated when orthogonal TH codes are used. But in the presence of asynchronous users, the interference remains significant [9]. Since using spread spectrum technology can suppress interference and achieve multiple access communication, the choice of spread spectrum code sequences directly affect the performance of systems, such as capacity, anti-interference ability, access and switching speed. Therefore, the generation and performance evaluation of spread spectrum code is the key technology in spread spectrum communication system. The time-hopping (TH) spread spectrum technology is usually used for pulse position modulation (PPM) UWB system [10]. However, compared with PPM and on-off keying (OOK) modulation schemes, the binary phase shift keying (BPSK) modulation has better spectral characteristics and better BER performance [11].

In this paper, in order to guarantee reliable communications and mitigate multi-access interference (MAI), the TH technology with BPSK modulation is introduced to TR-UWB system to realize multi-access communications. In the TH multiple access UWB communication system, time slots are allocated to each user using different TH sequences to control the transmission time of the pulse [12], so that different users can share the same frequency band. However, due to the incomplete orthogonality of TH sequences, there will still exist MAI [13]. Therefore, take the m and Gold sequences to realize multiple access communications and mitigate MAI.

The rest of this paper is organized as follows. Section 2 presents the system models of the TH-TR-UWB system. In Sect. 3, the time-hopping sequence selection is presented. Both simulation results and discussions are presented in Sect. 4. Finally, Sect. 5 gives the summary.

2 System Models

In the TH-TR-UWB systems, the multiple access technology is implemented by TH sequence, and it requires that the transmitter and the receiver use the same sequence. In this section, we introduce the signal model, IEEE 802.15.4a channel model and TH-TR receiver model, respectively.

2.1 Signal Model

The conventional TR system sends only one pair of pulses in each frame, we divide each frame into N_c time slots in this paper, each time slot sends a pulse pair, that is to achieve different users share the same frequency band. Consider multi-user system, the k-th user's transmitted signal can be expressed as [14]

$$s^{(k)}(t) = \sum_{i=-\infty}^{\infty} \left[g(t - iT_f - c_i^{(k)} T_c) + b_{\lfloor i/N_s \rfloor} g(t - iT_f - c_i^{(k)} T_c - T_d) \right], \quad (1)$$

where $g(t)$ is the transmitted pulse signal, T_f is the frame period, T_c is the time of a single time slot, $c_i^{(k)}$ represent the TH sequence (user k) in each frame corresponds to the position, and satisfy $0 \leq c_i^{(k)} \leq N_c - 1 (N_c \geq k)$. N_s is the number of frames used to transmit one bit information, $b_{\lfloor i/N_s \rfloor} \in \{-1, 1\}$ represents the information of the $b_k(i)$ after BPSK modulation in the i-th frame. When the k-th user transmit the i-th pulse pair, the additional time shift of the TH sequence to this pulse pair is $c_i^{(k)} T_c$. T_d is the delay between the reference pulse and the modulated data pulse. In order to avoid the reference pulse and the data pulses overlap each other, be sure to $T_d \gg T_p$, and T_d must be greater than the multi-path delay spread T_{mds} to prevent inter-symbol interference.

2.2 IEEE 802.15.4a Channel Model

The equivalent complex baseband impulse response of the multi-user IEEE channel model can be expressed as [15]

$$h^{(k)}(t) = \sum_{l=1}^{L} \alpha_l^{(k)} \delta(t - \tau_l^{(k)}), \quad (2)$$

where k represents the number of users, L is the total number of multi-path components, $\alpha_l^{(k)}$ and $\tau_l^{(k)}$ denote the complex fading coefficient and delay component of the l-th multipath signals, respectively. In general, we assume that the channel energy is normalized, i.e., $\sum_{l=1}^{L} \left[\alpha_l^{(k)}\right]^2 = 1$.

2.3 TH-TR Receiver Model

In multi-user TH-TR system, the transmitter uses the TH sequence to transmit the signal, through the UWB channel, and then after low-pass filter to filter out the noise, the signal received by the receiver is

$$\begin{aligned} r(t) &= \sum_{k=1}^{N} s^{(k)}(t) * h^{(k)}(t) + n(t) \\ &= \sum_{k=1}^{N} \tilde{s}^{(k)}(t) + n(t), \end{aligned} \quad (3)$$

where

$$\begin{aligned}\tilde{s}^{(k)}(t) = &\sum_{i=-\infty}^{\infty} \big[g(t - iT_f - c_i^{(k)}T_c) + b_l^{(k)} g(t - iT_f \\ &- c_i^{(k)}T_c - T_d)\big] * \sum_{l=1}^{L} \alpha_l^{(k)} \delta(t - \tau_l^{(k)}) \\ = &\sum_{l=1}^{L} \alpha_l^{(k)} \sum_{i=-\infty}^{\infty} \big[g(t - iT_f - c_i^{(k)}T_c - \tau_l^{(k)}) \\ &+ b_l^{(k)} g(t - iT_f - c_i^{(k)}T_c - T_d) - \tau_l^{(k)}\big],\end{aligned} \tag{4}$$

where $*$ represents linear convolution and the noise term $n(t)$ is the filtered complex zero-mean additive white Gaussian noise (AWGN) with power spectral density $N_0/2$. The receiver uses the TH sequence of each user to delay $r(t)$, then extract its reference pulse and data pulse carry on autocorrelation to get the corresponding decision variable $y_k(i)$, it can be expressed as:

$$y^{(k)}(i) = \sum_{j=0}^{N_s-1} \int_{ijT_f + T_d + c_i^{(k)}T_c}^{ijT_f + T_d + c_i^{(k)}T_c + T} r(t) r(t - T_d) dt, \tag{5}$$

where T is the integration time of integrator. In order to collect signal energy as much as possible, it must be set as $T \geq T_{mds}$. Accordingly, the received symbol is:

$$b^{(k)}(i) = sgn(y^{(k)}), \tag{6}$$

where $sgn(\cdot)$ represents sign function, and it is defined as

$$sgn(x) = \begin{cases} +1, \text{x} \geq 0 \\ -1, \text{x} < 0. \end{cases} \tag{7}$$

3 Time-Hopping Sequence Selection

In the TH-TR-UWB communication system, each user signal has no fixed time slot, and the slot selection is randomly determined by the TH sequence during the communication process. The TH sequence is the key to realize the TH-TR-UWB system. Using the TH sequence degrades the power spectral density of the signal and increases the anti-jamming capability of the system. At the same time, the TH sequence is also the way for the UWB system to realize the multiple access communication. And the number of users that the system can accommodate determined by the number of the TH sequence. So the performance of the TH sequence is directly related to the performance of the entire system. In this section, m and Gold sequences are used to realize multiple access communication.

3.1 Using m Sequence to Achieve Time-Hopping Code

The m sequence is an important pseudo-random sequence, which is easy to generate and reproduce, and it is widely used in spread spectrum communication and CDMA technology. The m sequence is generated by a linear shift register (LSR), its structure can be shown in Fig. 1.

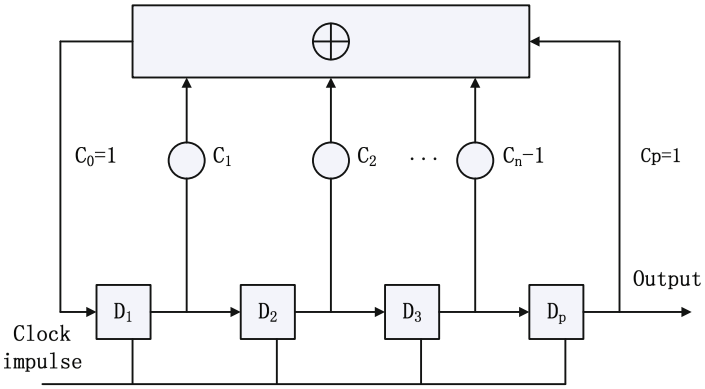


Fig. 1. The generator structure of m sequence.

From Fig. 1, D_i represents the i-th order LSR, and C_i is the feedback coefficient, which represents whether it has feedback at the i-th level (if $C_i = 1$, it has feedback. Otherwise, it has not). If the order of LSR is p, its period is $2^p - 1$ [16], and the characteristic polynomial which generate the m sequence can be expressed as [17]:

$$f(x) = C_0 + C_1 x^1 + \cdots + C_p x^p = \sum_{i=0}^{p} C_i x^i, \tag{8}$$

where $C_0 = C_p = 1$. The output of each shift register level can do modulo 2 add operation with the output of the last level shift register. We use the output of modulo 2 add operation as the input of next level.

The m sequence (which length is H) is divided into $\lfloor H/h \rfloor$ segments (the length of symbol is h), the residue length $(H - h \cdot \lfloor H/h \rfloor)$ is discarded. Then the binary symbol of each segment is converted to a 2^h-ary, that is, an m sequence is transformed into a 2^h-ary sequence, i.e., TH sequence. In this paper, the m sequence is segmented by the length of 3 and 4 symbols, and then be converted into multi-ary sequences. Set an m sequence (which generated by a 5-th order LSR)as follow: 1 0 0 0 0 1 0 0 1 0 1 1 0 0 1 1 1 1 1 0 0 0 1 1 0 1 1 1 0 1 0, and the specific method shown in Fig. 2.

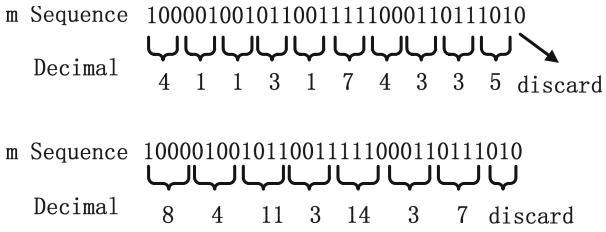


Fig. 2. The method of transforming m sequence into TH sequence.

3.2 Using Gold Sequence to Achieve Time-Hopping Code

Although the performance of m sequence is excellent, the cross-correlation values between m sequences are not good. The Gold sequence is a composite sequence of m sequence, and it is generated by two m sequences whose value of their cross-correlation function is extremely small [18]. And the cross-correlation function between any two sequences in the Gold code satisfy that [17]:

$$\left|R_{ab}(k)\right| \leq \begin{cases} 2^{\frac{p+1}{2}}+1, p \text{ is odd} \\ 2^{\frac{p+2}{2}}+1, p \text{ is even.} \end{cases} \tag{9}$$

In addition, it cannot generate optimization pairs when $R_{ab}(k)$ is an integer multiple of 4. Therefore, when the p is even, it must satisfy that p cannot be divisible by 4.

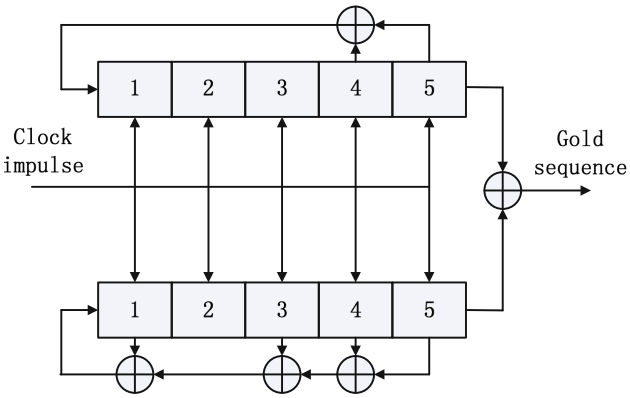


Fig. 3. The generator structure of Gold sequence.

The structure of Gold sequence generator is shown in Fig. 3. From the figure, we can see that the Gold sequence is generated by the two m sequences generators. If we only change the initial state of the second generator, both of them will generate an m sequence. Under the same clock controller, Gold sequence is generated by these two m sequences by doing modulo 2 add operation. That is to say, different initial states of the second generator result in different Gold sequences.

The method of converting the Gold sequence into a TH sequence is the same as the m sequence. However, the autocorrelation of Gold sequence is not as good as m sequence when the time delay is not zero, while the autocorrelation of Gold sequence is similar to m sequence when the time delay is zero. Therefore, the autocorrelation of m sequence is better than Gold sequence. Furthermore, the value of cross-correlation function of Gold sequence is smaller than m sequence, that is, Gold sequence has a better cross-correlation characteristic than the m sequence. Thus, the number of using gold sequence to achieve the TH sequence is much larger than the m sequence, i.e., it can achieve more users to transmit signal in the same frequency band, meanwhile the repetition rate of different users transmitted pulse is greatly reduced. It can avoid aliasing as much as possible, and degrade the BER performance.

4 Simulation Results and Analysis

In this section, the TR system based on TH technology is realized by simulation. The BER performance is evaluated in the IEEE802.15.4a CM1 channel models. CM1 represents strong residential line-of-sight (LOS) environments. The m and Gold sequences are used to realize the TH code. In this process, the two sequences are segmented by the length of 3 and 4 symbols, respectively. The BER performance is evaluated for m and Gold sequences with different length, and with different numbers of users, respectively.

Some important simulation parameters are set as follows: The pulse pair repeat period is 4, i.e., $N_s = 4$; the system sampling rate is 3952 MHz; and the length of data packet for each user is 128. All simulation results were averaged over 100 random channel realizations.

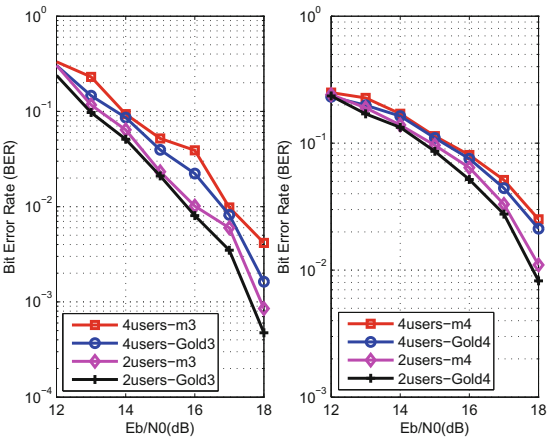


Fig. 4. The BER performance of m and Gold sequences which the length of symbol is 3 and 4, respectively.

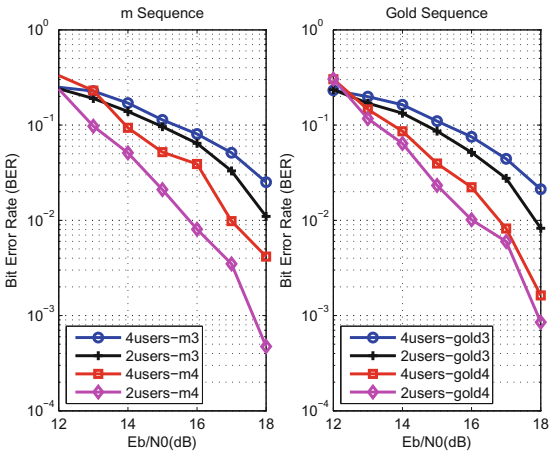


Fig. 5. The BER performance of m and Gold sequences, respectively.

Figure 4 shows that the BER performance in the CM1 channel model with the length of 3 and 4 symbols, respectively. As can be seen that the system BER gradually decreases with the increase of the SNR, and under the same number of users, the performance of Gold sequence is better than the m sequence. That is due to the number of Gold sequence is greater than the m sequence, and the cross-correlation characteristic is superior to the m sequence. From Fig. 5, it shown that the TH sequence which symbol length is 4 has better performance than the symbol length is 3 under the same conditions. However, the former is inevitably at the expense of lower data rate.

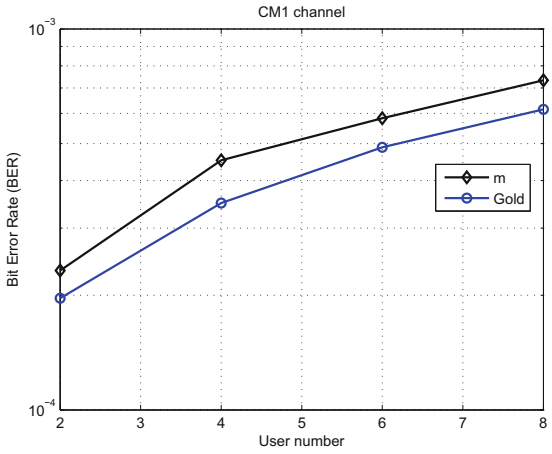


Fig. 6. The BER performance of m and Gold sequences under the different number of users, respectively.

As can be seen in Fig. 6, when the SNR is fixed at 20 dB and the length of symbol is 4, the system BER gradually increases with the increase of the number of users. And the performance of the Gold sequence is superior to the m sequence under the same number of users. For the same TH code, when the number of users increases, each frame is divided into more time slots, that is, each symbol needs more time to transmit, so the data transmission rate of each symbol is reduced, the BER of system become higher.

5 Conclusions

In this paper, the TH multi-access technology is introduced to the TR-UWB system, the m and Gold sequences are used to realize multi-access communications. The BER performance of the two TH sequences are compared by simulation. In the CM1 channels, the Gold sequence performance is superior to the m sequence. Moreover, the length of symbol affects the system performance, i.e., the TH sequence with symbol length 4 has better performance than the one with symbol length 3.

References

1. Wu, W., Rong, J., Hong, J., Wang, B.: Performance of UWB PPM-TH system with time reversal and its improved solution one-bit TR. In: Proceedings of IEEE International Symposium on Intelligent Signal Processing and Communication Systems, pp. 1–4, December 2010
2. Craciun, F., Halunga, S., Fratu, O.: Multi-user interference performances for PPM-TH-UWB systems. In: Proceedings of IEEE 19th Telecommunications Forum, pp. 360–363, November 2011
3. Wu, W., Wu, Z., Xie, W.: UWB PPM-TH and PAM-DS system with time reversal and its improved solution. In: Proceedings of IEEE 6th International Conference on Information and Automation for Sustainability, pp. 332–336, September 2012
4. Kouassi, K., Clavier, L., Doumbia, I., Rolland, P.-A.: Multiple-access interference management in TH-PPM UWB systems with pulse width randomisation. In: Proceedings of IEEE International Conference on Ultra-wideband, pp. 165–169, Sepember 2011
5. Dong, X., Xiao, L., Lee, A.: Performance analysis of dual pulse transmission in UWB channels. IEEE Commun. Lett. **10**, 626–628 (2006)
6. Baringbing, J., Witrisal, K.: Performance evaluation of multiple-access dual-pulse transmitted reference UWB systems. In: Proceedings of IEEE International Wireless Communications and Mobile Computing Conference, pp. 214–219, August 2008
7. Erseghe, T., Cipriano, A.M.: Maximum likelihood frequency offset estimation in multiple access time-hopping UWB. IEEE Trans. Wirel. Commun. **10**, 2040–2045 (2011)
8. Mary, P., Fijalkow, I., Poulliat, C.: Multi-rate resource allocations for TH-UWB wireless communications. Trans. Wirel. Commun. **11**, 4470–4481 (2012)
9. Li, L., Townsend, J.K., Ulman, R.J.: Multiple-access performance of transmitted reference UWB communications with M-ary PPM. In: Proceedings of IEEE Global Telecommunications Conference, pp. 1–6, November 2008

10. Shen, Y., Ueng, F., Chen, J., Huang, S.: A high-capacity TH multiple-access UWB system with performance analysis. IEEE Trans. Veh. Technol. **59**, 742–753 (2010)
11. Xu, Z., Sadler, B.M.: Multiuser transmitted reference ultra-wideband communication systems. IEEE J. Sel. Areas Commun. **24**, 766–772 (2006)
12. Young, D.J., Beaulieu, N.C.: Multiuser interference mitigation in time-hopped ultra-wideband receivers. In: Proceedings of IEEE International Conference on Ultra-wideband, pp. 125–129, Sepember 2012
13. Anis, N., Belghith, S.: Performance optimization of TH-UWB system in multipath channel. In: Proceedings of IEEE International Congress on Ultra Modern Telecommunications and Control Systems, pp. 31–37, October 2010
14. Jia, T., Kim, D.: Multiuser performance of balanced UWB transmitted-reference system in multipath. In: Proceedings of IEEE GLOBECOM, pp. 1–5, November 2006
15. Molisch, A.F., Balakrishnan, K., Chong, C.-C., Emami, S., Fort, A., Karedal, J., Kunisch, J., Schantz, H., Schuster, U., Siwiak, K.: IEEE 802.15.4a channel model - final report, IEEE P802.15-04-0662-00-004a. Technical report, October 2004
16. Khatter, A., Goyal, P.: Design and analyze the various m-sequences codes in MATLAB. Int. J. Emerg. Technol. Adv. Eng. **11**, 125–129 (2012)
17. Bao, Y., Li, X.: Spread Spectrum Communication Technology and Its Application, 2nd edn. Xidian University, Xi'an (2011)
18. Zhang, X.: Analysis of M-sequence and Gold-sequence in CDMA system. In: Proceedings of IEEE 3rd International Conference on Communication Software and Networks, pp. 466–468, May 2011

Check for updates

Improved Ultra-wideband Pulse Shaping Technique Based on Spectrum Shifted Gaussian Waveforms

Danli Liu, Zhonghua Liang(✉), La'ning Ma, and Huansheng Song

School of Information Engineering, Chang'an University,
Xi'an 710064, People's Republic of China
lzhxjd@hotmail.com

Abstract. In the ultra-wideband (UWB) communication system, Gaussian monocycle is the most frequently used due to its simplicity. However, in a Gaussian monocycle, the most of signal energy is concentrated in low frequency band, in which direct current components exist. Therefore, it cannot fully meet the U.S. Federal Communication Commission (FCC) on radiation template requirements. Meanwhile, in order to avoid the interference with other wireless systems, good pulse design is particularly important. Accordingly, several pulses were proposed, such as Scholtz's pulse, Hermite pulse, and spectrum shifted Gaussian waveforms (SSGW) pulse. However, it was shown that coefficient adjustment of SSGW pulse requires too many computations or iterations. In this paper, in order to overcome this drawback, random selection (RS) and Least Square Error (LSE) algorithms are introduced to the SSGW pulse design. Correspondingly, an improved SSGW waveform design is presented. Our analysis and results show that the improved approach has significant flexibility in designing and therefore it can meet the requirements of FCC spectrum mask better.

Keywords: Ultra-wideband (UWB)
Spectrum shifted Gaussian waveforms (SSGW)
Random selection (RS) · Least Square Error (LSE)

1 Introduction

Ultra-wideband (UWB) is a promising technology for high data rate wireless communications [1]. It uses nanosecond ultra-narrow pulses as transmitted signals, which carry the digital information to be transmitted. UWB system has low power

Z. Liang—This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61271262 and 61572083, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2017JM6099, and in part by the Fundamental Research Funds for the Central University under Grant 310824171004.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 135–145, 2018.
https://doi.org/10.1007/978-3-319-78130-3_15

consumption, low complexity, immunity to multipath fading and high frequency band utilization compared to the narrowband (NB) communication systems [2].

The Federal Communication Commission (FCC) has defined an intentional UWB device as one that has a bandwidth equal to or greater than 20% of the center frequency, or that has a bandwidth equal to or greater than 500 MHz [3]. In order to reduce potential interference with existing NB communication systems and to make efficient use of already scarce spectrum, the FCC has proposed spectrum mask for UWB indoor and outdoor systems which shown in Fig. 1, allowing a maximum transmitted spectral density of −41.3 dBm/MHz [4,5].

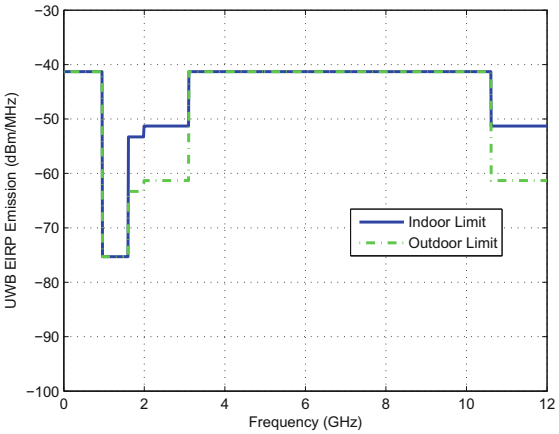


Fig. 1. FCC emission mask limit.

Various pulse shaping methods have been proposed in the literature. These pulse design methods can be divided into two categories [6]. One is the design method from time domain to frequency domain, that is, adjust the constraint parameters of a pulse's time-domain expression to guarantee that the corresponding spectrum meets the FCC spectral mask. Another is from the frequency domain to the time domain, where according to the spectrum masking in the frequency domain constraints, the pulse's time-domain expression can be obtained, and all filter design approaches belong to this kind of method.

Many researches have been done for design of UWB pulses based on standard Gaussian pulse [1,3,7–10]. However, the standard Gaussian monocycle does not meet the FCC spectral mask requirements and it contains direct current component. The second derivative of Gaussian function was proposed in [11] to avoid the existence of direct current component and low radiation efficiency. A flexible and systematic method was proposed in [12] to generate UWB pulses, in which the eigenvalue decomposition digital filter method was used to obtain two pulses, namely Parr pulses. Their energy is spread to the 3.1–10.6 GHz frequency band and both pulses meet the FCC spectral mask requirements. In order to reduce the interference between multi-user, the Hermite polynomial function was proposed [13].

The method of linear combination with differential Gaussian pulse was proposed in [14] to better fit the FCC spectrum mask. In this method, two algorithms were used to determine the combination coefficients, i.e., random selection (RS) algorithm and Least Square Error (LSE) algorithm. In order to solve the optimization problem of combination coefficients, a series of intelligent optimization algorithm were proposed [7,10,15]. More recently, a new pulse waveform design method, referred to as spectrum shifted Gaussian waveforms (SSGW) approach, was proposed in [9], and it has significant flexibility. Particularly, the SSGW pulse can be generated by superposition of Gaussian sub-carriers, and therefore it can be used to generate pulses meeting the FCC spectrum masks specified for UWB applications.

SSGW approach can be widely applied due to its flexibility, and it extremely useful in designing a time domain pulses with a predetermined spectral requirement. Therefore, this approach can also be widely used to generate the pulses for digital communication systems. But the analysis reported in [16] showed that coefficient adjustment requires too many computations or iterations. This could be a major drawback of the SSGW technique. In this paper, in order to overcome this drawback, the RS and LSE algorithms are introduced to the SSGW pulse design. Accordingly, two improved SSGW waveforms are presented.

The rest of this paper is organized as follows. Section 2 presents the RS and LSE algorithms based on the method of linear combination with differential Gaussian pulse. In Sect. 3, an improved pulse design method based on SSGW approach is presented. Both numerical results and discussions are presented in Sect. 4. Finally, Sect. 5 gives the summary.

2 Random Selection and LSE Algorithms

Several pulses have been proposed such as the Hermite pulses [13], Laplacian [17] and combination of Gauss pulses [8]. Two important parameters of Gaussian pulse are shape factor α and variance δ^2. These two parameters can be used to control Gaussian pulse's properties. The basic Gaussian pulse was defined as [14]

$$P(t) = \pm \frac{1}{\sqrt{2\pi\delta^2}} exp(-\frac{t^2}{2\delta^2}) = \pm \frac{\sqrt{2}}{\alpha} exp(-\frac{2\pi t^2}{\alpha^2}), \tag{1}$$

where $\alpha^2 = 4\pi\delta^2$ is the shape factor and δ^2 is the variance. Higher derivatives of Gaussian pulse are formed by the Gaussian pulse passing a suitable high pass filter. The n-th derivative of the Gaussian pulse can be written as [18]

$$P_n(t) = \pm \frac{\sqrt{2}}{\alpha} \frac{d^n}{dt^n} \left[exp(-\frac{2\pi t^2}{\alpha^2}) \right]. \tag{2}$$

The Gaussian derivative approach uses Gaussian derivatives and their optimal combination to generate pulses. And the combined pulse can be expressed as [8]:

$$f(t) = \sum_{i=1}^{n} w_i P_i(t, \alpha_i), \tag{3}$$

where the weighting coefficients are $W = [w_1, w_2, \cdots, w_n]$, α_i is the shape factor of the i-th order derivative of the Gaussian pulse and it is used as a tuning parameter, $P_i(t, \alpha_i)$ denotes the i-th order Gaussian pulse. In order to obtain the weighting coefficients W, we can use the RS algorithm, or error minimization such as the LSE algorithm.

2.1 Random Selection Algorithm

Random coefficient selection is performed by random number as a set of linear combination. Then under the conditions of radiation masking, choose a combination to meet the FCC requirements. A simple procedure for selecting the coefficients can be described as follows [14]:

Step 1: Choose a set of base functions $GP = [P_1(t, \alpha_1), P_2(t, \alpha_2), \ldots, P_n(t, \alpha_n)]$;
Step 2: Randomly generated a set of weighting coefficients, named W;
Step 3: Check the power spectral density (PSD) of the combined pulse which the weighting coefficient is W, whether satisfies the FCC emission limits;
Step 4: If the PSD of combined pulse does not meets the requirements of FCC spectral mask, return to *Step* 2;
Step 5: If the PSD of combined pulse meets the requirements of FCC spectral mask, and W is the first set of satisfied coefficient, then the initialization is set as $W' = W$; If the initialization has existed, then W is compared with W', based on the distance metrics that defined before, if the combined waveform generated by W is better than that of W', then set $W' = W$;
Step 6: Repeat *Step* 2–5 until the interval between the PSD of combined pulse and the FCC spectral mask is under the fixed threshold.

Specifically, we use the set of base functions GP composed by the first 15 derivatives of the Gaussian pulse ($n = 15$) with $\alpha = 0.714$ ns [14]. The PSD of the corresponding combined pulse shown in Fig. 2. From Fig. 2, we see that a combination of several base functions leads to a better approximation of the FCC emission mask compared with the basic Gaussian monocycle. But the major disadvantage lies in the lack of fitting in high frequency.

2.2 LSE Algorithm

Random coefficient selection is only a method for setting the linear combination coefficients. In the LSE algorithm, in order to optimize these coefficients, a standard process can be used to minimize the error, its objective function can be defined as [8]:

$$e(f) = min \int_{-\infty}^{+\infty} \left| P_M(f) - F(f) \right|^2 df, \tag{4}$$

where $P_M(f)$ denotes the FCC emission mask, and $F(f)$ is the PSD of the combined pulse. The time domain expression of emission mask $m(t)$ is equal to

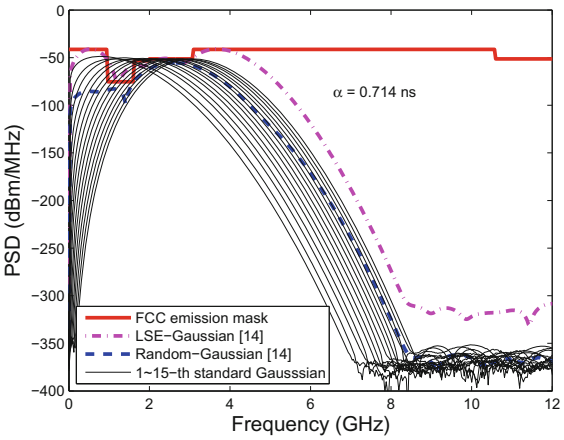


Fig. 2. PSD profiles using optimal combination of Gaussian derivatives (up to 15-th order) using RS and LSE algorithms [14], respectively.

the normalized power emission mask divided by free-space impedance, take its square root, then Inverse Fourier Transform (IFT). The objective function is rewritten as [19]:

$$e = min \int_{-\infty}^{+\infty} \left| m(t) - \sum_{i=1}^{n} w_i P_i(t, \alpha_i) \right|^2 dt. \tag{5}$$

Therefore, the pulse waveform design is converted to the minimum of the objective function.

The result using LSE algorithm is given in Fig. 2 (also see Fig. 2 in [14]). From Fig. 2, we see that the PSD of this pulse meets the requirement of FCC spectral mask. But in the low frequency band, there are out-of-band emissions, and it could cause interference to the existing narrow-band systems.

3 Improvement of UWB Pulse Shaping

The SSGW waveform [9] is one of the recent advances in the UWB pulse shaping design. This approach transforms Gaussian pulses from the time domain to the frequency domain by multiplying them with different weighted sinusoidal carriers. It provides good fitting properties of spectrum in terms of the FCC regulations. But this method has a major challenge, i.e., coefficient adjustment requires too many computations or iterations. In this section, we first introduce the SSGW approach [9], and then based on this method, the RS and LSE algorithms will be introduced to optimize weighting coefficients. Accordingly, two pulse waveforms are generated by simulation.

3.1 SSGW Algorithm

To accomplish the SSGW design process, the whole spectrum is divided into five different parts (as shown in Table 1), and in each part, a weighted sum of sinusoidal carrier modulated Gaussian pulses is used to generate a waveform. The corresponding pulse waveform expression is defined as [9]

$$X(t) = x(t)\{A_1 cos(2\pi f_{c1} t) + A_2 cos(2\pi f_{c2} t) + \cdots + A_m cos(2\pi f_{cm} t)\}, \quad (6)$$

where $x(t)$ is the Gaussian pulse defined by $x(t) = (1/\sqrt{2\pi})e^{-t^2/2}$, $F = [f_{c1}, f_{c2}, \cdots, f_{cm}]$ includes the center frequencies of modulated sinusoidal carriers, and $A = [A_1, A_2, \cdots, A_m]$ are the weighting coefficients, which should be determined in the UWB pulse design following to the FCC mask requirements.

Table 1. Spectrum segmentation and center frequency division

	Frequency	Center frequency (f_c)
First part	0–0.96 GHz	$f_{c1} = 0.48$ GHz
Second part	0.96–1.6 GHz	$f_{c2} = 1.28$ GHz
Third part	1.6–3.1 GHz	$f_{c3} = 2.35$ GHz
Fourth part	3.1–10.6 GHz	$f_{c4} = 6.85$ GHz
		$f_{c5} = 5.85$ GHz, $f_{c6} = 7.85$ GHz
		$f_{c7} = 4.85$ GHz, $f_{c8} = 8.85$ GHz
		$f_{c9} = 3.85$ GHz, $f_{c10} = 9.85$ GHz
Fifth part	10.6–12 GHz	$f_{c11} = 11.3$ GHz

As for the choice of center frequency, taking a Gaussian pulse with duration 2 ns as an example, we can see that sum of the PSD of the Gaussian pulses (the difference between two adjacent frequencies is 1 GHz, as shown in Fig. 3) will be a flat constant value, as shown in Fig. 3. Using the method presented in [9], we can obtain center frequency for the all Gaussian pulses, as shown in Table 1.

To determine the weighting coefficients, the SSGW method needs two steps, i.e., the spectrum synthesis of each part and the weighting adjustments of adjacent spectrum. The accuracy of adjustments depend on modification step-size, and set it as 0.001 in [9]. Therefore, it need at least 1000 times trials to obtain the optimal coefficients. Then adding up the synthesized pulses from all regions, the overall power spectrum still possibly exceeds the FCC power spectrum limit in the boundaries of two regions. So some coefficients need to be adjusted in these regions.

The output waveforms of this SSGW approach are reproduced from [9] as shown in Figs. 4 and 5. In which the pulse after weighting adjustments among adjacent spectrum is defined as $x^{(2ns)}_{sg_all(12345)}(t)$, and without adjustment is defined as $x^{(2ns)}_{sg_all}(t)$.

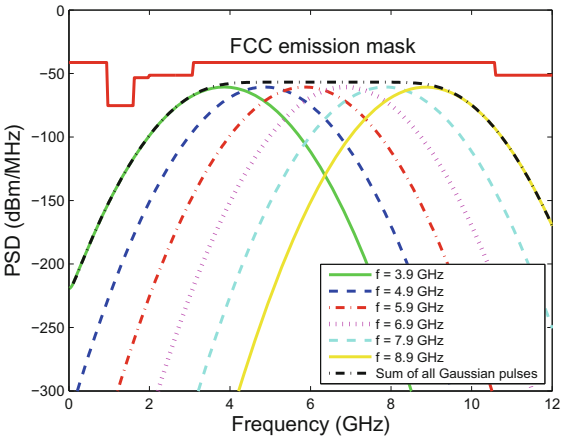


Fig. 3. PSD profiles Gaussian pulses with center frequencies 3.9, 4.9, 5.9, 6.9, 7.9, and 8.9 GHz, and the sum of all these Gaussian pulses, respectively.

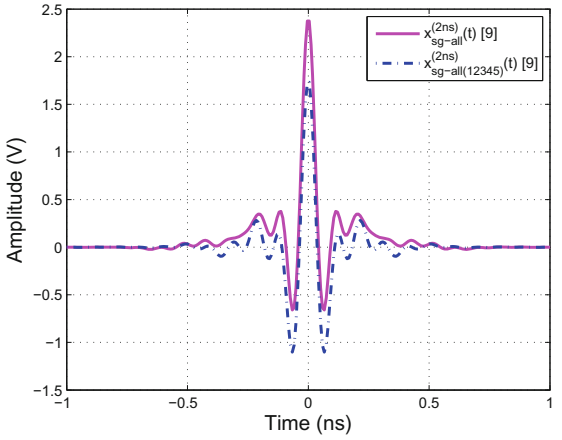


Fig. 4. Time domain waveform of the synthesized pulses with and without adjustment in SSGW technique [14].

3.2 Improved SSGW Algorithm

In order to improve the performance of the SSGW pulse, the RS and LSE algorithms are introduced to the SSGW approach. Accordingly, the pulse waveform can be rewritten as

$$\begin{aligned} X(t) &= x(t)\sum_{i=1}^{5} A_i \sum_{j=1}^{11} cos(2\pi f_{cj} t) \\ &= x(t)\{A_1 cos(2\pi f_{c1} t) + A_2 cos(2\pi f_{c2} t) + A_3 cos(2\pi f_{c3} t) \\ &+ A_4[cos(2\pi f_{c4} t) + \cdots + cos(2\pi f_{c10} t)] + A_5 cos(2\pi f_{c11} t)\}. \end{aligned} \tag{7}$$

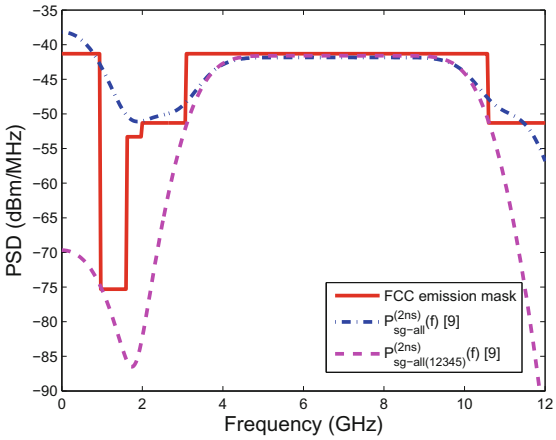


Fig. 5. PSD of synthesized pulses with and without adjustment in SSGW technique [14].

To perform the hybrid algorithm design process, we need to consider four steps as following:

Step 1: Spectrum segmentation. According to the FCC regulation on the UWB spectral mask, the whole UWB bandwidth is divided into five different segments, as illustrated in Table 1;

Step 2: Pulse selection. We select basic Gaussian pulse with a time duration of 2 ns;

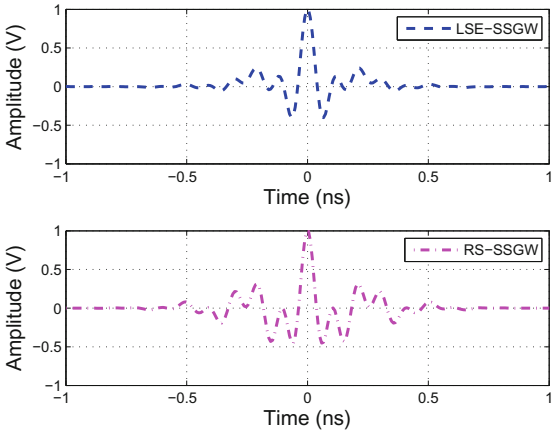


Fig. 6. Time domain waveforms using the RS and LSE algorithms based on SSGW approach, respectively.

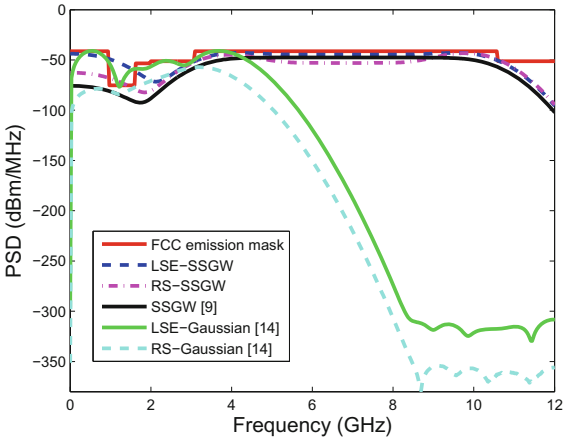


Fig. 7. Comparison of PSD with Gaussian derivative, SSGW and improved SSGW approaches.

Step 3: Center frequency selection. Using the method presented in [9], we can obtain center frequency for the all Gaussian pulses, as shown in Table 1;
Step 4: Spectrum weighting adjustment. In this part, we use the RS and LSE algorithms instead of the method provided in [9].

The time domain and frequency domain results using the improved SSGW pulses are given in Figs. 6 and 7.

4 Numerical Results and Discussions

In this section, we present the numerical results about SSGW [9], Gaussian derivative [14] and improved SSGW methods. Some important parameters are set as follows: $x(t) = (1/\sqrt{2\pi})e^{-t^2/2}$, the center frequencies are given in Table 1. For the RS algorithm, we set the number of iterations to 100.

The PSD profiles of the Gaussian derivative approach are shown in Fig. 2. We see the combination of several base Gaussian derivative functions lead to a better approximation of the FCC emission mask compared to the Gaussian monocycle. However, different minimum criteria have different combination results. The pulse spectrum obtained by the RS algorithm can match the FCC emission mask, and the power corresponding to each frequency point in the whole frequency range is less than the radiation limit. The LSE algorithm is used to minimize the total power spectrum error of the combined pulse. The optimization process is based on mean square distance, rather than point by point constraint. Therefore, the PSD profile exceeds the limits of FCC mask in some frequency bands. Moreover, in the high frequency band, both the two kinds of combined pulses cannot satisfy the FCC spectrum mask.

As described in Figs. 6 and 7, the SSGW pulse can be improved using the RS and LSE algorithms. Figure 6 shows the time-domain profiles of the improved SSGW waveforms using the RS and LSE algorithms, respectively. From Fig. 6, we see that there are less ripples in the time-domain profile using the LSE algorithm. Figure 7 shows that compared to the SSGW [9] and Gaussian derivative approaches [14], the methods presented in this paper cause much less gap with the FCC mask.

5 Conclusions

In this paper, the RS and LSE algorithms were introduced to the SSGW approach. Using the improved SSGW methods, overcome the drawback of the SSGW technique, i.e., coefficient adjustment requires too many computations or iterations. The improved SSGW methods not only offer efficiency in meeting the FCC spectral mask, but also effectively exploit the allowable bandwidth and power. Compared to the Gaussian derivative approach using the same algorithms, the improved SSGW methods fit better with the FCC spectral mask requirements.

References

1. Patil, P.A., Chiraniya, S.R., Patki, A.B.: Fifth derivative Gaussian pulse generator for UWB application. In: Proceedings of IEEE International Conference on Computing Communication Control and automation (ICCUBEA), pp. 1–4, August 2016
2. Yin, Z., Wu, S., Shi, Z., Wu, Z.: A new design of pulse waveform for waveform division multiple access UWB wireless communication system. In: Proceedings of IEEE International Conference on Communication Software and Networks (ICCSN), pp. 535–538, June 2016
3. Kumar, V.V., Ajith, V., Meenakshi, M.: Wavelet-based Gaussian impulse generation and optimization for UWB communication. In: Proceedings of IEEE International Conference on Advances in Computing and Communications (ICACC), pp. 214–218, June 2015
4. Barraj, I., Trabelsi, H., Masmoudi, M., Bouzid, A., Djebou, A.: A new UWB pulse shaping for IEEE 802.15.4a. In: Proceedings of IEEE International Multiconferences on Systems, Signals Devices (SSD-2013), pp. 1–5, March 2013
5. Gierl, C., Le, Q.T., Damm, C., Kppers, F.: Generation of UWB pulses utilizing directly modulated tunable MEMS-VCSEL. In: Proceedings of IEEE Topical Conference on Wireless Sensors and Sensor Networks (WiSNet), pp. 41–43, January 2015
6. Kim, S., Kim, Y., Li, X., Kang, J.: Orthogonal pulse design in consideration of FCC and IEEE 802.15.4a constraints. Commun. Lett. **17**, 896–899 (2013)
7. Keshavarz, S.N., Hamidi, M., Khoshbin, H.: A PSO-based UWB pulse waveform design method. In: Proceedings of IEEE International Conference on Computer and Network Technology, pp. 249–253, April 2010
8. Li, L., Wang, P., Wu, X., Zhang, J.: Improved UWB pulse shaping method based on Gaussian derivatives. In: Proceedings of IEEE IET International Communication Conference on Wireless Mobile and Computing (CCWMC 2011), pp. 438–442, Sepember 2011

9. Bai, Z., Liu, J., Chen, H.-H.: Design of ultra-wideband pulses based on spectrum shifted Gaussian waveforms. IET Commun. **7**, 512–520 (2013)
10. Qu, K., Chen, C., Chang, H., Niu, D., Wang, X.: UWB pulse design based on real-coded genetic algorithm. Fire Control Command Control 167–170 (2014). (in Chinese)
11. Win, M.Z., Scholtz, R.A.: Ultra-wide bandwidth time-hopping spread-spectrum impulse radio for wireless multiple-access communications. IEEE Trans. Commun. **48**, 679–689 (2000)
12. Parr, B., Cho, B., Wallace, K., Ding, Z.: A novel ultra-wideband pulse design algorithm. Commun. Lett. **7**, 219–221 (2003)
13. Jiang, G., Zhu, H., Cao, W.: Waveform design and performance analysis of ultra-wideband pulse based on Hermite functions. J. Microw. **2**, 9–13 (2007). (in Chinese)
14. Nardis, L.D., Giancola, G., Benedetto, M.-G.D.: Power limits fulfilment and MUI reduction based on pulse shaping in UWB networks. In: Proceedings of IEEE International Conference on Communications, pp. 3576–3580, June 2004
15. Li, Z., Zhou, Y., Sun, X., Zhou, Z.: A GA-based UWB pulse waveform design. Radio Eng. **1**, 14–16 (2012). (in Chinese)
16. Menon, M.B., Gopakumar, A., Iqbal, N.V.: A hybrid approach for UWB pulse shaping. In: Proceedings of IEEE International Conference on Electronics and Communication Systems (ICECS), pp. 373–377, June 2015
17. Yang, L., Ji, M., Zhang, W., Liu, Q.: Simulation study on UWB pulse signal based on IEEE 802.15.4a channel. J. Syst. Simul. 2172–2176 (2012). (in Chinese)
18. Sharma, S., Sharma, A., Bhatia, V.: Performance of pulse position modulation using various UWB pulses. In: Proceedings of IEEE International Advance Computing Conference (IACC), pp. 650–654, June 2015
19. Jia, Z., Chen, H., Cai, X., Chen, X.: A novel pulse design based on sinusoid Gaussian function for UWB communication. In: Proceedings of IEEE International Conference on Broadband Network and Multimedia Technology (IC-BNMT), pp. 1240–1244, October 2010



An Uplink SCMA Scheme Combined with Interleaving

Fei Yang(✉) and Zhisong Bie

Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China
{yfei, zhisongbie}@bupt.edu.cn

Abstract. SCMA is a novel, promising non-orthogonal multiple-access technique which adopts iterative detecting algorithms with affordable complexity. Current multiuser detection algorithms are mostly based on factor graph, which usually has a small number of variable nodes and function nodes and makes the girth of the factor graph small. Because the performance of an iterative detection algorithm based on factor graph is closely related to the minimum cycle length (girth) of the graph, we propose a novel uplink SCMA scheme where independent factor graphs corresponding to different SCMA blocks are connected together by using interleaving technique. Numerical results show the proposed scheme has a substantial gain over current detection algorithm over AWGN channels. Meanwhile, the diversity effect introduced by channel coefficients in different resource elements is also considered. The diversity effect combined with interleaving is analyzed. Simulation results demonstrate interleaving can affect BER performance under some circumstances related to channel coefficients diversity.

Keywords: 5G · SCMA · Factor graph · Girth · Diversity effect
Interleaving

1 Introduction

The fundamental requirements for the fifth generation (5G) wireless networks are capable of supporting massive capacity, massive connectivity and achieving the goal of low latency and extremely high reliability [1]. In order to satisfy these requirements, some essential techniques have to be innovated. In mobile communication system, multiple access technique is an essential method to make multiple users communicate with each other at the same time. In face of 5G, NOMA schemes have attracted considerable attention and become a key technology.

NOMA schemes can enhance spectral efficiency and serve more users simultaneously without occupying more resources. [2] introduced the solutions of NOMA in 5G and its challenges and opportunities. [3] introduced several candidate NOMA schemes, including SCMA, MUSA and PDMA, etc. SCMA [4, 5] derives from Low Density Signature (LDS) [6], which combines QAM modulation module and spreading module in LDS together and performs joint optimization. The joint optimization can bring extra shaping gains to SCMA. Because of the sparsity of SCMA codewords, iterative multiuser detection algorithms [6] with affordable complexity can be adopted.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 146–154, 2018.
https://doi.org/10.1007/978-3-319-78130-3_16

Currently, many multiuser detectors using iterative message passing algorithm (MPA) based on factor graph may achieve near maximum-likelihood performance, but they have a common property: MPA detector is performed for each SCMA block each time, which means the factor graph is small and the girth of the graph is small. Given that girth is an important factor which can affect the performance of the iterative detection algorithm, it's essential to improve the traditional iterative detector by constructing a big factor graph with a larger girth based on small factor graphs. As far as we know, there hasn't been any research working on generating a big factor graph to utilize the advantage of iterative detection algorithm fully. By using interleaving technique among independent factor graphs to generate a much bigger factor graph with large girth is our main contribution.

In [7, 8], with Rayleigh flat fading channel mode, resource block diversity effect is considered and performance is analyzed in which channel interleaving is used. In this paper, we have a further research about how interleaving can affect BER performance considering Rayleigh flat fading channel with resource element diversity. This is another contribution of us.

The rest of the paper is organized as follows. Section 2 presents the SCMA system model and structure. Section 3 presents the proposed SCMA scheme combined with interleaving. Simulation results are reported in Sect. 4 and analysis is made. Section 5 concludes the paper.

2 System Model

2.1 SCMA Introduction

We consider an SCMA uplink system as illustrated in Fig. 1 [9]. It directly maps binary input bits of each user (also called layer) to an SCMA codeword and spreads the codeword over multiple shared orthogonal resources, e.g. OFDMA tones [5].

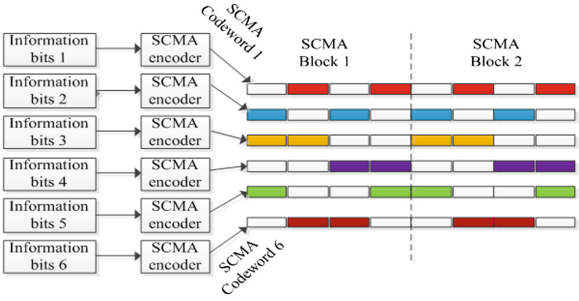


Fig. 1. Uplink SCMA transmission.

Each user has a unique K-dimensional complex codebook of size M, and each K-dimensional codeword is a sparse vector consisting of $N < K$ non-zero elements. Due to the sparsity, every $m = \log_2(M)$ bits of V users (layers) can be mapped to V SCMA

codewords, multiplexed over K orthogonal resources and it forms an SCMA block [4]. The overloading factor is defined as $\lambda = V/K$, if the number of multiplexing layers (V) is greater than spreading factor (K), SCMA can achieve higher transmission rate and more user connections.

After the transmission over a Rayleigh or AWGN channel, the received signal in vector form and the signal at the resource k in an SCMA block can be expressed respectively as [4]

$$\mathbf{y} = \sum_{v=1}^{V} diag(\mathbf{h}_v)\mathbf{x}_v + \mathbf{n}, \tag{1}$$

$$y_k = \sum_{v=1}^{V} h_{kv}x_{kv} + n_k, k = 1, 2, \ldots K, \tag{2}$$

where $\mathbf{x}_v = (x_{1v}, \ldots, x_{Kv})^T$ is the SCMA codeword corresponding to variable node (user node) v, $\mathbf{h}_v = (h_{1v}, \ldots, h_{Kv})^T$ is the channel vector consisting of channel coefficients (channel gains) corresponding to user v, and $\mathbf{n}$ is the complex additive white Gaussian noise whose mean is zero and variance is N_0.

In this paper, codebooks in [9] are adopted, which means $V = 6$, $K = 4$ and $M = 4$.

2.2 Factor Graph and Factor Graph Matrix

SCMA structure can also be represented by factor graph and factor graph matrix. In this paper, in order to make use of some independent SCMA blocks and associate them with each other, for convenience we take $n = 2$ SCMA blocks as example which means each user has a total number of $4 = n * \log_2(M)$ bits into SCMA encoder and the total number of resource elements is $8 = n * K$. Each user (layer) in an SCMA block is called a Variable Node (VN), each resource element is called a Function Node (FN). As a result, there are 8 function nodes and $12 = n * V$ variable nodes. The factor graph and factor matrix corresponding to our adopted codebooks are shown in Fig. 2.

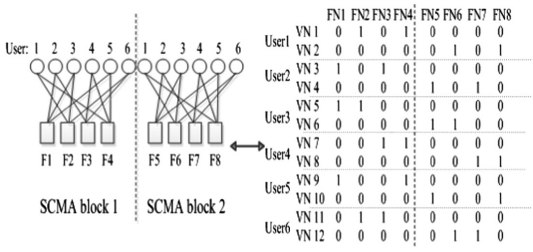


Fig. 2. Factor graph and factor matrix in traditional SCMA scheme.

From Fig. 2 we can see that each factor graph corresponding to an SCMA block is independent from each other, which means the girth of each factor graph is small. As is shown in Fig. 2, the girth of each graph is only 6.

The factor matrix corresponds to factor graph, it provides us another perspective to understand SCMA detection algorithm. The factor matrix in Fig. 2 is divided into 2 independent non-zero regions, the traditional SCMA detection algorithm for each block each time is performed based on each small independent factor matrix, but we can also understand in this way: several SCMA blocks can be detected at the same time similar to the general iterative detecting algorithm such as LDPC decoding although the structure of this factor matrix is unusual and consists of separated non-zero matrices. So our main goal is to break this independency among small factor matrices and make the factor matrix form a whole rather than several separated parts. This will be discussed later in detail in Sect. 3.

2.3 Rayleigh Channel Model

In this paper, we only consider SCMA detection in uplink channel. From Fig. 1 we know that each SCMA block consists of $K = 4$ resource elements, each resource element (RE) corresponds to a channel coefficient (channel gain). In this paper, Rayleigh flat block fading channel is considered when we need to analyze the resource element diversity effect combined with interleaving. The channel gain vector $\mathbf{h}_v$ for each user v over each fading block is assumed to be constant and may vary independently among different fading blocks [10]. In our research, we assume that each SCMA block represents one fading block which means the channel coefficients of this channel vector are all the same for all resource elements (RE) in this SCMA block. So, when $V = 6$ users transmit over $K = 4$ resource elements, there are only 6 different channel coefficients in this SCMA block. Unlike [7], which considers the diversity effect introduced by different allocations of SCMA codewords over resource blocks (RB), our assumption can also help us analyze this kind of diversity effect in a similar perspective. Over Rayleigh channel, whether the channel coefficients in one SCMA block are same does introduce diversity effect and can affect the BER performance. The detailed analysis about diversity effect associated with interleaving methods proposed by us will be demonstrated later in Sect. 4.

3 The Proposed SCMA

3.1 Algorithm Improvement

As we have mentioned in Sect. 2, traditional SCMA iterative receiver performs detection of each user's codeword in one SCMA block per time. This is a fixed pattern and its consequence is that the factor graph corresponding to a single SCMA block is small, and the graph's minimum length of cycle (girth) is small.

The performance of iterative detection algorithms based on factor graph such as LDPC decoding, has a relation to many factors, e.g. girth. Because short cycles are much susceptible to iterative detection errors, much work about how to create a factor graph with a large girth has been carried out [11].

The traditional SCMA detection algorithm iteratively passes and updates the extrinsic information between the variable nodes and function nodes, and the algorithm finally converges to reliable soft information (LLR) about the bits transmitted [4].

Usually by using this kind of iterative receiver over underlying factor graph, we can achieve near optimal performance. But the small girth of factor graph prevents us from getting a better performance of the current detection algorithm. So it's a key target for us to make use of the separate and independent factor graphs, break the independency among them and construct a connected as well as big graph based on these small graphs. When cycles are large, the estimates about the information bits will be much more accurate [11] and as a result, we can get a much better BER performance.

For convenience, we take Fig. 2 as example. In order to connect those separate factor graphs together, we adopt interleaving technique. How interleaving technique works is as follows:

n SCMA blocks means each user transmits $n * \log_2(M)$ bits and we regard these information bits as a frame. There are $n * V$ variable nodes and $K * n$ function nodes in a frame which can be illustrated in Fig. 2, the corresponding matrix has $n * V$ rows and $K * n$ columns.

For each kind of RE in each SCMA block, there are $d_f = 3$ users (variable nodes) linked to it, which means the corresponding position in matrix is 1. Then for these $d_f = 3$ users, we use a random interleaver for each user respectively and interleave the corresponding RE in all SCMA blocks. For example, the 2nd RE is linked to the 1st, 3rd and 6th user, for the 1st user, all the 2nd REs in each SCMA block are interleaved, which means the columns with number $2, 6, \ldots, K * N - 2$ in the matrix are interleaved. The 1st user's first codeword is allocated over the 2nd and 4th RE, so the coordinate (1, 2) in the original matrix stores 1, but after interleaving, this codeword may be allocated over the 4th and 6th RE, and the coordinate (1, 6) in the new matrix stores 1. The same rule goes for the rest users and other kinds of RE.

We need to use different random interleavers for each user because we need to reconstruct the fixed order of RE in original graph. The reason of interleaving K kinds of REs respectively is we want to keep the degree of variable node and function node in the original factor graph unchanged. So the generated new factor matrix is still regular. Figure 3 gives an example of what the new factor graph and matrix may be and it shows that the new factor graph is a whole with nodes connected with each other rather than separate parts like before. By using interleaving, the original order of RE are disorganized totally. By comparing Fig. 3 with Fig. 2, we can see that the girth becomes larger than previous graph's girth.

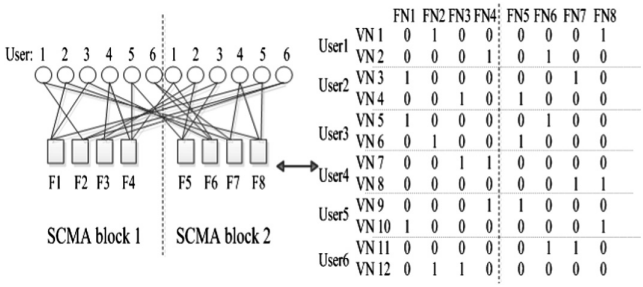


Fig. 3. Factor graph and factor matrix in our proposed SCMA scheme.

3.2 Detection Algorithm Combined with Interleaving

As we have mentioned earlier, we regard a frame as a unit and for each frame, we use the proposed interleaving scheme and get a factor matrix consisting of $n * V$ variable nodes and $K * n$ function nodes. Based on this factor matrix, we perform joint multiuser detection. The received signal $y_k, (k = 1, \ldots, K * n)$ at each RE (function node) can still be calculated using (2). And the iterative detection algorithm using the new factor matrix will be modified slightly based on the original algorithm [9, 12]. It's just similar to the general Sum-Product algorithm, e.g. LDPC decoding.

In this paper, we use $\mathbf{C}$ to represent factor matrix, and $x_{k,v}(m_v), m_v = 1, \ldots, M$ to denote the symbol on function node k if variable node v uses codeword m_v. We denote the function nodes connected to variable node v as $N(v) = \{k|(\mathbf{C})_{vk} = 1\}$, and the function nodes connected to node v excluding node k as $N(v) - \{k\}$. $N(k)$ and $N(k) - \{v\}$ have similar meaning. $L_{v\to k}$ represents the extrinsic information sent from v-th variable node to the k-th function node, the same meaning goes for $L_{k\to v}$. For this frame, the slightly modified algorithm based on [9, 12], is presented in detail in Algorithm 1.

Algorithm 1 . Modified Sum-Product algorithm based on the factor matrix combined with interleaving

1. **Initialization**: For $v = 1, \ldots, n * V$ and $k = 1, \ldots, K * n$, if $(\mathbf{C})_{vk} = 1$, set

$$L_{v\to k}(m_v) = \frac{1}{M}, m_v = 1, \ldots, M \quad , \tag{3}$$

$$\phi_k(\{m_{v_1}, m_{v_2}, m_{v_3}\}) = \exp\left[\frac{-\| y_k - \sum_{i=1}^{3} h_{k,v_i} x_{k,v_i}(m_{v_i})\|^2}{N_{0,k}}\right] , \tag{4}$$

v_1, v_2, v_3 links to node k, which equals to $N(k)$.

2. **FN update**: For each FN ($k = 1, \ldots, K * n$) with $(\mathbf{C})_{vk} = 1$, compute outgoing FN messages $L_{k\to v}$ for $m_v = 1, \ldots, M$ as follows:

$$L_{k\to v}(m_v) = \sum_{m_{v_1'}=1}^{M} \sum_{m_{v_2'}=1}^{M} L_ex \bullet \phi_k(\{m_v, m_{v_1'}, m_{v_2'}\}), L_ex = \prod_{\substack{v' \in N(k)-\{v\} \\ N(k)=\{v, v_1', v_2'\}}} L_{v'\to k}(m_{v'}) \quad , \tag{5}$$

3. **VN update**: For each VN ($v = 1, \ldots, n * V$) with $(\mathbf{C})_{vk} = 1$, compute outgoing VN messages $L_{v\to k}$ for $m_v = 1, \ldots, M$ as follows:

$$L_{v\to k}(m_v) = normalize\left(\frac{1}{M} \prod_{k' \in N(v)-\{k\}} L_{k'\to v}(m_v)\right) \quad , \tag{6}$$

4. **Stopping criteria**: The number of iterations reach the maximum limit.
5. **LLR total**: For $v = 1, \ldots, n * V$, for all m_v , compute

$$L_v(m_v) = \frac{1}{M} \prod_{k \in N(v)} L_{k\to v}(m_v) \quad , \tag{7}$$

By using Algorithm 1, we can obtain LLRs for each user's $n * \log_2(M)$ bits in this frame. Based on the factor matrix as a whole, we can achieve better performance which will be shown later.

4 Simulation Results

Simulations about our proposed SCMA combined with interleaving over AWGN channel were carried out. We also have a further research about how our proposed interleaving scheme can control the diversity effect over Rayleigh flat fading channel to some extent.

Both uncoded SCMA and turbo-coded SCMA are considered. We adopt turbo code with rate 1/2 and block length is 1000 bits. Turbo decoding was performed with 8 iterations.

4.1 Proposed SCMA Scheme Over AWGN Channel

Firstly, we evaluate the BER performance of our proposed SCMA scheme over AWGN channel. For uncoded SCMA, we choose $n = 21$ SCMA blocks as a frame, which means the factor matrix has 126 variable nodes and 84 function nodes. The case that all $K = 4$ kinds of RE are interleaved is considered. 4, 10 iterations are considered respectively. Figures 4 and 5 show the results for uncoded and turbo-coded SCMA respectively.

We can see that our proposed SCMA scheme outperforms the traditional SCMA scheme. Figure 4 shows that when the number of iterations is 10, the required Eb/N0 to achieve a BER of 10^{-4} is about 9.9 dB in our proposed SCMA while about 13.5 dB in traditional SCMA, obtaining 3.6 dB gain. Figure 5 shows with turbo coding, the proposed SCMA scheme can still achieve gain about 0.5 dB at low BER.

Because large girth can relieve the strong correlation among the messages passed throughout the detection process, our proposed scheme can achieve better performance than the traditional one.

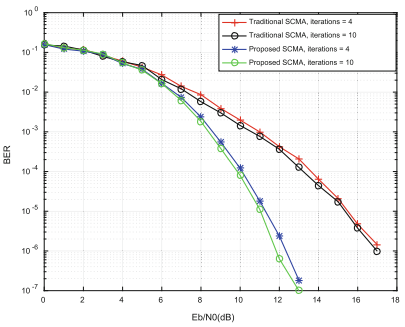


Fig. 4. BER performance of uncoded SCMA over AWGN channel.

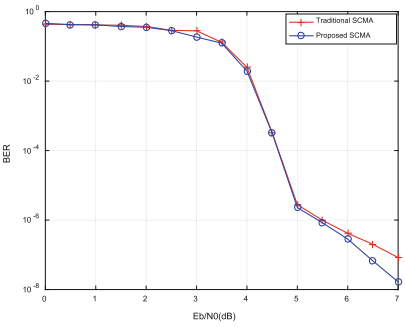


Fig. 5. BER performance of turbo-coded SCMA over AWGN channel with 10 iterations for SCMA detection.

4.2 Effect of Interleaving Over Rayleigh Channel

As mentioned in Sect. 2, we regard each SCMA block as one fading block in Rayleigh fading channel model. [7] drew some conclusions about the diversity effect on BER performance and it also considers channel interleaving. In this paper, we have a further research using our proposed interleaving scheme.

In the paper, we take all the channel coefficients same in an SCMA block as one case: without diversity effect, and channel coefficients different from each other as another case: with diversity effect. The simulation results without and with channel coding are shown in Figs. 6 and 7, respectively.

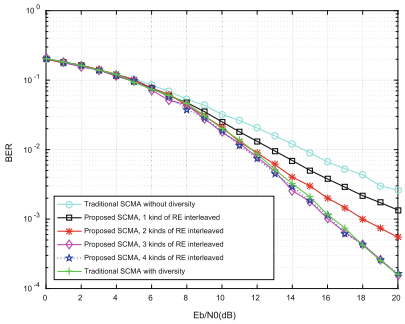


Fig. 6. BER performance of uncoded SCMA over Rayleigh channel.

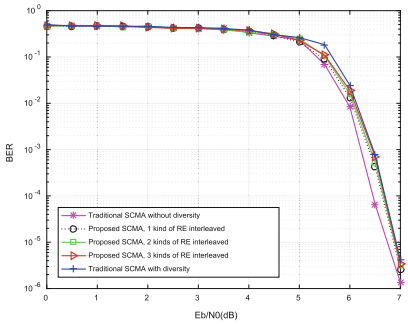


Fig. 7. BER performance of turbo-coded SCMA over Rayleigh channel with 10 iterations for SCMA detection.

Figure 6 shows that the performance of SCMA without diversity is worse than the scheme with diversity, losing 3 dB at $BER = 10^{-2}$. This conclusion is similar to [7]. As the case without diversity means the $K = 4$ channel coefficients are same, by using our proposed interleaving scheme, we can interleave 1 kind of RE, e.g. the 2nd RE in all SCMA blocks which makes each block have 1 channel coefficient different from the rest 3 coefficients. The same idea goes for interleaving 2 and 3 kinds of RE. Obviously, the case of interleaving 4 kinds of RE obtains the same performance as interleaving 3 kinds of RE because they both make the 4 channel coefficients different from each other. Over Rayleigh channel, the case of interleaving 3 or 4 kinds of RE in our proposed SCMA scheme equals to the traditional SCMA scheme apparently. Our proposed interleaving scheme changes the number of same channel coefficients in an SCMA block, and of course can control the degree of diversity effect. It can affect the performance progressively by using the number of different channel coefficients.

Figure 7 shows that with channel coding, the performance of the case without diversity is better than the case with diversity, obtaining about 0.3 dB gain at $BER = 10^{-3}$. We adopt the interleaving scheme to change the degree of diversity effect as well. The analytic process is similar.

5 Conclusion

In this paper, we presented an SCMA scheme combined with interleaving technique. Different from the traditional SCMA scheme whose detection algorithm is performed based on small factor graph, our proposed scheme can generate a big factor graph with large girth and introduce performance gain. Simulation results show that our proposed scheme outperforms the traditional one.

Acknowledgement. This paper is supported by the Defense Industrial Technology Development Program No. JCKY2016204B606.

References

1. 5G: A Technology Vision, Huawei Technologies Co., Ltd., Shenzhen, China, Whitepaper, November 2013. www.huawei.com/ilink/en/download/HW_314849
2. Dai, L., Wang, B., Yuan, Y., Han, S., Chih-Lin, I., Wang, Z.: Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. IEEE Commun. Mag. **53**(9), 74–81 (2015)
3. Wang, B., Wang, K., Lu, Z., Xie, T., Quan, J.: Comparison study of non-orthogonal multiple access schemes for 5G. In: 2015 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, Ghent, pp. 1–5 (2015)
4. Nikopour, H., Baligh, H.: Sparse code multiple access. In: 2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), London, pp. 332–336 (2013)
5. Taherzadeh, M., Nikopour, H., Bayesteh, A., Baligh, H.: SCMA codebook design. In: 2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall), Vancouver, BC, pp. 1–5 (2014)
6. Hoshyar, R., Wathan, F.P., Tafazolli, R.: Novel low-density signature for synchronous CDMA systems over AWGN channel. IEEE Trans. Signal Process. **56**(4), 1616–1626 (2008)
7. Sergienko, A.B., Klimentyev, V.P.: SCMA detection with channel estimation error and resource block diversity. In: 2016 International Siberian Conference on Control and Communications (SIBCON), Moscow, pp. 1–5 (2016)
8. Klimentyev, V.P., Sergienko, A.B.: Detection of SCMA signal with channel estimation error. In: 2016 18th Conference of Open Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT), St. Petersburg, pp. 106–112 (2016)
9. Altera Innovate Asia website, Presentation, 1st 5G Algorithm Innovation Competition-ENV1.0-SCMA. http://www.innovateasia.com/5g/en/gp2.html
10. Wang, Y., Zhou, S., Xiao, L., Zhang, X., Lian, J.: Sparse Bayesian learning based user detection and channel estimation for SCMA uplink systems. In: 2015 International Conference on Wireless Communications & Signal Processing (WCSP), Nanjing, pp. 1–5 (2015)
11. Ryan, W.E., Lin, S.: Channel Codes Classical and Modern. Cambridge University Press, Cambridge (2009)
12. Xiao, B., Xiao, K., Zhang, S., Chen, Z., Xia, B., Liu, H.: Iterative detection and decoding for SCMA systems with LDPC codes. In: 2015 International Conference on Wireless Communications & Signal Processing (WCSP), Nanjing, pp. 1–5 (2015)



Integer-Forcing Detection of Generalized Spatial Modulation

Guoquan Li[1,2(✉)], Ying Zhou[1,2], Xiangyun Zhou[1,2], Jinzhao Lin[1,2], Yu Pang[1,2], Jiacheng Wang[1,2], and Hongyu Zhu[1,2]

[1] Chongqing University of Posts and Telecommunications, Chongqing, People's Republic of China
ligq@cqupt.edu.cn

[2] Chongqing Key Laboratory of Photoelectronic Information Sensing and Transmitting Technology, Chongqing, People's Republic of China

Abstract. Generalized spatial modulation (GSM) activates more transmit antennas than spatial modulation to increase the spectral efficiency. Although the maximum likelihood (ML) detector is able to achieve the optimal performance, its exhaustive search leads to intractable computational complexity. In this paper, we design a detection scheme based on Integer-Forcing (IF) algorithm, which utilize IF receiver to decode the transmitted symbols for all possible antenna combinations, and then compare the Euclidean distances between the detected symbols and the received signals. Finally, the symbol vector with minimum Euclidean distance and the corresponding antenna combination are detected as the decided symbols and transmit antenna combination. Simulation results shows that the designed detection scheme achieves obviously better performance than the ZF and the MMSE detection with relatively lower complexity.

Keywords: GSM · Integer-Forcing · Detection

1 Introduction

With the development of wireless communication technology and high-speed multimedia services, multiple-input multiple-output (MIMO) has been widely used in wireless communication systems since it can increase the data rate without increase of extra bandwidth and transmit power [1,2]. It is well-known that increasing the number of antenna can significantly increase its capacity. Traditional MIMO technology needs to activate all transmit antenna at the same time to transmit data, which introduces Inter-Channel Interference (ICI). Additionally, MIMO also needs Radio frequency (RF) chain for each antenna and leads to high cost.

To solve the problems of MIMO, Spatial modulation (SM) was proposed by Meslesh, Haas etc. in 2006 and then attracted extensive attentions [3–5]. The basic idea is that only one transmit antenna is activated to transmit and the

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 155–165, 2018.
https://doi.org/10.1007/978-3-319-78130-3_17

rest remain silent at each moment, which can effectively avoid the (ICI) and inter-antenna synchronization (IAS). Therefore, only one RF chain is needed at the transmitter and the cost of implementation is greatly reduced. When transmitting, the data streams are divided into two parts, one of which is used to select the transmit antenna mapped to the spatial domain, and the other is mapped to the actual modulation information to be transmitted. For the SM with single active antenna, the increase of transmission rate can only be achieved by increasing the number of transmit antennas or modulation order. However, the former is limited by the cost and the physical size of equipments, and the latter would increase the error rate because of the reduction of Euclidean distance between different modulation symbols. In order to further improve the spectrum efficiency, generalized spatial modulation (GSM) [6] was proposed, which activates different transmit antenna combinations in different time slots depending on the bit patterns to be transmitted. Traditional SM technology can be considered as the case of GSM with one active antenna.

GSM signal detection has been studied by many researchers [7–9]. The Maximum Likelihood (ML) detection is the optimal detection algorithm, which jointly detects the transmit antenna combinations and transmit symbols, but the computational complexity increases exponentially with the number of active antennas and modulation order and thus restricts its application. Conventional linear detection algorithms, such as Zero-Forcing (ZF) and Minimum mean-squared error (MMSE) detection algorithm, have low computational complexity and degraded performance at the same time. Recently, a new linear receiver technique called Integer Forcing (IF) receiver for the MIMO systems has been proposed in [10–12]. Instead of decoding the transmitted symbols directly, IF receiver recovers an integer combination of the symbols by using a preprocessing matrix, which is designed to force the processed channel matrix to be full-rank and non-singular integer matrix with amplifying the noise as little as possible. Finally, the symbols are obtained by inverting the integer matrix. Rate and performance are proved to be similar to the ML decoding. Since the application of IF in the GSM systems is not well investigated, we adopt it in the detection of GSM and evaluate its performance in this paper.

The rest of paper is organized as follows. In Sect. 2, the system model is presented. In Sect. 3, IF algorithm is introduced and the decoding process is described. Section 4 presents the simulation results and analysis. Finally, a conclusion is made in Sect. 5.

Notations. Row vectors are presented by boldface letters and matrices are denoted by capital boldface letters. $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{H}}$ denotes transpose and Hermitian of a vector or matrix respectively. $\mathbf{I}$ and $\mathbf{0}$ represents identity matrix and zero matrix with proper dimension respectively. The sets $\mathcal{C}$ denotes the set of all complex numbers, where $i = \sqrt{-1}$. Let $\|.\|$ denotes the Euclidean norm of a vector.

2 System Model

The block diagram of GSM is shown in Fig. 1. Assume a GSM system with N_t transmit antennas and N_r receive antennas, in which N_a $(1 \leq N_a \leq N_t)$ antennas are activated in one time slot. Totally, there are $\binom{N_t}{N_a}$ possibilities for choosing the active antennas but only N_a antenna combinations are permitted and the other combinations are considered invalid.

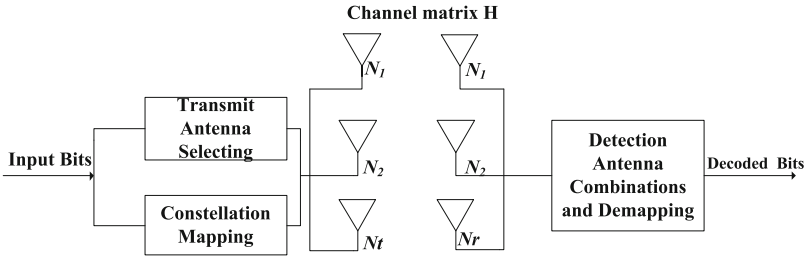


Fig. 1. Block diagram of GSM system.

In each time slot, the L bits to be transmitted are divided into L_1 antenna indexing bits and L_2 mapping bits. The first part is to select active transmit antenna combinations and

$$L_1 = \lfloor \log_2 \binom{N_t}{N_a} \rfloor. \tag{1}$$

The second part is used for M-QAM modulation mapping for active antennas and

$$L_2 = N_a \log_2 (M). \tag{2}$$

Therefore, the total number of bits

$$\begin{aligned} L &= L_1 + L_2 \\ &= \lfloor \log_2 (\tbinom{N_t}{N_a}) \rfloor + N_a \log_2 (M). \end{aligned} \tag{3}$$

Assume the active antenna combination is $\mathcal{K} = \{k_1, k_2, \ldots, k_{N_a}\}$ and the corresponding transmitted symbols are denoted by $\mathbf{s} = \{s_1, s_2, \ldots, s_{N_a}\}$, where $s_i (i = 1, 2, \cdots, N_a)$ is taken from a QAM constellation with unit average energy independently. Thus, the transmitted vectors $\mathbf{x} = [0, s_1, 0, \ldots, s_{N_a}, 0]^{\mathrm{T}}$ with zero value for the k-th $(k \notin \mathcal{K})$ entry.

In this paper, a flat fading channel model is assumed and the channel remains constant in each time slot. The channel matrix is denoted by $\mathbf{H} \in \mathcal{C}^{N_r \times N_t}$, whose entries follow complex Gaussian distribution with zero mean and unit variance respectively. Then the received vector $\mathbf{y} \in \mathcal{C}^{N_r \times 1}$can be formulated as

$$\mathbf{y} = \sqrt{\frac{\rho}{N_a}} \mathbf{H}\mathbf{x} + \mathbf{z}, \tag{4}$$

where ρ is the average signal-to-noise ratio (SNR) at each receive antenna. For convenience of expression, $\frac{\rho}{N_a}$ is defined as $\bar{\rho}$. $\mathbf{z} \in \mathcal{C}^{N_r \times 1}$ is the noise vector, whose entries are distributed as zero-mean additive white Gaussian noise with unit variance.

3 Detection Schemes of GSM

3.1 Known Detection Schemes

ML Detection: For GSM systems, the ML detection algorithm jointly detects the transmit antenna combinations and the modulated symbols by exhaustively searching all the possibilities. The output of the ML detector can be expressed as

$$(k, \mathbf{s}) = \arg\min_{k,\mathbf{s}} \|\mathbf{y} - \sqrt{\bar{\rho}}\mathbf{H}\mathbf{x}\|^2. \tag{5}$$

Totally, there are 2^{L_1} possibilities of transmit antenna combination. Considering the modulation order M, the needed times for the ML detection to calculate the Euclidean distance is $N_{ML} = 2^{L_1} M^{N_a}$. It can be seen that the computation complexity of the ML decoding grows with the number of active antennas and the modulation order exponentially, which limits its applications.

ZF Detection: ZF detection obtains the cancellation matrix $\mathbf{B}$ by inverting the channel matrix as

$$\mathbf{B}_{\mathrm{ZF}} = \frac{1}{\sqrt{\bar{\rho}}} \left(\mathbf{H}^{\mathrm{H}}\mathbf{H}\right)^{-1} \mathbf{H}^{\mathrm{H}}, \tag{6}$$

and then detects the transmitted signals by cancelling the interferences among them with it.

$$\begin{aligned} \tilde{\mathbf{y}}_{\mathrm{ZF}} &= \mathbf{B}_{\mathrm{ZF}}\mathbf{y} \\ &= \mathbf{B}_{\mathrm{ZF}}\mathbf{H}\mathbf{x} + \tilde{\mathbf{z}} \\ &= \mathbf{x} + \tilde{\mathbf{z}}. \end{aligned} \tag{7}$$

After this, the indices of the N_a entries with larger modulus of $\tilde{\mathbf{y}}$ will be chosen as the transmit antenna combination and decoded as the antenna indexing bits. The corresponding entries will be decoded as the modulation bits. For ZF decoding, the complexity is low with a performance loss because of noise amplification.

MMSE Detection: MMSE detection has the similar processing steps to ZF detection with a different cancellation matrix $\mathbf{B}_{\mathrm{MMSE}} = \frac{1}{\sqrt{\bar{\rho}}} \left(\mathbf{H}^{\mathrm{H}}\mathbf{H} + \bar{\rho}^{-1}\mathbf{I}_{N_{\mathrm{T}}}\right)^{-1} \mathbf{H}^{\mathrm{H}}$. Comparing to ZF detection, MMSE lightens the effect of noise enhancement and obtains a little better performance at low SNR.

3.2 Proposed Detection Scheme

IF receiver was first proposed in [12] as one linear receiver. Instead of detecting the transmitted symbols directly, IF receiver detects the integer combinations of them at first and reduces the noise amplification to some extent, as shown in Fig. 2. For the IF receiver, one of the key steps is to find a preprocessing matrix $\mathbf{B}$ to force the channel matrix $\mathbf{H}$ to be a nonsingular integer matrix $\mathbf{A}$ as approximate as possible, i.e. $\mathbf{BH} \approx \mathbf{A}$. However, since the active antennas are not fixed in different time slots, IF algorithm can not be adopted in a GSM system directly. In this paper, we use two methods based on IF receiver and evaluate their performance.

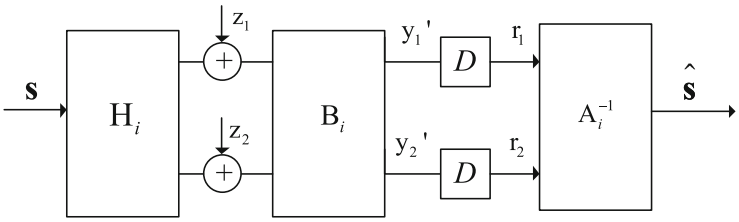


Fig. 2. Block diagram of IF

Suppose that the i-th $\left(i \in \{0, 1, \cdots, 2^{L_1} - 1\}\right)$ antenna combination is active. Based on (4), the channel model will be

$$\mathbf{y}_{\mathrm{IF}} = \sqrt{\bar{\rho}}\mathbf{H}_i\mathbf{s} + \mathbf{n}, \tag{8}$$

where $\mathbf{H}_i$ is built by taking the columns of the channel matrix $\mathbf{H}$ corresponding to the active transmit antennas. $\mathbf{n}$ is a noise vector with its entries distributed as i.i.d. zero-mean additive white Gaussian noise. By using the receive filter matrix $\mathbf{B}_i$, we have

$$\begin{aligned} \tilde{\mathbf{y}}_{\mathrm{IF}} &= \mathbf{B}_i\mathbf{y}_{\mathrm{IF}} \\ &= \sqrt{\bar{\rho}}\mathbf{B}_i\mathbf{H}_i\mathbf{s} + \mathbf{B}_i\mathbf{n} \\ &= \mathbf{A}_i\mathbf{s} + \left(\sqrt{\bar{\rho}}\mathbf{B}_i\mathbf{H}_i - \mathbf{A}_i\right)\mathbf{s} + \mathbf{B}_i\mathbf{n}. \end{aligned} \tag{9}$$

where $\mathbf{A}_i$ is the desired Gaussian integer matrix, and the effective noise is $\sqrt{\bar{\rho}}(\mathbf{BH}_i - \mathbf{A})\mathbf{x} + \mathbf{Bn}$. Accordingly, the effective noise power along the m-th row of $\mathbf{y}_{\mathrm{IF}}$ is

$$g\left(\mathbf{a}_m^i, \mathbf{b}_m^i\right) = \bar{\rho}\|\mathbf{b}_m^i\mathbf{H}_i - \mathbf{a}_m^i\|^2 + \|\mathbf{b}_m^i\|^2, \tag{10}$$

where $\mathbf{a}_m^i$ and $\mathbf{b}_m^i$ denote the m-th row of $\mathbf{A}_i$ and $\mathbf{B}_i$ respectively; Given $\mathbf{H}_i$ and P, we need to find the matrix $\mathbf{B}_i$ and $\mathbf{A}_i$ such that the maximum of all the noise powers $g\left(\mathbf{a}_m^i, \mathbf{b}_m^i\right)$ is minimized, and the corresponding matrix $\mathbf{A}_i$ is invertible.

In [13], the authors propose a method to find $\mathbf{A}$ by using the complex LLL lattice reduction algorithm (CLLL)[14]. CLLL aims to obtain the shortest vector

in the corresponding basis, and it can not be used for the minimization problem of IF receiver directly. Based on the algorithm in [13], an expanded matrix for the GSM detection is constructed as

$$\mathbf{L} = \begin{bmatrix} \sqrt{\bar{\rho}^{-(N_a+1)}}\mathbf{I} & -\sqrt{\bar{\rho}^{-N_a}}\mathbf{H}_i \\ \mathbf{0} & \mathbf{I} \end{bmatrix}. \tag{11}$$

The method of finding $\mathbf{A}_i$ and $\mathbf{B}_i$ can be summarized as follows:

---

1. Construct matrix

   $\mathbf{L} = \begin{bmatrix} \sqrt{\bar{\rho}^{-(N_a+1)}}\mathbf{I} & -\sqrt{\bar{\rho}^{-N_a}}\mathbf{H}_i \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$
2. $\mathbf{L}' = \mathrm{CLLL}(\mathbf{L})$
3. $\mathbf{V} = \mathbf{L}'\mathbf{L}^{-1}$
4. $\mathbf{v}_m = [\mathbf{d}_m \ \mathbf{c}_m]$
5. $\mathbf{a}_m^i \leftarrow \mathbf{c}_m$
6. $\mathbf{B}_i = \mathbf{A}_i\mathbf{H}_i^{\mathrm{H}} \left(P^{-1}\mathbf{I} + \mathbf{H}_i\mathbf{H}_i^{\mathrm{H}}\right)^{-1}$

In this way, we can detect $\mathbf{A}_i\mathbf{s}$ easily and obtain $\mathbf{s}$ by a integer matrix inversion accordingly. $\mathbf{s}$ is further demapped to recover the transmitted bits.

As seen from above, in order to recover the transmitted data, the transmit antenna combination is needed to be found before the detection of modulation symbols. First, we adopt the MMSE algorithm to find the antenna combination, and then finish the detection using IF algorithm. This process is named by MMSE-IF detection and listed as follows:

---

1. Find the filter matrix
   $\mathbf{B}_{\mathrm{MMSE}} = \frac{1}{\sqrt{\bar{\rho}}} \left(\mathbf{H}^{\mathrm{H}}\mathbf{H} + \bar{\rho}^{-1}\mathbf{I}_{N_{\mathrm{T}}}\right)^{-1} \mathbf{H}^{\mathrm{H}}$;
2. Obtain $\tilde{\mathbf{y}}_{\mathrm{MMSE}} = \mathbf{B}_{\mathrm{MMSE}}\mathbf{y}$;
3. Sort the entries of $\tilde{\mathbf{y}}_{\mathrm{MMSE}}$ in the descending order of modulus, and take the first N_a antennas as the active ones and the corresponding indexing bits;
4. Use CLLL algorithm to find the matrix $\mathbf{A}_i$ and $\mathbf{B}_i$ of IF algorithm;
5. Calculate $\tilde{\mathbf{y}}_{\mathrm{IF}} = \mathbf{B}_i\mathbf{y}_{\mathrm{IF}}$;
6. Round $\tilde{\mathbf{y}}_{\mathrm{IF}}$ to the nearest Gaussian integer to get the estimation of $\mathbf{A}_i\hat{\mathbf{s}}$ after mod operation;
7. Inverting the matrix $\mathbf{A}_i$ to $\mathbf{A}_i\hat{\mathbf{s}}$ to get the estimated $\hat{\mathbf{s}}$;
8. Demap $\hat{\mathbf{s}}$ to obtain the decoded bits.

Since MMSE has weak detection performance itself, there will be certain percentage of making wrong decision of the active antenna combination, and further resulting in a wrong detection with IF algorithm. In the following method, we

enumerate the possible antenna combinations, and then detect the corresponding symbols using IF algorithm. Next, we compare the Euclidean distances between the detected symbols and the received signals, and finally choose the symbol vector with minimum Euclidean distance and the corresponding antenna combination as the decided symbols and transmit antenna combination. The method is named by IF-GSM detection and the specific steps are given as follows:

---

1. Initialize the number of active antenna combination $i = 0$;
2. Use CLLL algorithm to find the matrix $\mathbf{A}_i$ and $\mathbf{B}_i$ of IF algorithm;
3. Calculate $\tilde{\mathbf{y}}_{\mathrm{IF}} = \mathbf{B}_i \mathbf{y}_{\mathrm{IF}}$;
4. Round $\tilde{\mathbf{y}}_{\mathrm{IF}}$ to the nearest Gaussian integer to get the estimation of $\mathbf{A}_i \hat{\mathbf{s}}$ after mod operation;
5. Inverting the matrix $\mathbf{A}_i$ to $\mathbf{A}_i \hat{\mathbf{s}}$ to get the estimated $\hat{\mathbf{s}}_i$;
6. Calculate the Euclidean distance

$$d_i = \|\mathbf{y} - \mathbf{H}_i \hat{\mathbf{s}}_i\|^2;$$

7. Let $i = i + 1$ and go to step 2) until $i = 2^{L_1}$;
8. Find the minimum of all d_is as the number of active antenna combination, and the corresponding $\hat{\mathbf{s}}_i$ as the detected symbol vector;
9. Output the decoded bits by demaping the symbols and converting the number of active antenna combination to bits.

Seen from the above detection scheme, the complexity is higher than the MMSE-IF detection. However, the final decision is not made based on the active antenna combinations directly which MMSE-IF detection does, and hence can achieve better performance as shown in the simulations. Comparing to the ML detection, this scheme does need the complexity of CLLL algorithm to obtain the matrix $\mathbf{A}_i$. However, the complexity increases linearly because of its linear property. The advantage of low complexity will be more obvious with the increase of the modulation order and the number of active antennas. Additionally, as long as the channel does not change, only 2^{L_1} kinds of $\mathbf{A}_i$ are needed to calculate for one time, but for the ML detection, all the Euclidean distances are needed to calculate for every time of detection.

4 Simulations Results

In this section, we present the simulation results by Bit Error Rate (BER) for several detection algorithms at the same spectral efficiency 6 bits/s/Hz. 4-QAM modulation is adopted and the constellation is $\{0, 1, 1i, 1 + 1i\}$ used to ensure the integer combinations of the constellation points are still in the constellation. We use $L_1 = 2$ bits to select the active antenna combinations and the number of active antennas $N_a = 2$. The antenna mapping rule is shown in Table 1.

Table 1. The mapping rule between antenna combination and bit patterns with $N_t = 4$ and $N_a = 2$

Bits of selecting transmit antenna	Antenna number
00	(1, 3)
01	(2, 4)
10	(2, 3)
11	(1, 4)

Figure 3 compares the bit error rate of the ML detection and IF detection of a GSM system with $N_t = 4$, $N_r = 2$. As shown in the Figure, the GSM IF detection can obtain good performance near the ML detection. There is only about 1 dB gap at BER 10^{-3}.

Figure 4 shows the BER performance of the ML detection and IF detection for a GSM system with $N_t = 5$, $N_r = 2$. Comparing Fig. 3 with Fig. 4, it can be seen that the performance of both schemes gets better by increasing the number of receive antennas.

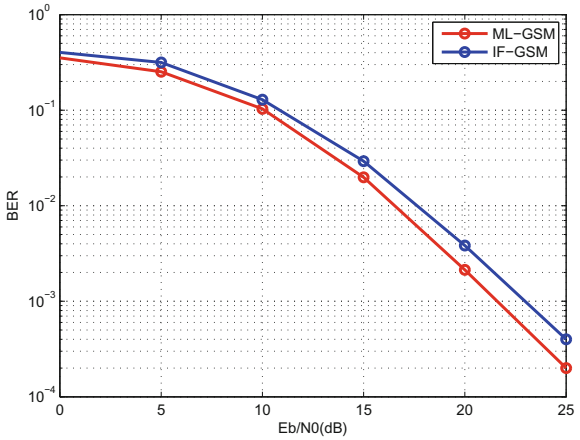


Fig. 3. BER comparison between ML and IF detection schemes at 6 bits/s/Hz, which $N_t = 4$, $N_r = 2$, $N_a = 2$.

Figure 5 shows the performance comparison of several detection schemes respectively for a GSM system with $N_t = 4$ and $N_r = 4$. It can be seen from the figure that the performance of the ML detection is better than the others. The IF detection achieves significantly better performance than the MMSE and ZF detection in spite of a little worse performance than the ML detection. Meanwhile, the MMSE-IF detection with MMSE algorithm to detection the active antenna combinations can not achieve performance improvement from the IF

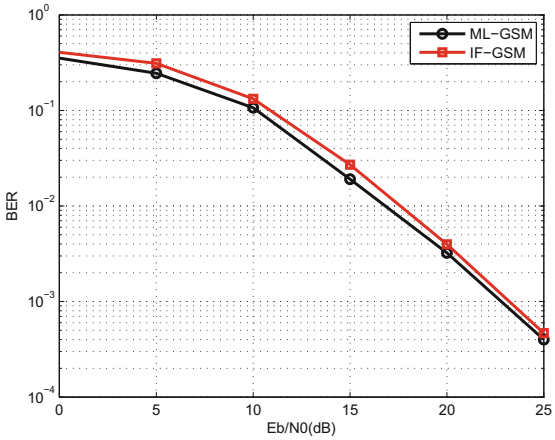


Fig. 4. BER comparison between ML and IF detection schemes at 6 bits/s/Hz, $N_t = 5$, $N_r = 2$, $N_a = 2$.

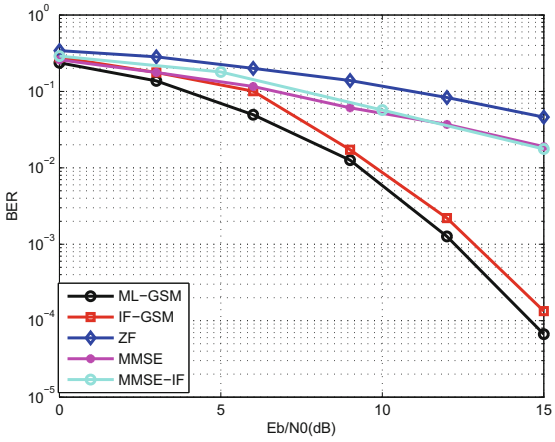


Fig. 5. Performance comparison of several detection schemes at 6 bits/s/Hz, $N_t = 4$, $N_r = 2$, $N_a = 4$.

receiver, just as what we talk about above. The IF detection scheme well balances the relations between MMSE, ZF and ML detection with complexity and performance.

5 Conclusion

Since traditional detection scheme of GSM systems, such as the ZF detection, the MMSE detection and the optimal ML detection cannot get good tradeoff between complexity and performance. To solve the problem, We design a scheme based

on IF algorithm, which utilize IF receiver to decode the transmitted symbols for all possible antenna combinations, and then compare the Euclidean distances between the detected symbols and the received signals. Finally, the symbol vector with minimum Euclidean distance and the corresponding antenna combination are detected as the decided symbols and transmit antenna combination. Monte Carlo Simulation is carried out to test the BER performance. The results shows that the designed detection scheme achieves obviously better performance than the ZF and the MMSE detection with relatively lower complexity.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (Grant No. 61301124, 61471075, 61671091), the University Innovation Team Construction Plan of Smart Medical System and Core Technology in Chongqing.

References

1. Arapoglou, P.D., Liolis, K., Bertinelli, M., Panagopoulos, A., Cottis, P., Gaudenzi, R.D.: MIMO over satellite: a review. Commun. Surv. Tutorials **13**, 27–51 (2011)
2. Larsson, E.G., Edfors, O., Tufvesson, F., Marzetta, T.L.: Massive MIMO for next generation wireless systems. IEEE Commun. Mag. **52**, 186–195 (2014)
3. Yang, P., Renzo, M.D., Xiao, Y., Li, S.: Design guidelines for spatial modulation. Commun. Surv. Tutorials IEEE **17**, 6–26 (2015)
4. Vo, B.T., Nguyen, H.H., Quoc-Tuan, N.: High-rate space-time block coded spatial modulation. In: 2015 International Conference on Advanced Technologies for Communications (ATC), pp. 278–282. IEEE Press, Ho Chi Minh City (2016)
5. Anoh, K.O.O., Jones, S.M.R., Abd-Alhameed, R.A.A.: A simple space-time coding technique for wireless communication systems. In: 2015 Internet Technologies and Applications, pp. 405–410. IEEE Press, Wrexham (2015)
6. Renzo, M.D., Haas, H., Ghrayeb, A., Sugiura, S., Hanzo, L.: Spatial modulation for generalized mimo: challenges, opportunities, and implementation. Proc. IEEE **102**, 56–103 (2013)
7. Narasimhan, T.L., Chockalingam, A.: On the capacity and performance of generalized spatial modulation. IEEE Commun. Lett. **20**, 252–255 (2016)
8. Wang, C., Cheng, P., Chen, Z., Zhang, J.A., Xiao, Y., Gui, L.: Near-ML low-complexity detection for generalized spatial modulation. IEEE Commun. Lett. **20**, 618–621 (2016)
9. Wang, J.T., Jia, S.Y., Song, J.: Generalised spatial modulation system with multiple active transmit antennas and low complexity detection scheme. IEEE Trans. Wirel. Commun. **11**, 1605–1615 (2012)
10. Zhan, J., Nazer, B., Erez, U., Gastpa, M.: Integer-forcing linear receivers. IEEE Trans. Inf. Theory **60**, 7661–7685 (2014)
11. Zhan, J., Nazer, B., Erez, U., Gastpa, M.: Integer-forcing linear receivers: a new low-complexity MIMO architecture. In: 2010 IEEE 72nd Vehicular Technology Conference-Fall, pp. 1–5. IEEE Press, Ottawa (2010)
12. Zhan, J., Nazer, B., Erez, U., Gastpar, M.: Integer-forcing linear receivers. In: 2010 IEEE International Symposium on Information Theory, pp. 1022–1026. IEEE Press, Austin (2010)

13. Sakzad, A., Harshan, J., Viterbo, E.: Integer-forcing MIMO linear receivers based on lattice reduction. IEEE Trans. Wirel. Commun. **12**, 4905–4915 (2013)
14. Sakzad, A., Harshan, J., Viterbo, E.: On complex LLL algorithm for integer forcing linear receivers. In: Communications Theory Workshop (AusCTW), pp. 13–17. IEEE Press, Adelaide (2013)



Design of Dynamic TDMA Protocols for Tactical Data Link

Lei Liu(✉), Yiming Liu, Zhaowei Wang, and Chunxu Liu

China Academy of Electronics and Information Technology,
Beijing 100041, China
liulei_ioa@163.com

Abstract. The Link-16 tactical data link is a wireless data broadcasting network using fixed TDMA protocol. However, the fixed TDMA protocol cant adjust the time slot adaptively according to users demand, resulting in waste of network resources. In this paper, two dynamic TDMA protocols for tactical data link is studied, which are designed to meet the transmission requirements of some network nodes to send large quantities of messages abruptly. In the reservation-based dynamic TDMA protocol, dynamic reserved time slots are allocated by the master node according to the real-time level requirements and the time-out degree of the messages. In the contention-based dynamic TDMA protocol, other nodes can compete with the time slot when there is no data transmission requirement in itself fixed time slot. The two dynamic TDMA protocols improves the utilization rate of time slots and increase the throughput of the network. The simulations results show that two dynamic TDMA protocols are able to achieve much higher throughput, lower average delay than the fixed TDMA protocol.

Keywords: Tactical data link · Fixed TDMA protocol
Reservation-based dynamic TDMA protocol
Contention-based dynamic TDMA protocol

1 Introduction

Tactical data link is the basic means to realize the information battlefield [1]. The transmission of the battlefield situation, reconnaissance information, the coordination of weapons and other co-formatted messages [2] are inseparable from the support of the data link system. Therefore, the tactical data link has significance research value. However, battlefield wireless network has the characteristics of non-center, large capacity, high mobility, long distance, distributed and so on [3], which pose research challenges in the area of multiple access control (MAC) protocol in tactical data link system.

The MAC protocol is used to coordinate the access of the shared wireless channel among network nodes [4], which is very important in the area of tactical data link. Generally speaking, MAC protocol is divided into scheduling-based

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 166–175, 2018.
https://doi.org/10.1007/978-3-319-78130-3_18

MAC protocol and contention-based MAC protocol according to whether the resource is pre allocated or not. In addition, the scheduling-based MAC protocol includes TDMA [5–7], FDMA, CDMA and so on. And the contention-based MAC protocol includes ALOHA based MAC protocol [8] and handshaking based MAC protocol [9,10].

The Link-16 tactical data link is a wireless data broadcasting network [11] using time division multiple access (TDMA) [12,13]. Each node take turns to occupy a time slot to broadcast its own platform information, according to the network management regulations [14,15]. When not broadcasting, the member in the network receive broadcast information from other members. However, fixed TDMA protocol cant adjust the time slot adaptively according to users demand, resulting in waste of network resources. To solve this problem, it is necessary to study dynamic TDMA protocols.

In this paper, two dynamic TDMA protocols for tactical data link is studied, which are designed to meet the transmission requirements of some network nodes to send large quantities of messages abruptly. The two dynamic TDMA protocols improves the utilization rate of time slots and increase the throughput of the network.

The rest of this paper is organized as follows. In Sect. 2, we briefly introduce the original fixed TDMA protocol scheme. Then we respectively propose a reservation-based dynamic TDMA protocol and a contention-based dynamic TDMA protocol in Sect. 3 and Sect. 4. Section 5 describes the simulations that were carried out. Finally, we give a conclusion in Sect. 6.

2 Original Fixed TDMA Protocol Overview

In the fixed TDMA protocol, all the time slots are allocated to each node before the network runs according to the node's business requirements. For a network with N nodes, the length of the frame used by the TDMA protocol is N time slots, and each node is divided into unique slots, as shown in Fig. 1. Because there is only one node can be accessed in each time slot, there is no threat of collisions for any data of nodes. Each node in the network occupies one time slot to send data in one time frame period, and in other time slots to receive data. The interval at which any node send data is a time frame period. Once every network node of a cycle is sent once, it enters the next time frame period, and so on, until the entire network stops running.

The fixed TDMA protocol scheme is simple to implement and does not require overhead network resources for new slot allocation. At the same time, since the time slots used by each node are fixed and do not overlap each other, there is no interference in the network. Each node in the network occupies one time slot to send data in one time frame period, and receive data in other time slots. If a node or several nodes have a large number of messages to be sent, which have special transmission needs. Since each node can only send data once per time frame period and the number of message words sent in each slot is limited, so a large number of messages are stored in the buffer of the network

node which have a large number of messages to be sent. The messages to be sent must wait until the next time frame period. If timeout, these messages will be deleted. Therefore, the disadvantage of fixed TDMA protocol is the low channel utilization, which cant adaptively adjust the time slot according to users demand, resulting in waste of network resources. To solve this problem, it is necessary to study the dynamic TDMA protocol.

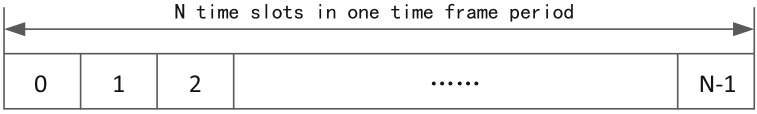


Fig. 1. The frame structure of the fixed TDMA protocol

3 A Reservation-Based Dynamic TDMA Protocol

Based on the fixed TDMA protocol, this section designs a reservation-based dynamic TDMA protocol to meet the transmission requirements of some network nodes to send large quantities of messages abruptly. This dynamic TDMA protocol avoids the deletion of large number of messages because of timeout, reduces the packet loss rate and improves the network throughput. In the network, a master node is needed to manage and allocate dynamic reserved time slots, while other nodes act as slave nodes.

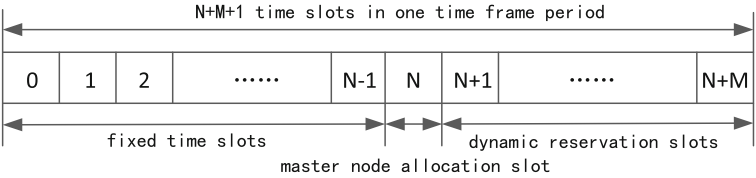


Fig. 2. The frame structure of the reservation-based dynamic TDMA protocol

The frame of the reservation-based dynamic TDMA protocol is divided into three parts: fixed time slots, master node allocation slot and dynamic reservation slots, as shown in Fig. 2. If there are N nodes in the network, there are N fixed time slots in a time frame period, and each node occupies a fixed time slot. The master node allocation slot is occupied by the master station to send the message about dynamic reserved time slots. Dynamic reservation slots are assigned to network nodes that require additional time slots during this time frame period. This section assumes that there are M slots reserved as dynamic allocation slots during the time frame period.

In the allocation process of dynamic reserved time slot, the four-step management strategy is adopted which is slot application-allocation-occupancy-release. The principle is as follows.

When the network is initialized, the master node and the slave nodes numbered 0, 1, and 2... N−1, in which number 0 is the master node, 1, 2... N-1 are the slave nodes. There are N + M + 1 time slots in one time frame period. The time slot number is 0, 1, 2... N + M. In the fixed slot section, the master node numbered 0 occupies slot 0, the slave node numbered 1 occupies slot 1, and so on. All the fixed time slots are assigned to the corresponding number of nodes. The dynamic reservation slots are not allocated temporarily. When the master node receives the capacity request message, the master node send the dynamic reservation slot allocation message using the N slot. All reserved slots after N are used as the master node to dynamically allocate the time slots to the slave nodes.

Once the network starts, each node sends its own message in the corresponding fixed time slot. The maximum number of message sent per time slot is a fixed value. When the slave node sends data, it calculates whether the message in the sending queue can be sent in its fixed time slot. If cant, the remaining life of all message words in the sending queue is calculated. Then the number of message words to be timed out is calculated. When it is greater than 0, the number of dynamic reservation slots required to apply is calculated. And the slot reservation number along with data messages sent in fixed time slot. After receiving the data packet, the master node analyzes the number of additional time slots required by the slave node.

When all nodes transmit data packet in the fixed time slots, the master node calculates the demand for reserved time slots from the slave nodes and determines the allocation strategy of the reserved time slot according to the demand and the remaining ratio. If there is enough reserved time slots, the reserved time slots is allocated according to the number of applications requested by the slave node. If the number of reserved slots is less than the number of slots requests from the slave node, the reserved slots are allocated according to the real-time level requirements and the time-out degree of the messages. The reserved slots are firstly allocated to the nodes who have high real-time services and high time-out degree messages. After the reserved slots are successfully assigned, a time slot allocation message is generated and sent in the master allocation slot N to inform the corresponding slave nodes how to occupy the reserved time slot.

After receiving the time slot allocation message from the master node, the slave node sets itself time frame structure according to the starting position and length of the allocated reserved slot. And then the data packet is retransmitted when the time slots arrive. When all time slots are used in this time frame period, slave nodes release the occupied dynamic reservation slots so that the reserved slots can be applied and used by other slave nodes in the next time frame period.

4 A Contention-Based Dynamic TDMA Protocol

In the fixed TDMA protocol and the reservation-based dynamic TDMA protocol, a fixed time slot is assigned for each node in the time frame period, which can only be occupied by the corresponding node. Even if the node does not have transmission requirement in the corresponding time slot, the slot cant be occupied by other nodes who have transmission requirement. In view of the above questions, a dynamic TDMA protocol based on contention is designed, which allows other nodes to compete with the time slot when there is no data transmission requirement in itself fixed time slot. The contention-based dynamic TDMA protocol improves the utilization rate of the time slot and increase the throughput of the network.

The contention-based dynamic TDMA protocol is designed based on the fixed TDMA protocol and MACA protocol, which is used to avoid conflicts caused by hidden terminals. The protocol assigns a fix time slot to each node in the network. The fixed time slot is the main time slot of each node, and the node is the master node of the time slot.

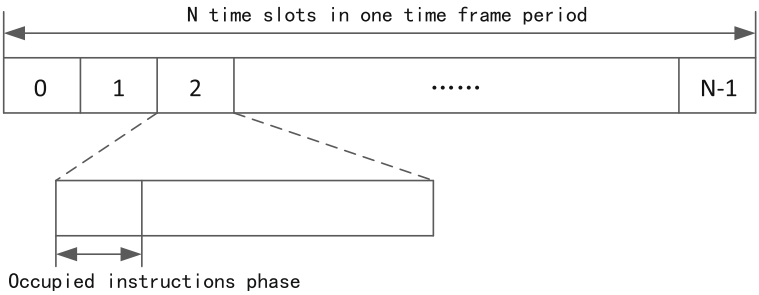


Fig. 3. The frame structure of the contention-based dynamic TDMA protocol

Figure 3 shows the frame structure of the contention-based dynamic TDMA protocol. A time frame period is divided into N time slots. N is the number of nodes in the network. A main time slot is assigned to each node at the time of network initialization. At the beginning of each time slot, there is an occupied instructions to indicate whether the master node needs to use this time slot. As long as the node has transmission requirement, its main time slot can be used itself. When the node does not need to use its main time slot, other nodes can occupy the time slot by means of contention.

When the node needs to use the main time slot, the main time slot is divided into occupied instructions phase and data phase, the frame structure is shown in Fig. 4. The Occupy packet is sent during the occupied instructions phase of the main slot. At the same time the other nodes receive the Occupy packet which indicts the current slot is used by the master node. In the data phase, the master node send data packet by the way of unicast, multicast or broadcast.

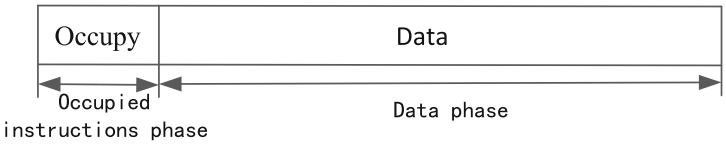


Fig. 4. The slot structure when the node uses its main slot

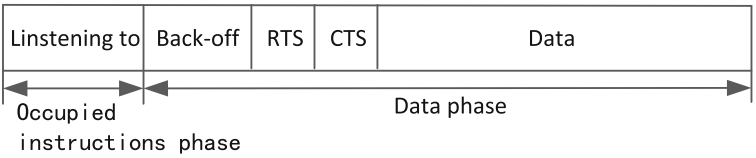


Fig. 5. The slot structure occupied by other nodes when the node does not use its main slot

When the node does not use the main time slot, that is, other nodes in the occupied instruction phase does not receive the Occupy packet, the main slot is divided into the occupied instruction phase, contention phase and data phase. The frame structure is shown in Fig. 5. No packets are sent during the occupied instruction phase of the main slot. Other nodes listen to the time slot and find that the current time slot is not used by the master node. The nodes who need to send data can compete for the current time slot through the random back-off mechanism and RTS/CTS interaction during the contention phase. The node who compete successful can send data packets in the data phase of the current slot. If the contention fails, data packets cant be sent during the data phase of the current slot. In the data phase, the competing nodes can choose to send data by unicast, multicast or broadcast. When a source node S needs to send data, the RTS is sent after a random back-off in the contention phase. If the source node S will send data in unicast mode, the destination node D responds to the CTS after receiving the RTS. If the source node S needs to send the data in multicast or broadcast way, one of the destination nodes is set up in the RTS to reply to the CTS. The source node S sends data after receiving the CTS.

5 Simulations and Results

5.1 Simulation Model

The network topology of the simulation is shown in Fig. 6, 50 nodes are randomly distributed in a circular area with a diameter of 300 km. The maximum communication distance of the nodes is the diameter of the circular area. That is, all nodes can listen to each other. The 50 nodes generate data packets at different data rates according to the Poisson distribution and send data packets to any other node in the way of unicast, multicast or broadcast. The network data

rate is set to 200 kbps. To highlight the performance of multiple access protocol in data link layer, it is assumed that the physical layer error bit rate is zero in this simulation.

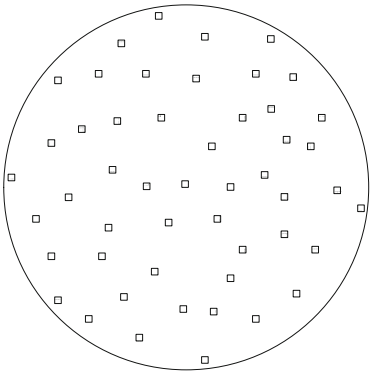


Fig. 6. Network topology

In the fixed TDMA protocol simulation, a time frame period takes 0.5 s. And a time frame period is divided into 50 time slots which corresponds to 50 nodes. Therefore each time slot is 10 ms. The maximum length of the packet sent per time slot is 1500 bit.

In the reservation-based dynamic TDMA protocol, a time frame period takes 0.5 s. And a time frame period is divided into 100 time slots. Therefore each time slot is 5 ms. The first 50 time slots (No. 0–49) are fixed time slots, corresponding to 50 nodes respectively. No. 50 slot is the master node allocation slot. And No. 51–99 slot is dynamic reservation slots. The maximum length of the packet sent per time slot is 800 bits.

In the contention-based dynamic TDMA protocol, a time frame period takes 0.5 s. And a time frame period is divided into 50 time slots. Therefore each time slot is 10 ms. If the master node has data to be sent, the 16 bit Occupy packet is sent directly at the beginning of the time slot. Then the data packet is sent subsequently in current time slot. The maximum length of the packet the master node sent in its main time slot is 1500 bits. Other nodes are listening in the first 1 ms of the time slot. If the master node has no data to send in its main time slot, there is no Occupy packet can be listened by other nodes. Then other nodes compete for the right to use current slot. The back-off time is randomly chosen in 0–1 ms. RTS and CTS packet are taken 72 bit respectively. And the maximum length of the packet sent by the completion node in one time slot is set to 800 bit.

5.2 Simulation Results

In this section, the performance of the fixed TDMA protocol, reservation-based dynamic TDMA protocol and the contention-based dynamic TDMA protocol is

examined. Offered load, throughput and average delay are defined in the following formula (1)–(3).

$$offered\ load = \frac{total\ data\ packets\ transmitted * data\ packet\ length}{data\ rate * simulation\ duration} \quad (1)$$

$$throughput = \frac{total\ data\ packets\ receiveded * data\ packet\ length}{data\ rate * simulation\ duration} \quad (2)$$

$$average\ delay = \frac{\sum time\ of\ receiving\ data - time\ of\ transmitting\ data}{total\ data\ packets\ received} \quad (3)$$

The simulation results of the offered load-throughput of the three protocols are shown in the Fig. 7. As can be seen from Fig. 7, when the offered load is low, the throughput of the three protocols increases with the increase of the offered load. When the load increases to a certain value, the throughput reaches the maximum and remains essentially unchanged. At the same time, as shown in the diagram, the throughput of the fixed TDMA protocol is relatively low. Because the slot division is fixed, the node can only send data in its own time slot. When the node has no data transmission demand at its main time slot, it leads to no data transmission in the time slot, resulting in low utilization of network resources. The throughput of the contention-based dynamic TDMA protocol is relatively bigger. Because when the node has no data transmission demand in its main time slot, other nodes can use RTS/CTS to compete this time slot and send data, to avoid the waste of time slots. The throughput of the reservation-based dynamic TDMA protocol is largest of the three protocols. The master node allocates the dynamic reservation slots according to transmission demands of nodes, which improve the utilization ratio of the system resource.

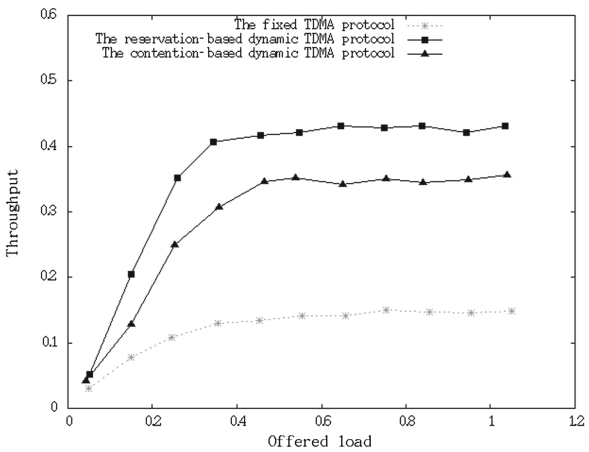


Fig. 7. Offered load C throughput of three protocols

The Fig. 8 shows the simulation results of the offered load- average delay of the three protocols. As can be seen from Fig. 8, the average delay of the reservation-based dynamic TDMA protocol is the lowest of the three protocols. When the offered load is low, the average delay of the contention-based dynamic TDMA protocol is only lower a little than the fixed TDMA protocol, due to back-off time and RTS/CTS interactive messages. When the offered load increases, the average delay of the contention-based dynamic TDMA protocol becomes more and more lower than that of the fixed TDMA protocol, which is due to the increase of the network throughput, thus reducing the waiting time of packet.

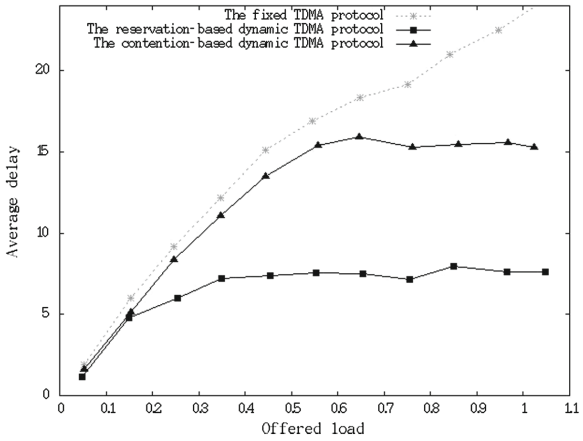


Fig. 8. Offered load C average delay of three protocols

6 Conclusion

In this paper, two dynamic TDMA protocols for tactical data link are studied, which are designed to meet the transmission requirements of some network nodes to send large quantities of messages abruptly. In the reservation-based dynamic TDMA protocol, dynamic reserved time slots are allocated by the master node according to the real-time level requirements and the time-out degree of the messages. In the contention-based dynamic TDMA protocol, other nodes can compete with the time slot when there is no data transmission requirement in itself fixed time slot. The two dynamic TDMA protocols improves the utilization rate of time slots and increase the throughput of the network. The simulations results show that two dynamic TDMA protocols are able to achieve much higher throughput, lower average delay than the fixed TDMA protocol.

Acknowledgement. This research is supported in part by the National Science Foundation of China, under Grand No. 61501415.

References

1. Jaime, C.A., Edgar, P., John, H.: Augmenting attack graphs to represent data link and network layer vulnerabilities. In: IEEE Military Communications Conference, MILCOM 2016, pp. 1010–1015. IEEE Conference Publications, Baltimore (2016)
2. North, R., Browne, N., Schiavone, L.: Joint tactical radio system-connecting the GIG to the tactical edge. In: IEEE Military Communications Conference, pp. 1–6. IEEE Press, Washington, DC (2006)
3. CJCSM 6120.01C.: Joint multi-tactical data link (TDL) operating procedures (JMTOP). Joint staff, Washington, DC (2002)
4. Rao, A., Stoica, I.: An overlay MAC layer for 802.11 networks. In: The 3rd International Conference on Mobile Systems, Applications, and Services, New York, pp. 135–148 (2005)
5. Kanzaki, A., Uemukai, T., Hara, T., et al.: Dynamic TDMA slot assignment in ad hoc networks. In: 17th International Conference on IEEE Advanced Information Networking and Applications, AINA, pp. 330–335 (2003)
6. Arivudainambi, D., Rekha, D.: Broadcast scheduling problem in TDMA ad hoc networks using immune genetic algorithm. Int. J. Comput. Commun. Control **8**(1), 18–29 (2012)
7. Gronkvist, J., Karlsson, J., Sterner, U., et al.: Adaptation delay and its impact on application performance for TDMA ad hoc networks. In: The 11th Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net), pp. 55–60. IEEE Conference Publications (2012)
8. Norman, A.: The Aloha system. In: Proceedings of AFIPS 1970 Fall Joint Computer Conference, pp. 281–285 (1970)
9. Qi, W., Katia, J.R., Yongjun, X., et al.: TDMA versus CSMA/CA for wireless multihop communications: a stochastic worst-case delay analysis. IEEE Trans. Industr. Inform. **13**(2), 877–887 (2016)
10. Jin, L., Huang, D.F.: A slotted CSMA based reinforcement learning approach for extending the lifetime of underwater acoustic wireless sensor networks. Comput. Commun. **36**(9), 1094–1099 (2013). Original Research Article
11. Kihyoung, K., Kyuman, L., Jaesung, L.: A RELNAV enhancement for reducing cumulative position error in Link-16 without GRU. In: 2016 IEEE Military Communications Conference, MILCOM 2016, pp. 942–947. IEEE Conference Publications, Baltimore (2016)
12. Cruz, C.I.: NETWARS based study of joint stars link-16 network. Air University (2004)
13. Hao, C., JingMing, K., ZunWen, H.: Modeling and simulation of link 16 network using OPNET. In: Proceedings of the International Conference on Communication and Information (2005)
14. Link-16-Product-Development-Group. SISO Standard for Link 16 Simulations. http://www.sisostds.org/
15. Tong, L., Xiaojing, L., Zheng, D.: An optimal design of time division multiple access protocol for data link network. In: Electronic Information and Communication Technology (ICEICT), pp. 158–161. IEEE Conference Publications, Harbin (2016)



Estimation Under Staggered Frame Structure for Multi-cell Multi-user Massive MIMO Systems

Mengjun Zhang(✉), Gang Xie, and Jinchun Gao

Beijing University of Posts and Telecommunications, Beijing, China
243173821@qq.com, xiegang@bupt.edu.cn

Abstract. This paper describes an interesting method to address the problem of pilot contamination, which adopts the staggered frame structure. The core idea is that different users in the same cell send pilot symbols in different time intervals, and users who send pilot sequences simultaneously in different cells use orthogonal pilot sequences. Furthermore, two channel estimation methods were utilized, which respectively are the Orthogonal Projection Least Square (OPLS) method and the Subtract Interference Least Square (SILS) method. Both two methods don't directly estimate the whole channel parameters. Alternatively, they estimate the sectional channel response that is orthogonal to the other users' channel response. The OPLS method is implemented by orthogonal projection operation, whose complexity is higher. So we proposed the SILS method, which combines signal detection to reduce the computational complexity. The simulation results show the staggered frame structure and the channel estimation methods can solve the pilot contamination effectively.

Keywords: Massive MIMO · Pilot contamination · Channel estimation
Staggered frame

1 Introduction

Massive MIMO makes a great improvement about the communication system's performance due to a great quantity of antennas deployed in the base station [1]. The extra antennas make the energy focus on much smaller space than conventional communication systems. The detailed description is given in the reference [2, 3]. In Massive MIMO system, it does not require the rich scattering environment. All users make use of the same time-frequency resource block, so that the resource allocation is simplified. And the user terminals employ cheap single-antenna. When the base station acquires the perfect channel state information (CSI), it can completely suppress the inter-user interference if the base station's server antenna tends to be infinite. Hence, the research activities are motivated to a great degree on channel estimation in Massive MIMO.

The eigenvalue-decomposition (EVD) method was proposed in [4], it estimates the channel by using the received signals' sample covariance matrix. However, a great many of antennas makes the EVD scheme quite costly. Moreover, the sample

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 176–185, 2018.
https://doi.org/10.1007/978-3-319-78130-3_19

covariance matrix can't fully replace real covariance matrix, which will result in an inaccurate estimation. Furthermore, if use pilot symbols to estimate the channel, the phenomenon of pilot contamination will not be ignored. To solve this problem, Haifan presented a coordinated channel estimation techniques in [5]. They utilized the user channels' covariance information and adopted a covariance-aware pilot assignment strategy in different cells. However, this method requires channel covariance matrix to satisfy specific conditions. And it is not easy to implement. Reference [6] proposed a data-aided channel estimation scheme. They investigated a scheme that partially decoded data and pilot symbols both are used to aid channel estimation. But iterative operation also increases the computational complexity.

Reference [7] addressed the pilot contamination problem from frame structure point of view, it proposed the conception of the staggered frame structure. Under the staggered frame structure, user terminals belonging to the same cell begin their frames at staggered time points, as well as these users transmit their pilot signals at non-overlapped time periods. But they only investigated the Massive MIMO system with single cell. This paper extended the staggered frame structure to multicell multiuser Massive MIMO systems. In our scheme, each cell uses the same frame structure as the paper [7]. Users belonging to different cells could start their frames and send pilot at the same time intervals but the pilot sequences must be orthogonal to avoid pilot contamination. Then the Orthogonal Projection Least Square (OPLS) and the Subtract Interference Least Square (SILS) were investigated. The OPLS method projects the received signal to the orthogonal complement space of other interfering user channels before utilizing the least squares to estimate the channel parameters. But its computational complexity is high. We proposed the SILS method to reduce the complexity. In this method, the interference data signals will be detected and subtracted from the received signal before the channel estimation.

In this paper, the notations conform to the following descriptions.

We use lowercase bold-type letter (e.g., $\mathbf{x}$) to represent the vector, and uppercase, bold-type letter (e.g., $\mathbf{X}$) to indicate matrix. $\mathbf{X}^T, \mathbf{X}^H$ and $\mathbf{X}^*$ respectively indicate the transpose, the conjugate transpose and conjugate of the matrix $\mathbf{X}$. Let $\ell(\mathbf{X})$ denote the column space of the matrix $\mathbf{X}$. Specifically, $\mathbf{I}_M$ means $M \times M$ identity matrix. $C^{m \times n}$ indicates the space of $m \times n$ matrix with complex entries. Furthermore, we denote orthogonal projection onto $\ell(\mathbf{X})$ with the matrix $\Pi_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^H\mathbf{X})^{-1}\mathbf{X}^H$ and $\Pi_{\mathbf{X}}^{\perp} = \mathbf{I} - \Pi_{\mathbf{X}}$.

2 Multi-cell Multi-user Massive MIMO System and Frame Structures

In this chapter, we first present the multicell multiuser Massive MIMO system model and introduce the characteristics of synchronized frame structures and staggered frame structures in detail. Meanwhile a novel pilot transmission strategy is proposed for staggered frame structure.

2.1 Multi-cell Multi-user Massive MIMO System

Massive MIMO communication system discussed in this paper has L cells. There are K user terminals and one base station spread over every cell. All terminals are single-antenna device, as well as the base station deploys M antennas, and $M \gg K$. Furthermore, we assume all user terminals share the same bandwidth. In order to easy to express, we specify the first cell as the targeted cell. So the targeted base station's received signal, which is a $M \times 1$ vector, is

$$\mathbf{y} = \sum_{i=1}^{L} \sum_{k=1}^{K} \mathbf{g}_k^i x_{k,t}^i + \mathbf{n}, \tag{1}$$

where $x_{k,t}^i$ is the signal symbol sent by the k-th user terminal in the i-th cell at the time instant t; $\mathbf{n}$ is a $M \times 1$ vector that represents complex Gaussian additive white noise whose elements' mean is zero and variance is σ_n^2; and $\mathbf{g}_k^i$ is a $M \times 1$ vector that indicates the channel response between the k-th user in the i-th cell and the targeted BS. The channel vector's element $g_{mk}^i \triangleq [\mathbf{g}_k^i]_m$ indicates the channel parameter between the k-th user terminal in the i-th cell and the m-th antenna deployed on targeted base station, and is given by

$$g_{mk}^i = h_{mk}^i \cdot \sqrt{\beta_k^i},\ m = 1, \cdots, M \quad i = 1, \cdots, \mathrm{L}\ k = 1, \cdots, K, \tag{2}$$

where h_{mk}^i is the fast fading coefficient obeyed the Gaussian distribution. $\sqrt{\beta_k^i}$ contains the log-normal shadow fading and geometric attenuation, and we assume it is a prior constant independent of the antenna index m. These assumptions are based on the condition that the distance from the BS to the user is much more far than the base station's antenna spacing, and the value of β_k^i is going to be very slowly changing in time.

2.2 Synchronized Frame Structure and Pilot Contamination

This section introduces the typical synchronized frame structure used in multicell multiuser Massive MIMO system. Furthermore, we analyze the cause of pilot contamination.

As exhibited from Fig. 1(a), the coherence time interval T consists of three phases which is the pilot symbols sending, the uplink data sending and the downlink data sending. In the pilot sending phase, user terminals transmit pilot symbols $\mathbf{s}_k^i = [s_{k,1}^i, s_{k,2}^i, \cdots, s_{k,\tau}^i]^T \in C^{\tau \times 1}$ to the base station synchronously. $s_{k,t}^i$ is the pilot symbol sent by the k-th user terminal in the i-th cell at the time instant t, $k = 1, 2, \cdots, K$, $i = 1, 2, \cdots, L$ and τ is the pilot length.

In general, orthogonal pilot sequences are adopted in one cell, and the same pilot symbols are reused in all L cells. In other words, if define $\mathbf{S}^i \triangleq [\mathbf{s}_1^i, \mathbf{s}_2^i, \cdots, \mathbf{s}_K^i]^T \in C^{K \times \tau}$ as pilot symbols used by all user terminals located in i-th cell, then $\mathbf{S}^i \mathbf{S}^{iH} = \mathbf{I}_K$ and $\mathbf{S}^1 = \mathbf{S}^2 = \cdots = \mathbf{S}^L$. Under this scenario, the base station's received signal matrix $\mathbf{Y} \in C^{M \times \tau}$ is

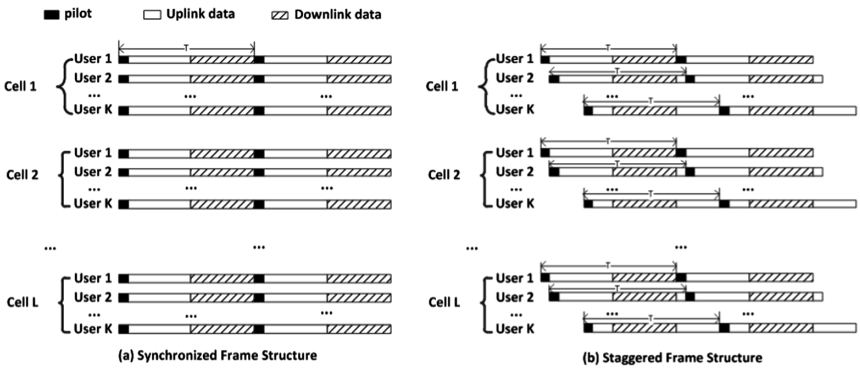


Fig. 1. Synchronized and staggered frame structure in multicell multiuser Massive MIMO system. T represents the coherence time interval.

$$\mathbf{Y} = \sum_{i=1}^{L} \mathbf{G}^i \mathbf{S}^i + \mathbf{N}, \tag{3}$$

where $\mathbf{G}^i \triangleq [\mathbf{g}_1^i, \mathbf{g}_2^i, \cdots, \mathbf{g}_K^i] \in C^{M\times K}$ indicates the channel matrix between K user terminals located in the i-th cell and targeted BS, $\mathbf{N} \in C^{M\times\tau}$ is the additive Gaussian complex noise in the targeted base station. If utilize the least square (LS) approach, the base station will get the channel estimation matrix as follows

$$\widehat{\mathbf{G}}^1 = \mathbf{Y}\mathbf{S}^{1H}\left(\mathbf{S}^1\mathbf{S}^{1H}\right)^{-1} = \mathbf{G}^1 + \sum_{i\neq 1}^{L} \mathbf{G}^1 + \mathbf{N}\mathbf{S}^{1H}\left(\mathbf{S}^1\mathbf{S}^{1H}\right)^{-1}. \tag{4}$$

We can see from the above formula, the LS method and the reused operation of pilot sequences makes the desired channel estimate leaked by the interfering channels, which give rise to "pilot contamination".

2.3 Staggered Frame Structure

Figure 1(b) illustrates the staggered frame structure applied to multicell multiuser Massive MIMO system. As shown here, user terminals belonging to the same cell begin their frames at staggered time points, and these user terminals transmit their pilot symbols at non-overlapped time periods. And the pilot sequences could be the same. User terminals belonging to the different cells could begin their frames at the same time points, but the pilot sequences these users transmitted at the same time periods should be orthogonal. The channel estimations of all the users should be accomplished before the downlink data phase started, and all users' downlink data phases should share the same time slot. Owing to the fact that it has difficulty for the base station to receive the uplink data and transmit the downlink data at the same time. In addition, there would also have serious interference.

In the staggered frame structure, we assure that every cell's k-th user send orthogonal pilot symbols to the targeted base station. During the pilot phase of the k-th user, the k-th user in every cell transmit $\mathbf{x}_k^i = \mathbf{s}_k^i = [s_{k,1}^i, s_{k,2}^i, \cdots, s_{k,\tau}^i]^T \in C^{\tau\times 1}$ ($\forall k = 1, 2, \cdots, K$ and $\forall i = 1, 2, \cdots, L$) at overlapped time intervals with the power of $p_P = (\mathbf{s}_k^{iH}\mathbf{s}_k^i)/\tau$. While the other user terminals send their uplink data $\mathbf{x}_j^i = [x_{j,1}^i, x_{j,2}^i, \cdots, x_{j,\tau}^i]^T \in C^{\tau\times 1}$ with the power of p_d, where $j = 1, \cdots k-1, k+1, \cdots, K$ and $i = 1, 2, \cdots, L$. Define $\mathbf{S}_k \triangleq \left[\mathbf{s}_k^1, \mathbf{s}_k^2, \cdots, \mathbf{s}_k^L\right]^T \in C^{L\times\tau}$ as pilot sequences used by the users who belonging to different cells and transmit pilot symbols at overlapped time intervals. And the pilot sequences used in the Massive MIMO system satisfy the following properties: $\mathbf{S}_k\mathbf{S}_k^H = (\tau p_p)\mathbf{I}_L$ and $\mathbf{S}_1 = \mathbf{S}_2 = \cdots = \mathbf{S}_K$. Therefore the received signal matrix $\mathbf{Y} \in C^{M\times\tau}$ is written as

$$\mathbf{Y} = \mathbf{G}_k\mathbf{S}_k + \sum_{i=1}^{L} \widetilde{\mathbf{G}}_k^i \widetilde{\mathbf{X}}_k^i + \mathbf{N} = \mathbf{G}_k\mathbf{S}_k + \widetilde{\mathbf{G}}_k\widetilde{\mathbf{X}}_k + \mathbf{N}, \tag{5}$$

where $\mathbf{G}_k \triangleq [\mathbf{g}_k^1, \mathbf{g}_k^2, \cdots, \mathbf{g}_k^L] \in C^{M\times L}$ is the channel matrix of the users who transmits the pilot sequence at the current time instant and

$$\widetilde{\mathbf{G}}_k^i \triangleq [\mathbf{g}_1^i, \mathbf{g}_2^i, \cdots, \mathbf{g}_{k-1}^i, \mathbf{g}_{k+1}^i, \cdots, \mathbf{g}_K^i] \in C^{M\times(K-1)}, \tag{6}$$

$$\widetilde{\mathbf{G}}_k \triangleq \left[\widetilde{\mathbf{G}}_k^1, \widetilde{\mathbf{G}}_k^2, \cdots, \widetilde{\mathbf{G}}_k^L\right] \in C^{M\times L(K-1)}, \tag{7}$$

$$\widetilde{\mathbf{X}}_k^i \triangleq \left[\mathbf{x}_1^i, \mathbf{x}_2^i, \cdots, \mathbf{x}_{k-1}^i, \mathbf{x}_{k+1}^i, \cdots, \mathbf{x}_K^i\right]^T \in C^{(K-1)\times\tau}, \tag{8}$$

$$\widetilde{\mathbf{X}}_k \triangleq \left[\widetilde{\mathbf{X}}_k^{1T}, \widetilde{\mathbf{X}}_k^{2T}, \cdots, \widetilde{\mathbf{X}}_k^{LT}\right]^T \in C^{L(K-1)\times\tau}. \tag{9}$$

The Massive MIMO system under the staggered frame structure will solve the problem of pilot contamination because the pilots sent using the same time intervals are pairwisely orthogonal. And it is worth mentioning that the only possible interference is the other channels that sent the data at the same time, which would affect the accuracy of the estimation. In next section, we adopt Orthogonal Projection Least Square (OPLS) to suppress interference from other users in the Massive MIMO system. Meanwhile, we proposed Subtract Interference Least Square (SILS) method to reduce the complexity of OPLS.

3 Channel Estimation Under Staggered Frame Structure

3.1 Least Square Estimation

For the comparison, we first present the Least Square estimation which is a typical channel estimation method. In the staggered frame structure, the k-th user located in every cell so there are in total L users can be estimated at the same time

$$\begin{aligned}
&\widehat{\mathbf{G}}_k^{LS}\mathbf{Y}\mathbf{S}_k^H\left(\mathbf{S}_k\mathbf{S}_k^H\right)^{-1}\\
&=\mathbf{G}_k\mathbf{S}_k\mathbf{S}_k^H\left(\mathbf{S}_k\mathbf{S}_k^H\right)^{-1}+\widetilde{\mathbf{G}}_k\widetilde{\mathbf{X}}_k\mathbf{S}_k^H\left(\mathbf{S}_k\mathbf{S}_k^H\right)^{-1}+\mathbf{N}\mathbf{S}_k^H\left(\mathbf{S}_k\mathbf{S}_k^H\right)^{-1},\\
&=\mathbf{G}_k+(\tau p_P)^{-1}\widetilde{\mathbf{G}}_k\widetilde{\mathbf{X}}_k\mathbf{S}_k^H+(\tau p_P)^{-1}\mathbf{N}\mathbf{S}_k^H
\end{aligned}\tag{10}$$

The LS method's computational complexity is very low, and its complexity is $O(ML\tau)$. But the channel estimations obtained through this method are not accurate, because the data signals sent from the other users will cause severe interference. Therefore, we investigate OPLS method and SILS method that suppress the interference, are presented next.

3.2 OPLS and SILS Estimation

The proposed OPLS method and the SILS method don't directly estimate the whole channel parameters. Alternatively, they estimate the sectional channel response that is orthogonal to the other users' channel response. So the interference from the other user terminals could be eliminated as much as possible.

Under the staggered frame structure, when the k-th user terminal located in every cell start transmitting their pilot sequences, the other $L(K-1)$ users, who are sending the uplink data signals, whose channel matrix estimations have already been acquired. We denote $\ell\left(\widehat{\widetilde{\mathbf{G}}}_k\right)$ as the subspace spanned by the known user terminals' channel matrix estimation. Then we project the received signal onto the orthogonal complement of $\ell\left(\widehat{\widetilde{\mathbf{G}}}_k\right)$ to get the $\mathbf{Y}'$, and it is written as,

$$\mathbf{Y}'=\Pi^{\perp}_{\widehat{\widetilde{\mathbf{G}}}_k}\mathbf{Y}=\left(\mathbf{I}_M-\widehat{\widetilde{\mathbf{G}}}_k\left(\widehat{\widetilde{\mathbf{G}}}_k^H\widehat{\widetilde{\mathbf{G}}}_k\right)^{-1}\widehat{\widetilde{\mathbf{G}}}_k^H\right)\mathbf{Y}.\tag{11}$$

Then, use the LS method to estimate every k-th user's in total L users' channel parameters from the projected signal matrix $\mathbf{Y}'$. So the output of OPLS method of every k-th user's in total L users' channel matrix estimation can be written as

$$\widehat{\mathbf{G}}_k^{OPLS}=\mathbf{Y}'\mathbf{S}_k^H\left(\mathbf{S}_k\mathbf{S}_k^H\right)^{-1}=(\tau p_P)^{-1}\Pi^{\perp}_{\widehat{\widetilde{\mathbf{G}}}_k}\mathbf{Y}\mathbf{S}_k^H.\tag{12}$$

We can see if the targeted base station has the perfect CSI of the $L(K-1)$ user terminals who transmit uplink data, that is to say, $\widehat{\widetilde{\mathbf{G}}}_k=\widetilde{\mathbf{G}}_k$, the OPLS estimation in Eq. (12) could also be given by

$$\begin{aligned}
&\widehat{\mathbf{G}}_k^{OPLS}=(\tau p_P)^{-1}\Pi^{\perp}_{\widetilde{\mathbf{G}}_k}\mathbf{Y}\mathbf{S}_k^H\\
&=(\tau p_P)^{-1}\Pi^{\perp}_{\widetilde{\mathbf{G}}_k}\left(\mathbf{G}_k\mathbf{S}_k+\widetilde{\mathbf{G}}_k\widetilde{\mathbf{X}}_k+\mathbf{N}\right)\mathbf{S}_k^H.\\
&=\mathbf{G}_k+(\tau p_P)^{-1}\Pi^{\perp}_{\widetilde{\mathbf{G}}_k}\mathbf{N}\mathbf{S}_k^H.
\end{aligned}\tag{13}$$

Equation (13) shows that after the orthogonal projection, the interference from the other user terminals is suppressed entirely. However, it should be mentioned that the computational complexity is higher with the complexity of $O(M^2(LK+\tau))$, because the calculation of the matrix $\mathbf{Y}'$ contains matrix inversion and projection operations.

Next, we will show that if signal detection combined with LS channel estimation, the computational complexity will be drastically reduced. And we call this method SILS method. Assume at the time instant t $(1 \le t \le \tau)$ of pilot phase of the K-th user, every K-th user sends the pilot symbols and the other user (that is the j-th user $\forall j = 1, 2, \cdots, K-1$) sends the uplink data. Then the received signal at the time instant t is

$$\mathbf{y}_t = \mathbf{G}\mathbf{x}_t + \mathbf{n}_t, \tag{14}$$

where $\mathbf{G} = \left[\mathbf{g}_1^1, \mathbf{g}_2^1, \cdots, \mathbf{g}_K^L\right] \in C^{M \times LK}$ indicates the channel matrix from all LK user terminals to the targeted BS,

$\mathbf{x}_t = \left[x_{1,t}^1, x_{2,t}^1, \cdots, x_{K-1,t}^1, s_{K,t}^1, x_{1,t}^2, \cdots, x_{K-1,t}^L, s_{K,t}^L\right]^T \in C^{LK \times 1}$ and $x_{j,t}^i$ is the uplink data symbol sent by the j-th user in the i-th cell at the time instant t, $(\forall j = 1, 2, \cdots K-1$ and $\forall i = 1, 2, \cdots, L)$; $s_{K,t}^i$ is the pilot symbol sent by the K-th user terminal in the i-th cell at the time instant t.

Here we adopt MRC receivers to detect the data symbols. In the pilot phase of the K-th user, we have estimated the other users' channel $\hat{\mathbf{g}}_j^i (\forall i = 1, 2, \cdots L,\ \forall j = 1, 2, \cdots K-1)$. So we can detect the $L(K-1)$ users data signal from the received signal $\mathbf{y}_t$

$$\hat{x}_{j,t}^i = \frac{1}{M} \hat{\mathbf{g}}_j^{iH} \mathbf{y}_t. \tag{15}$$

After detecting the data signal, the data signal can be subtracted from the received signal to obtain

$$\mathbf{y}_t' = \mathbf{y}_t - \sum_{i=1}^{L} \sum_{j=1}^{K-1} \hat{\mathbf{g}}_j^i \hat{x}_{j,t}^i, \tag{16}$$

$$\mathbf{Y}' = \left[\mathbf{y}_1', \mathbf{y}_2', \cdots, \mathbf{y}_\tau'\right]. \tag{17}$$

Using the MRC detection, we take advantage of the following orthogonal relationship about Massive MIMO,

$$\frac{1}{M} \mathbf{g}_{j_1}^{i_1 H} \mathbf{g}_{j_2}^{i_2} = \begin{cases} 1 & i_1 = i_2,\ j_1 = j_2 \\ 0 & i_1 \ne i_2,\ j_1 \ne j_2 \end{cases}. \tag{18}$$

We can see that the maximum complexity of the OPLS channel estimation comes from the orthogonal projection when calculating the $\mathbf{Y}'$. So we combine the data detection with the channel estimation to replace the orthogonal projection operation,. In this method, the other users' data signals can also be subtracted from the received signal. There are only $ML(K-1)\tau$ complex multiplications, so the complexity is very

low. Therefore, the proposed SILS estimation that combined with MRC detection brings a significant reduction in the computational complexity of OPLS estimation.

4 Simulation Results and Discussions

This paper simulates a Massive MIMO system with $L = 4$ cells and every cell contains $K = 4$ user terminals. We consider the first cell as the targeted cell. We simulated the OPLS method to compare the performance, the SILS method and the LS method under the staggered frame structure, LS method and the EVD method mentioned in [4] under the synchronized frame structure. We evaluate the performance of these estimation algorithms according to the normalized mean square error (NMSE) of the users' channel estimation in the targeted cell. And our simulation adopted the BPSK modulation.

We choose $D_{11} = diag\{0.98, 0.78, 0.59, 0.48\}$, $D_{12} = diag\{0.47, 0.36, 0.26, 0.12\}$, $D_{13} = diag\{0.32, 0.25, 0.17, 0.10\}$, and $D_{14} = diag\{0.24, 0.19, 0.12, 0.04\}$. For the pilot-based method, we use $\tau = 4$ symbols for pilots. Under the staggered frame structure, the pilot transmission strategy proposed in paper is used, that is, the user terminals located in the same cell use the same pilot symbols, while the users belonging to different cells but start their frame at the same time point use the pairwisely orthogonal pilot symbols. The pilot power of each user is $p_P = 8$ dB while the uplink data signal power is $p_d = 0$ dB. Furthermore, the number of the antenna array deployed BS $M = 100$ is considered. The uplink time interval is set to be $T = 200$, which consists of the pilot phase and the uplink data phase. We simulate for 10000 successive frames. About the EVD-based method, this paper chose pilot symbol $\tau_{EVD} = 1$ to resolve the multiplicative factor ambiguity.

Figure 2 exhibits the channel estimation NMSE versus the signal to noise ratio. $EVD, N = T$ means that the EVD-based algorithm that we adopt the sample covariance

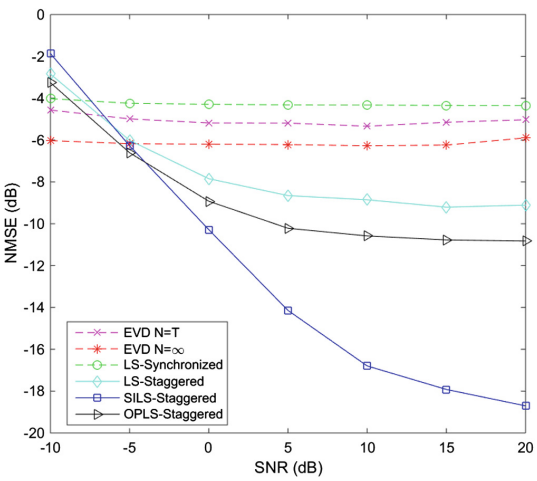


Fig. 2. Channel estimation NMSE versus SNR for $M = 100$, $T = 200$, $p_P = 8$ dB

matrix with the number of samples $N = T$. $EVD, N = \infty$ represents that we assume the covariance matrix is perfect. From the figure, it is found that the performance of LS, OPLS and SILS under the staggered frame structures is better than LS algorithm based on synchronized frame structure and EVD algorithm. Further, the performance of the SILS algorithm is significantly improved.

Figure 3 shows NMSE of these algorithms under the staggered frame structures versus the number of BS antennas M at $SNR = 8$ dB. We can see that the NMSE of the LS algorithm does not change because of the presence of pilot contamination, but the NMSE of our OPLS and SILS methods decreases with the number of BS antennas M increasing. The reason is that the processed received signal, whether by orthogonal projection or subtracting the interference, is closer to the pure received pilot signal, which is advantageous for estimating the real channel response. It is also observed that the performance of SILS is better than the OPLS in the same number of antennas. That's because some of the characteristics of the targeted channel characteristic will be lost when the orthogonal projection operation is performed based on the incompletely correct channel estimation, but the SILS algorithm merely subtracts the received interfering signal and does not affect the channel parameters that we want to be estimated.

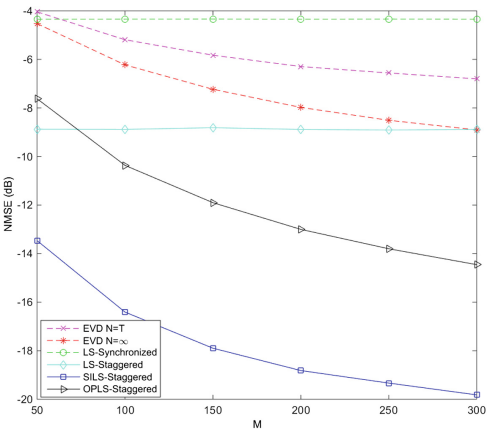


Fig. 3. Channel estimation NMSE versus M for $T = 200$, $SNR = 10$ dB, $p_P = 8$ dB

5 Conclusion

This paper presents the staggered frame structure used in multicell multiuser Massive MIMO systems. In the staggered frame structure, the paper proposes a unique pilot transmission strategy, that is, different users within the same cell transmit the same pilot symbols at different time intervals, but users belonging to different cells send orthogonal pilot symbols at the same time intervals. For the staggered frame structure, we propose that not directly estimate the whole users' channel parameters, alternatively estimate the part of the channel response that is orthogonal to the other users'. So our

method which includes OPLS and SILS greatly inhibited the users' channel interference, making the estimate more accurate.

Acknowledgments. This work was supported by National Natural Science Foundation of China [grant number 61531007].

References

1. Marzetta, T.L.: Noncooperative cellular wireless with unlimited numbers of base station antennas. IEEE Trans. Wirel. Commun. **9**(11), 3590–3600 (2010)
2. Larsson, E.G., et al.: Massive MIMO for next generation wireless systems. IEEE Commun. Mag. **52**(2), 186–195 (2013)
3. Rusek, F., Persson, D., Lau, B.K., et al.: Scaling up MIMO: opportunities and challenges with very large arrays. Mathematics **30**(1), 40–60 (2013)
4. Ngo, B.Q., Larsson, E.G.: EVD-based channel estimation in multi-cell multi-user MIMO systems with very large antenna arrays, pp. 3249–3252 (2012)
5. Yin, H., Gesbert, D., Filippou, M., et al.: A coordinated approach to channel estimation in large-scale multiple-antenna systems. IEEE J. Sel. Areas Commun. **31**(2), 264–273 (2012)
6. Ma, J., Ping, L.: Data-aided channel estimation in large antenna systems. IEEE Trans. Signal Process. **62**(12), 3111–3124 (2014)
7. Kong, D., Qu, D., Luo, K., et al.: Channel estimation under staggered frame structure for massive MIMO system. IEEE Trans. Wirel. Commun. **15**(2), 1469–1479 (2016)



A Time-Domain Combined Anti-jamming Preprocess Method for DOA Estimation in DSSS System

Shuyun Li[1(✉)], Sheng Ke[1], and Yuan Chai[2]

[1] Beijing Institute of Technology, Beijing 100081, China
2439246076@qq.com, keshengbit@vip.qq.com
[2] Institute of Telecommunication Satellite,
China Academy of Space Technology, Xi'an 710000, China
crazy061235@gmail.com

Abstract. In this paper, we propose an improved time-domain combined anti-jamming structure applied ahead of DOA estimation in DSSS communication system. It makes full use of the weak correlation of DSSS signals on the premise of the accurate use of the DOA estimation algorithm. Compared with the traditional anti-jamming method used in beam space, it influences much less useful signals' peak after anti-jamming and it has better resolution. In addition, this method can suppress the interference that has the same arrival direction of the useful signals.

Keyword: Time-domain anti-jamming · The beam space MUSIC algorithm DSSS · DOA estimation

1 Introduction

Direction of arrival (DOA) estimation has been widely applied in many fields, such as radar, electronic counter measure (ECM), sonar, communication, etc. [1, 2]. However, many kinds of interference are unavoidable for DOA estimation in practice, such as active jamming and interaction of useful signals, etc. [3]. The spreading spectrum system has a strong anti-jamming ability, but sometimes in the presence of strong jammers, it is difficult to suppress the jammers effectively via only the spreading gain [4]. In order to improve anti-jamming performance, a group of time-domain combined ant-jamming filters are used by making full use of the weak correlation of DSSS signals. In this paper, we put forward an improved time-domain anti-jamming algorithm in the beam space. The algorithm has a simpler structure than other methods used in the array space due to it is adopted in the beam space and has less channels. Compared with the transformation domain techniques, the linear phase characteristic of anti-jamming filter in time-domain guarantees the accuracy of DOA estimation, and it can be transformed to an iterative structure easily. Moreover, the improved algorithm fixes the problem that the traditional LMS filter banks will hurt the peak of the desired signals in the spatial spectrum. And more importantly, when the arrival direction of interference and signal are the same, the filter also plays a role in anti-jamming [5, 6].

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 186–194, 2018.
https://doi.org/10.1007/978-3-319-78130-3_20

The paper is organized as follow: the model and the assumption of the whole system are proposed in Sect. 2. Then the algorithm of DOA estimation in beam space is introduced and the constraint that the spatial spectrums of useful signals are not affected after filtering is presented. Then we put forward our improved method of the time-domain combined anti-jamming filter according to the constraint condition. Finally, the MATLAB simulation results of the DOA estimation and the resolution of the combined anti-jamming and the traditional time-domain anti-jamming are shown. And when the interference and the desired signal come from the same direction, by analyzing the correlation function of the received signal and the pseudorandom code, we know this method also has worked.

2 Model and Assumption

Figure 1 shows the schematic of DOA estimation added the time-domain combined filter. The first-order beamforming adopts the conventional multi-beam forming that covers the possible arrival direction of signals. The antenna array signals are composed to several beam spaces, and then the directions of arrival are estimated using the composing beam signals. The estimated signal directions are used for second-order adaptive beamforming. In order to get accurate direction of signal, we add a combined time-domain anti-jamming preprocess in beam spaces. Compared with the anti-jamming in antenna array, the progress saves more resources in beam space for there are always more channels in the array space than in the beam space.

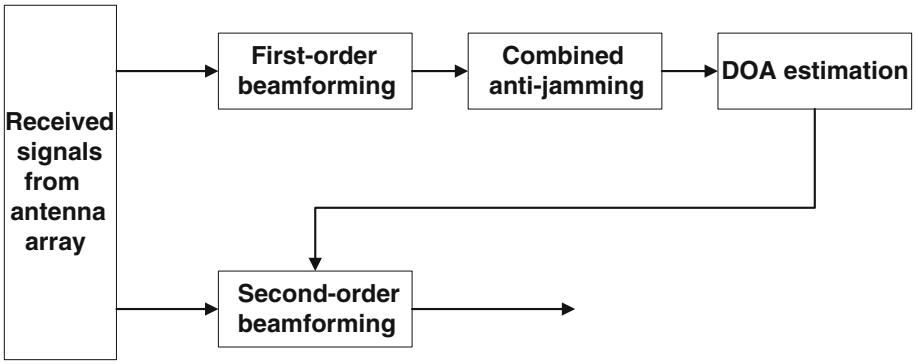


Fig. 1. Schematic of DOA estimation added the time-domain combined filter

The mathematical model of received signal that is narrowband and far-field from antenna array can be written as

$$\boldsymbol{X}(t) = \boldsymbol{A}(\theta_s)\boldsymbol{S}(t) + \boldsymbol{N}(t) + \boldsymbol{A}(\theta_i)\boldsymbol{I}(t) \tag{1}$$

For convenience, we suppose a Uniform Linear Array (ULA) with M linear equispaced omni-directional sensors. $\boldsymbol{S}(\mathrm{t}) = [\boldsymbol{s}_1(\mathrm{t}), \boldsymbol{s}_2(\mathrm{t}), \ldots, \boldsymbol{s}_j(\mathrm{t})]^T$ are the received J

signals from antenna array. $\boldsymbol{A}(\boldsymbol{\theta_s}) = [\boldsymbol{a}(\theta_1), \boldsymbol{a}(\theta_2), \ldots, \boldsymbol{a}(\theta_j)]$ is a M * J matrix, and the steering vector of signal is written as:

$$\boldsymbol{a}\left(\theta_j\right) = [1, e^{-j2\pi dsin\theta_j/\lambda}, \ldots, e^{-j2\pi(M-1)dsin\theta_j/\lambda}]^T \tag{2}$$

Where $[.]^T$ stands for the transpose of a matrix. $\boldsymbol{I}(t) = [\boldsymbol{i}_1(t), \boldsymbol{i}_2(t), \ldots, \boldsymbol{i}_k(t)]^T$ is the received K interferences from antenna array, $\boldsymbol{A}(\boldsymbol{\theta_i}) = [\boldsymbol{a}(\theta_1), \boldsymbol{a}(\theta_2), \ldots, \boldsymbol{a}(\theta_k)]$ is a M * K matrix, and the steering vector of interference is written as:

$$\boldsymbol{a}(\theta_k) = [1, e^{-j2\pi dsin\theta_k/\lambda}, \ldots, e^{-j2\pi(M-1)dsin\theta_k/\lambda}]^T \tag{3}$$

$\boldsymbol{N}(t)$ is a Gaussian noise vector which is generated by array with a zero mean and variance of σ^2.

There are **B** first-order beam formers that form **B** beam spaces cover the possible arrival direction of signals. Only with the dimension of the beams is larger than the number of signals to be estimated, the directions of all incoming signals can be estimated, at the same time, we hope that the computation in beam space is less than in array space, so there is **J < B < M**. The beam transformation matrix is **T.** The **B** beam space signals are written in vector form as:

$$\boldsymbol{Y}(t) = \boldsymbol{T}^H \boldsymbol{X}(t) \tag{4}$$

As for the MUSIC algorithm in beam space, the covariance matrix of array signal $\boldsymbol{X}(t)$ is given by:

$$\begin{aligned} \boldsymbol{R}_{xx} = E\left[\boldsymbol{X}\boldsymbol{X}^H\right] &= \boldsymbol{A}(\theta_s)E\left[\boldsymbol{S}\boldsymbol{S}^H\right]\boldsymbol{A}(\theta_s)^H + E\left[\boldsymbol{N}\boldsymbol{N}^H\right] + \boldsymbol{A}(\theta_i)E\left[\boldsymbol{I}\boldsymbol{I}^H\right]\boldsymbol{A}(\theta_i)^H \\ &= \boldsymbol{A}(\theta_s)\boldsymbol{R}_{ss}\boldsymbol{A}(\theta_s)^H + \sigma^2\boldsymbol{I} + \boldsymbol{A}(\theta_i)\boldsymbol{R}_{ii}\boldsymbol{A}(\theta_i)^H \end{aligned} \tag{5}$$

The first-order beamforming changes M × 1 array signals into B × 1 beam signals. The covariance matrix of beam signal Y(t) is given by:

$$\begin{aligned} \boldsymbol{R}_{yy} = \boldsymbol{T}^H\boldsymbol{R}_{xx}\boldsymbol{T} &= \boldsymbol{T}^H\left(\boldsymbol{A}(\theta_s)\boldsymbol{R}_{ss}\boldsymbol{A}(\theta_s)^H + \sigma^2\boldsymbol{I} + \boldsymbol{A}(\theta_i)\boldsymbol{R}_{ii}\boldsymbol{A}(\theta_i)^H\right)\boldsymbol{T} \\ &= \boldsymbol{B}(\theta_s)\boldsymbol{R}_{ss}\boldsymbol{B}(\theta_s)^H + \sigma^2\boldsymbol{I} + \boldsymbol{B}(\theta_i)\boldsymbol{R}_{ii}\boldsymbol{B}(\theta_i)^H \end{aligned} \tag{6}$$

where $\boldsymbol{B}(\theta_s) = \boldsymbol{T}^H\boldsymbol{A}(\theta_s) = \left[\boldsymbol{T}^H\boldsymbol{a}(\theta_1)\ldots\boldsymbol{T}^H\boldsymbol{a}\left(\theta_j\right)\right] = \left[\boldsymbol{b}(\theta_1)\ldots\boldsymbol{b}\left(\theta_j\right)\right]$, $\boldsymbol{B}(\theta_i) = \boldsymbol{T}^H\boldsymbol{A}(\theta_i) = \left[\boldsymbol{T}^H\boldsymbol{a}(\theta_1)\ldots\boldsymbol{T}^H\boldsymbol{a}(\theta_k)\right] = [\boldsymbol{b}(\theta_1)\ldots\boldsymbol{b}(\theta_k)]$, $\boldsymbol{T}^H\boldsymbol{T} = \boldsymbol{I}$.

When the interference-to-signal ratio (ISR) is low, the power of interference is dispersed in entire bandwidth though DSSS signals can be recovered due to the correlation with the local PN code after the dispreading operation. So by using the anti-jamming capability of DSSS signals, we can suppress interference effectively. In this case, the covariance matrix of beam signal $\boldsymbol{Y}(t)$ can be expressed approximatively as:

$$\begin{aligned} \boldsymbol{R}_{yy} &= \boldsymbol{B}(\theta_s)\boldsymbol{R}_{ss}\boldsymbol{B}(\theta_s)^H + \sigma^2\boldsymbol{I} + \boldsymbol{B}(\theta_i)\boldsymbol{R}_{ii}\boldsymbol{B}(\theta_i)^H \\ &\approx \boldsymbol{B}(\theta_s)\boldsymbol{R}_{ss}\boldsymbol{B}(\theta_s)^H + \sigma^2\boldsymbol{I} \end{aligned} \tag{7}$$

The eigenspace of R_{yy} can be decomposed into signal subspace and noise subspace. The beam space MUSIC spectrum is given as follows:

$$\boldsymbol{P}_{b_music} = \frac{1}{\boldsymbol{b}^H(\theta)\boldsymbol{U}_N\boldsymbol{U}_N^H\boldsymbol{b}(\theta)} = \frac{1}{\boldsymbol{a}^H(\theta)\boldsymbol{T}\boldsymbol{U}_N\boldsymbol{U}_N^H\boldsymbol{T}^H\boldsymbol{a}(\theta)} \tag{8}$$

However, when the interference-to-signal ratio (ISR) is high, the effect of interference cannot be ignored. Not only does the interference influence the beam space MUSIC spectrum, but also affects the dispreading operation. In this case, adopting a suitable anti-jamming process is essential.

3 Time-Domain Combined Anti-jamming Based on Beam Space Music Algorithm

3.1 Constraint Condition of Time-Domain Filters in Beam Space

The beam spaces signals are written as: $\boldsymbol{Y}(t) = \boldsymbol{T}^H\boldsymbol{X}_S(t) + \boldsymbol{T}^H\boldsymbol{N}(t) + \boldsymbol{T}^H\boldsymbol{X}_I(t) = \boldsymbol{Y}_S(t) + \boldsymbol{Y}_N(t) + \boldsymbol{Y}_I(t)$, where $\boldsymbol{X}_S(t) = \boldsymbol{A}(\theta_s)\boldsymbol{s}(t)$, $\boldsymbol{X}_I(t) = \boldsymbol{A}(\theta_i)\boldsymbol{I}(t)$. Considering that the DSSS signal $\boldsymbol{Y}_S(t)$ is submerged in the noise owing to the use of high rate PN sequences to spread spectrum before the dispreading operation, we can use time-domain adaptive filter to suppress interference that has the high correlation compared with DSSS signals and noise. In this paper we adopt the Wiener filter that is the optimal estimator based on the minimum mean square error criterion [7].

There is a fixed amplitude and phase relationship [8] that is used for DOA estimation among the beam space desired signals $\boldsymbol{Y}_S(t)$. However,the traditional Wiener filter [9, 10] bank cannot guarantee the relationship because all the filters in different channels are formed uncooperatively. In order to ensure that the spatial spectrums of useful signals are not affected after filtering, we must keep the amplitude and phase relationship of the desired component in the channel unchanged [11, 12]. Based on this prerequisite, this issue can be described as:

Find $\boldsymbol{h}_a$ to

$$\min\left\{\mathrm{E}\left[\left(d_a(\mathrm{t}) - \boldsymbol{h}_a^T\boldsymbol{y}_{Sa}(\mathrm{t})\right)^2\right]\right\} \tag{9}$$

Subject to

$$\frac{\boldsymbol{h}_a^H\boldsymbol{y}_{Sa}(\mathrm{t})}{\boldsymbol{h}_b^H\boldsymbol{y}_{Sb}(\mathrm{t})} = \frac{\boldsymbol{y}_{Sa}(\mathrm{t})}{\boldsymbol{y}_{Sb}(\mathrm{t})} \quad \forall t = 0, 1, 2, \ldots, \forall 1 \le a, b \le B \tag{10}$$

Where the $\boldsymbol{y}_{Sa}(t), \boldsymbol{y}_{Sb}(t)$ are the desired signal components in the a-th and b-th beam space channels and the $\boldsymbol{h}_a, \boldsymbol{h}_b$ are the filter for a-th and b-th beams. In the expression (10), we focus on the constraint. As it must hold for any a, b and t, we have to get

$$\boldsymbol{h}_1 = \boldsymbol{h}_2 = \ldots = \boldsymbol{h}_B = \boldsymbol{h} \tag{11}$$

Unfortunately there is not such a filter **h** that makes all the channels satisfying the minimum condition (9). So we offer the issue a concession that we want to find a filter **h** to minimize function J.

$$J = \sum_{a=1}^{B} \lambda_a E\left[d_a(\mathrm{t}) - \boldsymbol{h}_a^T \boldsymbol{y}_{Sa}(\mathrm{t})\right]^2 \tag{12}$$

where λ_a is the weighting coefficient for each beam. By calculating the partial derivative of J to **h**, we get such a filter h to minimize function J. This expression of h is written as:

$$\boldsymbol{h} = \sum_{a=1}^{B} (\lambda_a \boldsymbol{P}_a) / \sum_{a=1}^{B} (\lambda_a \boldsymbol{R}_a) \tag{13}$$

where $\boldsymbol{P}_a = E[\boldsymbol{y}_{Sa} d_a^H]$, $\boldsymbol{R}_a = E[\boldsymbol{y}_{Sa} \boldsymbol{y}_{Sa}^H]$.

3.2 Design of the Time-Domain Filter for Multi-beam

From the expression (13), the weighting coefficient λ_a has a great influence on the performance of the filter **h**. So how to select the weighting coefficient λ_a is the key question of filter design. A natural thought is that the channel containing larger interference component contributes more to forming the coefficients of the filter to guarantee the effect of anti-jamming. In this way, we firstly smooth the signal with different filters that are designed by the respective data. Then by computing the power variation of before filtering and after filtering, we get the interference power in each beam space and the $\boldsymbol{P}_a, \boldsymbol{R}_a$ of respective filter. In order to make the filter suppress the interference more thoroughly, the larger the power of interference is in certain beam, the bigger the proportion of filter design is. The structures of designed filter is showed

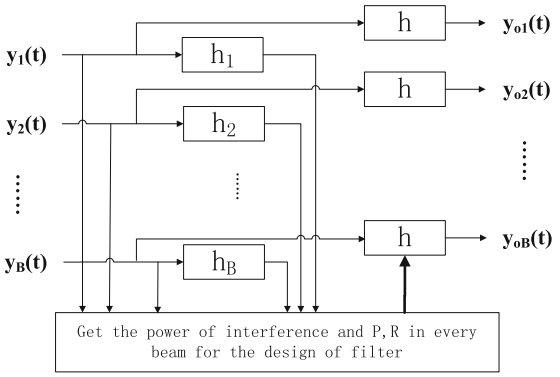


Fig. 2. Structure of the design of time-domain combined filter

in Fig. 2, and the selected method of λ_a is written as expression (14), where $|y_{Ia}|^2$ is the power of interference in a-th beam space. Compared with many other weighting coefficients, this method outperforms.

$$\lambda_a = \frac{|\boldsymbol{y}_{Ia}|^2}{\sum_{a=1}^{B} |\boldsymbol{y}_{Ia}|^2} \tag{14}$$

4 Simulation Results

We adopt the MATLAB as simulation software. Consider a uniform linear array consisting of 16 omni-directional sensors with interspacing d = λ/2. The number of snapshots is 10000. The signals adopt PN sequences to spread spectrum, spreading factor is 255. The beam formers adopt the method of DFT beamforming [13] and generate 8 beam spaces, whose point angle is $[-60^{\circ}, 60^{\circ}]$. The beam pattern is showed in Fig. 3. In this system, we suppose the number of received signal is four at most. Then the number of actual received signal and the direction of arrival are determined by analyzing the height of the spectral peak.

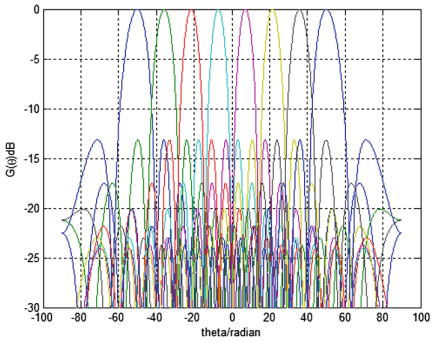
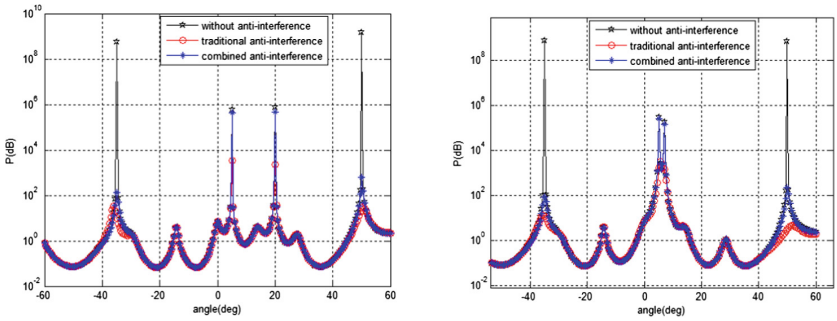


Fig. 3. Directional pattern of 16 arrays and 8 beams

Experiment 1: The Results of DOA Estimation
In the first experiment, there are two narrowband interferences from directions of −35° and 50°, and SNR = −17 dB, ISR = 50 dB. The DOA estimation results are as follows as Fig. 4. In the Fig. 4(a) the directions of signal are 5° and 20°, though in the Fig. 4(b) the directions of signal are 5° and 7°.

As for the simulation Fig. 4(a), we can find that the interferences at −35° and 50° are mistaken as signals without anti-jamming, but with anti-jamming the signals are detected accurately only at 5° and 10°. Figure 4 shows that the time-domain combined anti-jamming method does not influence the signal peak at the same time of suppressing the interference, but the time-domain traditional anti-jamming method cut



(a) signal angle=[5° 20°] (b) signal angle=[5° 7°]

Fig. 4. DOA estimation for SNR = −17 dB, ISR = 50 dB

down the signal peak through it restrains the interference peak more. From the Fig. 4(b), we can also know that when two signals are closed to each other, the traditional method cannot distinguish two signals and only mistakes one signal whose estimated direction of arrival deviates from the any actual signal's direction.

Experiment 2: The Comparison of Resolution and the Influence on Second-Order Beam

In the second experiment, there are two narrowband interferences from directions of −35° and 50°, the one DSSS signal is from direction of 5°, the other signal's direction is from 5° to 10°, and SNR = −17 dB, ISR = 50 dB. The resolution of combined anti-jamming and traditional anti-jamming is showed in Fig. 5. The error rate of the second-beam due to the deviation of DOA estimation is showed in Fig. 6.

From Fig. 5, we can get that the traditional method only can successfully distinguish two signals whose directions of arrival are 4° apart though the combined method can distinguish two signals whose directions of arrival are 1° apart. So the combined anti-jamming method has the better resolution than the traditional method. The simulation in Fig. 6 shows that the results of DOA estimation in first-order beam have a great influence on the second-order beam. When deviation is more than 1°, the communication process is unable to achieve in second-beam. So the high resolution that

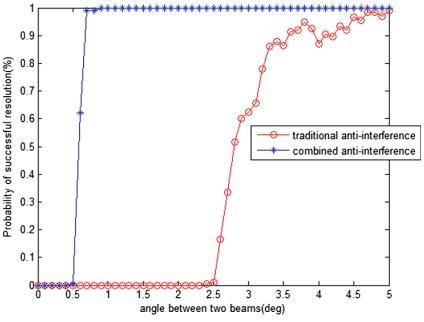


Fig. 5. The comparison of resolution

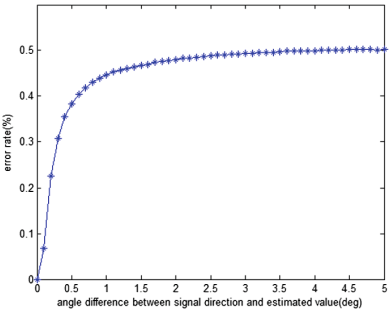
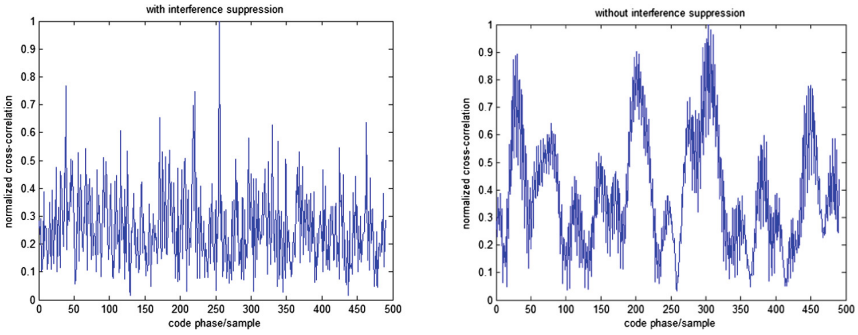


Fig. 6. Error rate of second-beam

provides the accurate number of signals and the direction of arrival is necessary in the whole system.

Experiment 3: The Comparison of Correlation Function
In the third experiment, there are two narrowband interferences from directions of 20° and 50°, the two DSSS signals are from direction of 5° and 20°, and SNR = 0 dB, ISR = 30 dB. The correlation functions of after combined anti-jamming and without interference suppression are showed in Fig. 7.



(a) combined anti-jamming (b) without interference suppression

Fig. 7. Correlation function

In this experiment, there are two interferences of which the one has the same direction as DSSS signals, the other has different direction with DSSS signals. The Fig. 7(b) illustrates that the operation of anti-jamming is so necessary in high ISR because that the correlation peak is overwhelmed by strong interference. By comparing the two figures with anti-jamming and without anti-jamming in the Fig. 7, we can know that the correlation of signal after filtering is improved greatly. At the same time, the results in the Fig. 7(a) indicates that the time-domain combined anti-jamming method not only suppress the interference that has different direction of signals, but also suppress the interference that has the same direction of signals.

5 Conclusion

According to the model of DOA estimation and beamforming, this paper proposes the time-domain anti-jamming in beam space that saves more resources than in antenna array. We get the constraint condition in order to ensure that the spatial spectrums of useful signals are not affected after filtering by analyzing the algorithm of DOA estimation. And we put forward a designed method of the time-domain combined anti-jamming filter according to the constraint condition. In the simulation results, by compared with the traditional anti-jamming method, the combined method does not influence useful signals' peaks and has better resolution that can successfully

distinguish two signals whose directions of arrival are 1° apart. And more importantly, when the arrival directions of interference and signal are the same, the filter also plays a role in anti-jamming.

References

1. Kai, Y.: The research of beam-space DOA estimation based on non-uniform array. Southeast University (2015)
2. Kim, M., Ichige, K., Arai, H.: Design of Jacobi EVD processor based on CORDIC for DOA estimation with MUSIC algorithm. In: The IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, vol. 1, pp. 120–124. IEEE (2002)
3. Wang, G., Pan, Y., Peng, S., et al.: A new anti-jamming preprocess method for DOA estimation based on uniform circular array. In: CIE International Conference on Radar, pp. 1–4. IEEE (2006)
4. Qi, P., Lv, H., Zhang, A., et al.: Design and simulation of multi-user direct sequence spread spectrum system based on m sub-sequence. In: International Conference on Optical Communications and Networks, pp. 1–4 (2013)
5. Zhou-Hua Z.: The fourth order cumulants based modified MUSIC algorithm for DOA in colored noise. In: Wearable Computing Systems, pp. 345–347. IEEE (2010)
6. Devendra, M., Manjunathachari, K.: DOA estimation of a system using MUSIC method. In: International Conference on Signal Processing and Communication Engineering Systems, pp. 309–313 (2015)
7. Masry, E.: Closed-form analytical results for the rejection of narrow-band interference in PN spread-spectrum systems-Part I: linear prediction filters. IEEE Trans. Commun. **32**(8), 888–896 (1984)
8. Zhang, Z., Zhang, H., Zhu, G.: Effect and analysis of amplitude and phase errors in modified MUSIC algorithm direction-finding. Mod. Electron. Tech. **30**(5), 48–50 (2007)
9. Gong, W., Sun, X., Wu, S.: Research on adaptive reduced-dimension STAP anti-interference based on multistage nested Wiener filter. J. Astronaut. **31**(10), 2360–2366 (2010)
10. Xiong, Z.: Research on key technologies of the anti-jamming GNSS receiver. Beijing Institute of Technology (2016)
11. Wang, D., Lu, Y., Pei, W., Pang, C.: Application of space algorithm based on MUSIC in GPS anti-jamming. Mod. Defence Technol. **41**(6) (2013)
12. Huang, J.: Research on spatial anti-jamming algorithms of GNSS and FPGA implementation. Chongqing University (2016)
13. Wei, X., Wu, C., Yao, C.: A novel blind multi-target beam-forming algorithm in guidance data link system. In: Chinese Control and Decision Conference, pp. 2966–2969 (2014)

An Efficient Wireless Backhaul Algorithm for User-Centric Ultra-Dense Networks

Hong Feng(✉), Xi Li, and Hong Ji

Key Laboratory of Universal Wireless Communication, Ministry of Education,
Beijing University of Posts and Telecommunications,
100876 Beijing, People's Republic of China
{fenghong, lixi, jihong}@bupt.edu.cn

Abstract. Currently, user-centric ultra-dense network (UUDN) is recognized as a promising candidate to deal with the challenge of very high area throughput density. However, the limited wireless backhaul capability restricts the performance of UUDN and needs further study. In this paper, we propose a score and location-aware distributed (SLD) wireless backhaul routing algorithm to transmit users' data flexibly and efficiently. We consider channel state information, interference between access points (APs) and load of APs in the algorithm. The joint optimization of load of APs and throughput in backhaul links is formulated as a weighted rate maximization problem and resolved by iterative algorithm. The proposed wireless backhaul routing algorithm balances backhaul tasks on APs in UUDN and has high success rate in finding backhaul paths. Simulation results show the effectiveness of the proposed algorithm.

Keywords: UUDN · Wireless backhaul · Routing · Access point group

1 Introduction

Nowadays, mobile internet is changing people's daily life. The increasing numbers of high traffic data rate services lead to the explosive mobile traffic growth in recent years. Moreover, it is predicted that mobile data demand will increase up to 1000-fold by 2020 [1]. In order to meet the explosive data traffic demand, user-centric ultra-dense network (UUDN) is emerging as a promising candidate [2]. Compared with the traditional heterogeneous network, UUDN has a new network architecture which break through the traditional cellular architecture of the network controlling users [3]. In UUDN, access points (APs) are densely deployed in busy areas, such as shopping malls, subways, railway stations and so on. Every terminal has its own unique access point group (APG) in which there are some APs transmitting the data to user efficiently. With the increasing traffic demands of users, the throughput requirement of UUDN in downlinks is also growing enormously. However, backhaul is acknowledged as a bottleneck in UUDN because of the lack of ideal backhaul links. It is difficult and inflexible to connect all the APs to local service center (LSC) with fibers to deal with the backhaul problem in UUDN [4]. Therefore, it's feasible to forward the backhaul traffic of APs into core network in a flexible wireless way.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 195–205, 2018.
https://doi.org/10.1007/978-3-319-78130-3_21

There are many studies in wireless backhaul for different networks. Compared with wired backhaul, wireless backhaul has advantages of cost and flexibility. Small base stations establish wireless backhaul links directly in heterogeneous small cell network. Many wireless backhaul studies are limited in one hop wireless transmissions and the node, which is responsible to forward data back to the core network, is set as backhaul node specially. Meanwhile, multi-gateway wireless backhaul is researched based on wireless multi-hop relay. The macro base station controls all wireless backhaul routing in the research. The limited throughput of wireless backhaul links restrict the increment of users' data rate. Because of dense distributed deployment of APs, short distances between APs and strong interferences, the existing research results are not appropriate in UUDN.

As to the wireless backhaul design, researchers have proposed many approaches. The backhaul topology considering deployment cost and reliability in Wi-Fi network has been studied [5]. In microwave wireless backhaul network, the design of backhaul takes some factors into account, such as hop count on backhaul link, the total length of backhaul link, cross count between adjacent links and so on [6]. And a LOS SM-MIMO system concept is proposed in [7] which has a good performance in system capacity and energy efficiency. In the aspect of increasing backhaul throughput, a joint optimization multicast beamforming in backhaul links and access links in cloud-RANs is studied in [8]. In [9], the authors analyze the energy efficiency of a two-tier MIMO heterogeneous network with wireless backhaul and prove the importance to control the network load. Meanwhile, the backhaul energy consumption of small cells under different backhaul architectures have been analyzed in [10].

However, in the aforementioned wireless backhaul network studies, the backhaul topologies are fixed and could not change along with the load of APs and channel conditions. Moreover, almost all of the researches on wireless backhaul have dedicated nodes to control the wireless backhaul routing in a centralized way. In this paper, we propose a score and location-aware distributed (SLD) wireless backhaul routing algorithm in UUDN. The joint optimization of load of APs and throughput in backhaul links is formulated as a weighted rate maximization problem in every hop on the backhaul path. We consider channel state information (CSI), interference between APs and load of APs in the algorithm. The proposed algorithm balances backhaul tasks on APs and has high success rate in finding backhaul paths when satisfying the traffic demands of users. Simulation results prove the performance of the proposed algorithm compared to the existing schemes.

The rest of the paper is organized as follows. Section 2 introduces the system model and Sect. 3 investigates the problem. The SLD wireless backhaul routing algorithm is proposed in Sect. 4. Simulation results are presented in Sect. 5. Finally, Sect. 6 provides the conclusions.

2 System Model

As shown in Fig. 1, the network is composed by a LSC and *S* APs with some mobile users in the areas. The LSC connects to core network with fiber links which have a feature of wide bandwidth and can provide services with high data rate. APs build access links with users which are in these APs' coverage areas. Several APs make up an access points group (APG) and serve a user together so that the user can reach high data access rate and get prefect quality of experience. Every user has its own unique APG in UUDN. In an APG, some APs communication with the user on downlink but others for uplink. The user gets information from APs on the downlink. While, the user also needs to transmit huge traffic data to core network when enjoying high data rate service. So APs are also responsible for gathering data generated by the user and sending the data to LSC which works as gateway. In this paper, we consider the problem about how to transmit the data generated from users to core network in a cost-effective and efficient way.

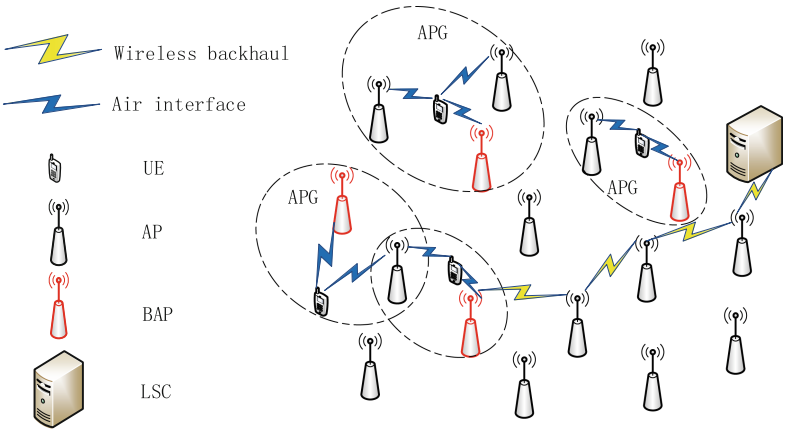


Fig. 1. The wireless backhaul network in UUDN

According to the scenario in Fig. 1, a unique APG is created when a user is attached to the network. We define users' traffic data that need to be sent to core network as backhaul task of APs and define the AP which has backhaul tasks as backhaul AP (BAP). The AP which has no access link with users is defined as unoccupied APs. In this paper, we assume APs providing downlink services to users don't undertake backhaul tasks and only BAPs and unoccupied APs are responsible for backhaul tasks. So wireless backhaul communications take place between BAPs or

unoccupied APs. Ultimately, BAPs pass data to LSC in a "hop-by-hop" way, and then LSC forwards the data to core network through fiber links at a high data rate.

3 Problem Model and Solution

We assume there is only a BAP in an APG, and then focus on wireless backhaul links between BAPs and simplify access links between users and APs in UUDN. The number of backhaul tasks of BAPs changes with time and load conditions. When sending traffic data to next BAP or unoccupied AP, a BAP occupies a sub-channel. Moreover, if having more than one task, a BAP transmits data on some different sub-channels at the same time. In order to construct flexible wireless backhaul network and make full use of network resources, we put forward a score and location-aware distributed (SLD) wireless backhaul routing algorithm to plan wireless backhaul paths.

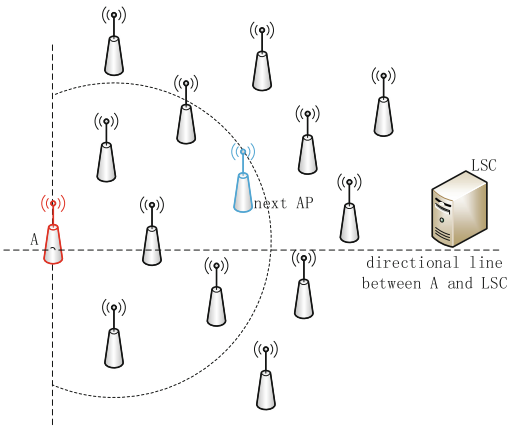


Fig. 2. The selection of next AP in a hop

As shown in Fig. 2, when a user sends its traffic data to the BAP, which is denoted by red color and as *A*, *A* broadcasts a backhaul request signal to APs around it. After receiving the backhaul request signal, BAPs and unoccupied APs calculate their own scores by themselves and send response signals to *A*. The response signal contains the information of score and location. Through analyzing response signals it receives, *A* finds its candidate APs set and selects the AP with the highest score as the next AP of *A*. It is denoted by blue color in Fig. 2. And then, the next AP of *A* performs the same process like *A* to find its next AP. Ultimately, traffic data is forwarded to LSC and the whole backhaul path is planned.

Here we explain how to find the next AP of A as follows. A is the starting BAP in Fig. 2. We assume wireless backhaul communications between APs share a same spectrum resource. The limited system bandwidth is divide into a fixed number of sub-channels. We assume that the set F represents all sub-channels in the network. The size of F is N. And F is formulated as $F = \{f_1, f_2, \ldots, f_N\}$. We assume that the k-th AP is one of APs in the semicircle region of Fig. 2. We assume the number of k-th AP's backhaul tasks is b_k. b_k could be zero here. The set J represents sub-channels occupied by backhaul tasks of k-th AP. As different backhaul tasks occupy different sub-channels to transmit data, the size of J is b_k. And J is formulated as $J = \{j_1, j_2, \ldots, j_{b_k}\}$. The k-th AP may has its own users and we assume that the set L represent sub-channels occupied by access links of k-th AP. We assume that the number of users of k-th AP is a_k. And L is formulated as $L = \{l_1, l_2, \ldots, l_{a_k}\}$.

The score takes two factors into account. One is the maximum rate that k-th AP can achieve. Another is the band resources which is occupied by k-th AP at present. For the first factor we consider, k-th AP gets interference information on each sub-channels. The total interference power received by k-th AP for all the other APs in the network at sub-channel n is

$$I_k^n = \sum_{i \neq A} P_i^n h_{ik}^n d_{ik}^{-\alpha}, \quad n \in F \text{ and } n \notin J \text{ and } n \notin L \tag{1}$$

where P_i^n is transmit power of the i AP at sub-channel n and i represents the interfering AP that utilizes the sub-channel n.h_{ik}^n is the path gain between k-th AP and i AP and it follows $h_{ik}^n \sim \exp(1)$.d_{ik} is the distance between k-th AP and i AP. $\alpha > 2$ is the path loss coefficient.

According to the formula (1), the received SIN of k-th AP at sub-channel n is

$$SINR_{Ak}^n = \frac{P_A^n h_{Ak}^n d_{Ak}^{-\alpha}}{I_k^n + \sigma_k^n} \tag{2}$$

Here, we assume A communicates with k-th AP through the maximum power P_A^n.σ_k^n is the power spectral density of the Additive Gaussian White Noise and $\sigma_k^n = N_0 B$. B is the band width of a sub-channel.

Therefore, according to Shannon's law, the data rate of k-th AP over sub-channel n is denoted as

$$r_{Ak}^n = B \log_2\left(1 + SINR_{Ak}^n\right) \tag{3}$$

For the second factor, band resources occupied by k-th AP is selected. We let x_k^n be a binary variable where $x_k^n = 1$ if sub-channel n is occupied by k-th AP and $x_k^n = 0$ otherwise. The score of k-th AP calculated by itself is denoted as

$$score_k = \max\ x_k^n r_{Ak}^n \left(1 - \frac{a_k + b_k}{N}\right) \tag{4}$$

$$\text{s.t.} \qquad \sum_n x_k^n = 1 \tag{4a}$$

$$x_k^n = \{0, 1\}, \forall n \tag{4b}$$

$$P_i \leq P_m, \forall i \tag{4c}$$

Our object is to find the sub-channel that could achieve the maximum data rate of the k-th AP in all system sub-channels. After calculating the score, the k-th AP send its response signal, including score, location and data rate between k-th AP and A on selected sub-channel, to A.

When receiving response signals from surrounding APs, A extracts the score and location information from these signals. A selects some APs from APs that send response signals as candidate APs on the basis of location and data rate information. We assume the set C represents candidate APs of A and the size of C is m. And C is formulated as $C = \{c_1, c_2, \ldots, c_m\}$. The decision of next AP is

$$\max\ score_{c_i},\ c_i \in C \tag{5}$$

As the formula (5) shows, A selects the AP whose score is the maximum in the candidate set. A section of path is successfully planned so far.

4 SLD Wireless Backhaul Routing Algorithm

For the purpose of finding a backhaul path from initial BAP to LSC, we propose a SLD wireless backhaul routing algorithm (SLDWBRA). The general sequence of SLDWBRA is as follows.

The initial BAP broadcasts a backhaul request signal to surrounding APs, and then enter a wait state. After receiving backhaul request signal, APs calculate their own scores and send response signals to the initial BAP. The AP can be a candidate AP of initial BAP, only if its distance between LSC is less than the distance between initial BAP and LSC. The initial BAP chooses some candidate APs and selects the AP, with maximum score in candidate APs, as next AP. With this, the initial BAP forwards traffic data to the next AP. Then, the next AP conducts the same process as initial BAP does. As this way, traffic data is forwarded to LSC and the whole backhaul is planned. Explained specially, to ensure traffic data could be forwarded to LSC, we set LSC is the next AP of current BAP directly once LSC becomes one of candidate APs.

The detailed algorithm is presented in Algorithm 1.We use an array task to denote sub-channels occupied by backhaul tasks.

Algorithm 1 SLD Wireless Backhaul Routing Algorithm

Input:
contention graph **G**
existing tasks array on BAP **task**
access point ***A***
Local Service Center **LSC**
Output:
the path from ***A*** to **LSC** and its score
Initialization: Initialize **G**, **task**, ***A***, **LSC**
A broadcast request signal and APs around ***A*** form set ***S***
Definition:
T: the set of sub-channels occupied by *Si*;
N: the set of all sub-channels in the network;
for all *Si* ∈ ***S*** **do**
  for all *Ni* ∈ ***N*** **do**
    if *Ni* ∉ ***T***
    calculate the data rate **r** and the score of *Si* on sub-channel *Ni*, and then add the score in set ***E***
    end if
  end for
  score of *Si* = the maximum score in ***E***
  add the score and location information in response signal of *Si*
end for
APs send response signals（set ***B***）to ***A***
for all *Bi* ∈ ***B*** **do**
  if the location information in *Bi* is wrong abandon *Bi*
  else if the data rate is not satisfied abandon *Bi*
  else
  save score of *Bi* in ***A***'s score statistics, these APs constitute candidate set ***C***
  end if
end for
if ***C*** = ∅
  fail to find the path
else if LSC ∈ ***C*** **do**
  A connect to **LSC** directly and send data to **LSC**
  find the path successfully and return score of the selected path
else
  select the AP (***K***) with the maximum score in set ***C*** as the next AP of ***A***
  A connect to ***K*** and send traffic data to ***K***
  K go to find next AP like ***A*** until the path from ***A*** to **LSC** is found
end if

5 Simulation Results

In this section, we evaluate the performance of the SLD wireless backhaul routing algorithm and analyze the influence of important parameters by simulation. We assume flat Rayleigh fading channels and the path loss coefficient $\alpha = 4$. We deploy 150 APs in

a 100 m × 100 m square space. The location of these APs is randomly distributed evenly. The LSC is positioned at the center of the square space. Unless being mentioned elsewhere, we use this scenario as the standard model for the numerical simulation, where the maximum transmit power of AP is $P_m = 0.3$ W. The white noise density is set −174 dBm/Hz and total bandwidth is 20 MHz.

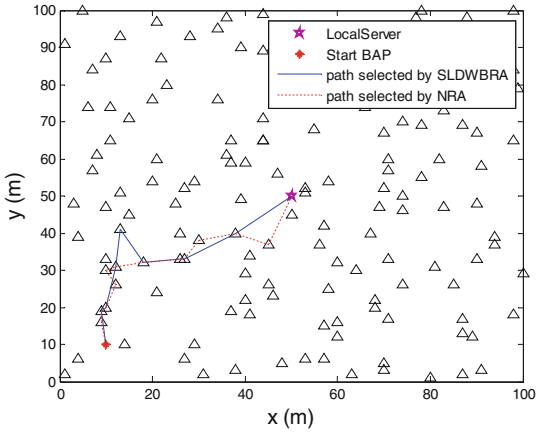


Fig. 3. The backhaul path planned by SLDWBRA and NRA

The backhaul paths planned by SLDWBRA and nearest routing algorithm (NRA) are displayed in Fig. 3. The Fig. 3 also shows the deployment scenarios of APs in simulation. Comparing with the two paths, we can find some sections of two backhaul paths are same and the feature of NRA is selecting the nearest AP in every hop.

Figures 4 and 5 show the loads and throughputs of candidate APs in a hop. The selected AP is the second AP in Fig. 4 and the fourth AP in Fig. 5. The throughputs of APs in these two figures is calculated with maximum power of AP. In Fig. 4, it is clear the throughput of the second AP is the highest and could meet data rate demand of the users. Meanwhile, the second AP has no backhaul tasks. It illustrates the load of the second AP is the lightest in these five candidate APs. Considering these two factors, it is wise to choose the second AP as the next AP. In Fig. 5, the third AP has no backhaul tasks but its throughput is so low that the data rate requirement of the users can't be satisfied. While, the throughput of the sixth AP is maximum in candidate APs, but the sixth AP has four backhaul tasks. However, the fourth AP has one backhaul task and its throughput also satisfies the data rate requirement. Balancing the two factors, it's a good choice to select the fourth AP as the next AP.

Figure 6 illustrates the relationship between throughput of backhaul path and the user's data rate requirement. Although data rate requirement of the user changes with different services, the throughput of backhaul path planned by SLDWBRA meets the user's demand well. This shows that SLDWBRA is reliable and could satisfy various internet services.

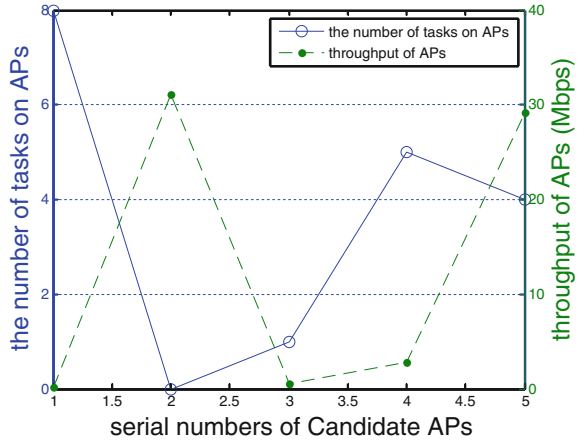


Fig. 4. The loads and throughputs of candidate APs in a hop

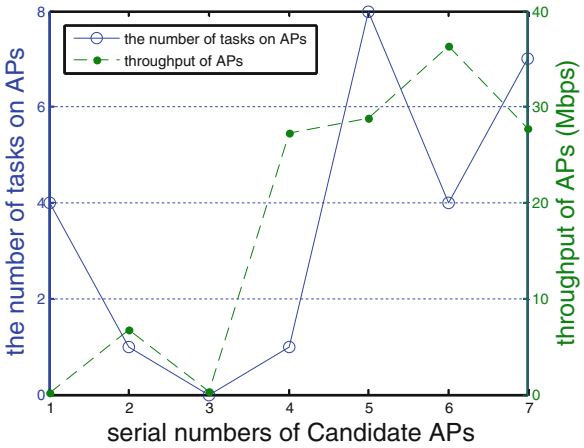


Fig. 5. The loads and throughputs of candidate APs in a hop

Figure 7 illustrates the relationship between data rate requirement of users and success rate of selecting next AP in a hop. We compare the performance of success rate of selecting next AP between SLDWBRA and NRA. From Fig. 7, we can get the success rate of selecting next AP shows a trend of decline with the increasing number of users' data rate demand. Compared with NRA, the success rate of SLDWBRA is always higher than NRA and the gap between them is obvious when users' data rate demand is large. It shows that the SLD wireless backhaul routing algorithm proposed in this paper can get better effect in the success rate of selecting next APs.

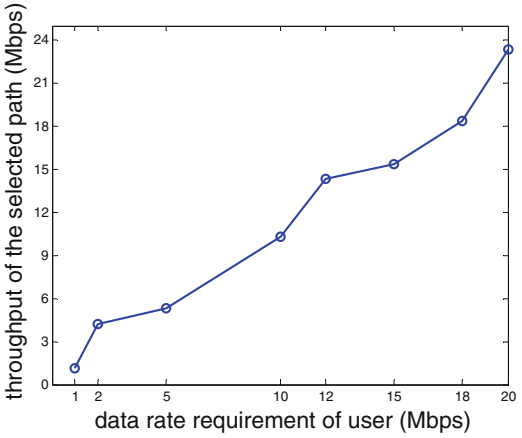


Fig. 6. Relationship between throughput of backhaul path and the user's data rate requirement

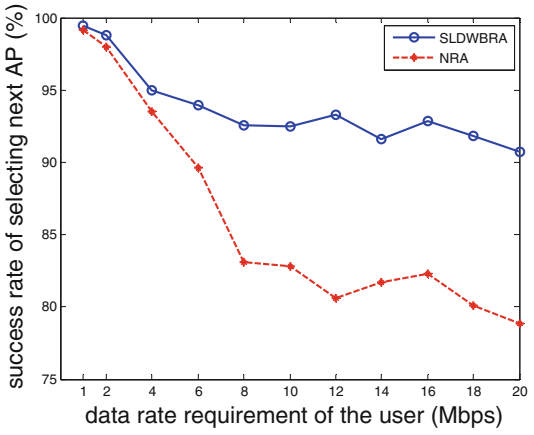


Fig. 7. Relationship between success rate of selecting next AP and user's data rate requirement

6 Conclusion

In this paper, we propose SLD wireless backhaul routing algorithm in the user-centric ultra-dense network. In the algorithm, we consider channel conditions, interference and band resources occupied by AP when selecting the next AP. At the same time, channel selection is also completed in the algorithm. The SLD wireless backhaul routing algorithm aims at planning a whole backhaul path between initial BAP and LSC, and throughput of the backhaul path could satisfy the traffic data demand of users. The simulation results show that the scoring mechanism of APs provides a tradeoff of throughput and load of candidate APs. We have shown that the proposed algorithm performs considerable gains in success rate of selecting next AP compared with the nearest routing algorithm and thus, could become a suitable candidate backhaul algorithm for future ultra-dense networks.

Acknowledgments. This paper is jointly sponsored by the National Science and Technology Major Project of the Ministry of Science and Technology of China (Grant No. 2016ZX03001017) and the National Natural Science Foundation of China (Grant No. 61671088).

References

1. Ran, C., Wang, S., Wang, C.: Balancing backhaul load in heterogeneous cloud radio access networks. IEEE Wirel. Commun. **22**, 42–48 (2015)
2. Robert, B., Tim, I., Kumar, B., Erik, D., Gunnar, M., Yngve, S., Stefan, P., Michael, M., Afif, O.: Ultra-dense networks in millimeter-wave frequencies. IEEE Commun. Mag. **53**, 202–208 (2015)
3. Chen, S., Qin, F., Hu, B., Li, X., Chen, Z.: User-centric ultra-dense networks for 5G: challenges, methodologies, and directions. IEEE Wirel. Commun. **23**, 78–85 (2016)
4. Ge, X., Pan, L., Tu, S.: Wireless backhaul capacity of 5G ultra-dense cellular networks. In: IEEE 84th Vehicular Technology Conference (VTC-Fall), pp. 18–22, September 2016
5. Ma, Y., Zhang, J.: Optimisation of Wi-Fi mesh network for small cell backhaul. In: International Workshop on High Mobility Wireless Communications (HMWC) (2014)
6. Li, Y., Cai, A., Qiao, G.: Multi-objective topology planning for microwave-based wireless backhaul networks. IEEE Access **4**, 5742–5754 (2016)
7. Darko, C., Tim, H., Berthold, L.: Next generation mm-Wave wireless backhaul based on LOS MIMO links. In: German Microwave Conference (GeMic), German (2016)
8. Hu, B., Hua, C., Chen, C., Guan, X.: Multicast beamforming for wireless backhaul with user-centric clustering in Cloud-RANs. In: IEEE International Conference on Communications (ICC), May 2016
9. Yang, H.H., Giovanni, G., Quek, T.Q.S.: MIMO HetNets with wireless backhaul: An energy-efficient design. In: IEEE International Conference on Communication (ICC), May 2016
10. Sibel, T., Paolo, M., Fabricio, F.: Is backhaul becoming a bottleneck for green wireless access networks? In: IEEE International Conference on Communication (ICC), August 2014



Stochastic Optimal Control for Power Allocation and Link Selection in Energy-Harvesting Three-Node Relay Network

Qianyun Gong[1], Qinbo Chen[1], Hui Cai[1], Maoxin Tian[1], and Sai Zhao[2(✉)]

[1] School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, Guangdong, China
{gongqy3,chenqb6,caih8,tianmx}@mail2.sysu.edu.cn
[2] School of Mechanical and Electrical Engineering, Guangzhou University, Guangzhou 510006, Guangdong, China
zhaosai@gzhu.edu.cn

Abstract. In this paper, we consider a three-node relay network with energy harvesting (EH). Assuming the EH source equipped the energy queue, and the EH relay has energy and data queues, unlike the conventional half duplex relay, the relay experience link selection at each slot. We design a optimization approach aiming at maximizing the long-term averaged transmission rate under the energy queue and data queue length constraints. Considering the link selection and power allocation, we propose a Lyapunov optimization algorithm to solve the energy causality, then obtain the asymptotically optimality. Without requiring any knowledge of the statistics of energy arrivals and channel states, simulations show our proposed algorithm outperform the greedy algorithm and conventional relay algorithm.

Keywords: Energy harvesting · Power allocation · Link selection
Lyapunov optimization

1 Introduction

The original three-node relay network was put by van der Meulen [1], with a source, a decode-and-forward (DF) relay and a destination. For conventional half-duplex relay network, the data transmit from source to relay at first slot, and then transmit to destination at second slot. While a relay protocol employing adaptive link selection have been investigated on the three-node relay network by [2]. To solve the increasing energy cost, the EH techniques, which allow buffer operated on wireless communication systems to harvest energy from surrounding environment, is promising on energy-constrained wireless networks. In [3], the authors work out the rate maximization problem on the EH buffer-aided link adaptive relaying systems by dynamic programming (DP) approach.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 206–215, 2018.
https://doi.org/10.1007/978-3-319-78130-3_22

However, the numerical solutions by DP are obtained with high computational complexity, therefore, a low-complexity Lyapunov approach is proposed in [4] firstly with energy harvesting problem. Based on Lyapunov functions, the energy causality constraint was solved and the online algorithm was proposed in hybrid energy supply networks [5–7]. To the best of our knowledge, considering energy queue and data queue at the same time, the Lyapunov optimization approach for power allocation and link selection on three-node relay network has not been considered.

In this paper, we consider a three-node relay network consisting of a EH source, a half-duplex EH DF relay and a destination. Our goal is to maximize the long-term average transmission rate under the energy queue causality and data queue stability. We apply the low-complexity Lyapunov optimization approach to solve our problem, verify its asymptotically optimality, and compare the performance with greedy algorithm and conventional relay algorithm.

2 System Model and Problem Formulation

In this section, we consider a three-node relay channel consisting of a EH source, a EH relay and a destination node as shown in Fig. 1. There are energy and data queues at the EH relay, while only energy queue at EH source. The relay operates in a half-duplex mode, i.e. it couldn't receive and sent data packet within the same slot. Unlike conventional relaying, the system will choose one link work at each slot. Assume the two link's channel state is perfect knowledgable. Denote $l_k^t \in \{0, 1\}$, where 0 means the link is off, and 1 means the link is on, with $k \in \{s, r\}$, where link s means source-relay link, and link r means relay-destination. We denote Δt the time slot length and T = {0,1,2...} as the set of time slot. Therefore, the link selection are subjected to the following constraint

$$l_s^t + l_r^t = 1, \ k \in \{s, r\}, \forall t \in T, \tag{1}$$

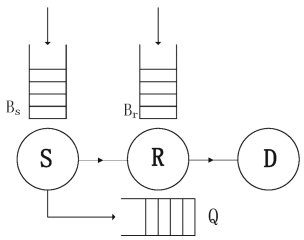


Fig. 1. S-R-D energy harvesting network

At each slot, the energy packets with E_s^{max} and E_r^{max} arrive at the source and relay respectively with independent and identically distribution, and part of the energy, denoted as e_j^t, will be stored in the energy queue that can be transmit in the later slots, satisfying

$$0 \leqslant e_k^t \leqslant E_k^{max}, k \in \{s, r\}, \forall t \in T, \tag{2}$$

Assume the source and relay energy queue length is infinite, denoted as $B_s^t < +\infty$ and $B_r^t < +\infty$. Considering the transmission power should just conclude the energy that has been harvested, the energy causality constraint should be

$$P_k^t \Delta t \leqslant B_k^t \leqslant +\infty, k \in \{s, r\}, \forall t \in T, \tag{3}$$

Thus, the energy queue dynamics is

$$B_k^{t+1} = B_k^t - l_k^t P_k^t \Delta t + e_k^t, k \in \{s, r\}, \forall t \in T, \tag{4}$$

Due to the link selection, the relay node need to have a data queue for keeping data packets until transmission. We assume the data queue length is infinite. Let Q^t denote the data queue length in slot t, the data queue dynamics is

$$Q^{t+1} = Q^t - l_r^t R_r^t B \Delta t + l_s^t R_s^t B \Delta t, \tag{5}$$

where B means bandwidth, $R_k^t = log_2\ (1 + \frac{P_k^t |h_k^t|}{\sigma^2})$. and σ^2 is the noise power at receiver. The data queue length should be satisfied:

$$0 < Q^t < +\infty, \forall t \in T \tag{6}$$

Our objective is to design a energy harvesting, link selection and power control algorithm for e_k^t, l_k^t and P_k^t to maximize the long-time averaged transmission rate. It can be formulated as the following optimization problem:

$$\begin{aligned} P1: \max_{\mathbf{l^t}, \mathbf{P^t}, \mathbf{e^t}} \quad & \frac{1}{T} \lim_{T \to \infty} \sum_{t=0}^{T-1} E(l_s^t R_s^t + l_r^t R_r^t) \\ s.t. \quad & (1), (2), (3), (6), \\ & l_k^t \in \{0, 1\},\ k \in \{s, r\}, \qquad (7) \\ & 0 \leq P_k^t \leq P_k^{max}, k \in \{s, r\}, \qquad (8) \end{aligned}$$

where the constraint (7) is the zero-one indicator for link selection, constraint (8) indicates that the transmit power's upper bound.

3 Lyapunov Optimization Algorithm

In this section, we present the Lyapunov optimization algorithm to solve our problem [4]. The main idea is to construct a Lyapunov optimization algorithm with perturbed weights for determining the energy harvesting, power allocation and link selections. Therefore, we first choose the perturbation parameters θ_s for $link_s$ and θ_r for $link_r$, and then define the quadratic Lyapunov function as

$$\mathrm{L}_t = \frac{1}{2} \sum_k (B_k^t - \theta_k)^2 + \frac{1}{2} (Q^t)^2, k \in \{s, r\}, \tag{9}$$

We will later define proper perturbation parameters, the energy causality constraint (3) will always meet, therefore we can drop this constraint in the problem. As the Lyapunov function (9) grows large as the data queue length increase or the energy queue length decreases. Therefore, we can stabilize the Lyapunov function to stabilize the queue length. The Lyapunov drift is the change of the Lyapunove between two slots. We can define as follows:

$$\triangle t = E[L^{t+1} - L^t \mid \mathbf{Z^t}], \tag{10}$$

where $\mathbf{Z^t} = (\mathbf{B_k^t}, \mathbf{Q^t})$, and the expectation is taken over the channel, energy queue and data queue states.

We aim to minimize the expected sum rate while stabilizing the queues. Therefore, the Lyapunov drift-plus-penalty function can be expressed as:

$$\Delta V^t = \Delta t - VE[l_s^t R_s^t + l_r^t R_r^t \mid \mathbf{Z^t}], \tag{11}$$

We have the following proposition holds for the upper bound.

Proposition 1. *For any feasible values* $\mathbf{l^t}, \mathbf{P^t}, \mathbf{e^t}$*, the drift-plus-penalty function has the following upper bound:*

$$\begin{aligned}\Delta V^t \leqslant & E[\sum_{k=s,r}(B_k^t - \theta_k)(e_k^t - l_k^t P_k^t \Delta t) + Q^t(l_s^t R_s^t \Delta t B - l_r^t R_r^t \Delta t B) \\ & - V[l_s^t R_s^t + l_r^t R_r^t] \mid \mathbf{Z^t}] + c,\end{aligned} \tag{12}$$

where c is a constant that

$$c = \frac{1}{2}\sum_{k=s,r}[(E_k^{max})^2 + (P_k^{max}\Delta t)^2 + (R_k^{max}\Delta t B)^2] \tag{13}$$

Proof: For the dynamics of B_k^t in (4), we have:

$$\begin{aligned}&\frac{1}{2}[\sum_{k=s,r}(B_k^{t+1} - \theta_k)^2 - \sum_{k=s,r}(B_k^t - \theta_k)^2] \\ &= \frac{1}{2}\sum_{k=s,r}[(e_k^t - l_k^t P_k^t \Delta t)^2 - 2(B_k^t - \theta_k)(e_k^t - l_k^t P_k^t \Delta t)] \\ &\leq \frac{1}{2}\sum_{k=s,r}[(E_k^{max})^2 + (P_k^{max}\Delta t)^2] - 2(B_k^t - \theta_k)(e_k^t - l_k^t P_k^t \Delta t)]\end{aligned} \tag{14}$$

Using the similar approach, for the dynamics of Q^t, we get:

$$\begin{aligned}&\frac{1}{2}[(Q^{t+1})^2 - (Q^t)^2] = \frac{1}{2}[(Q^t - l_r^t R_r^t B\Delta t + l_s^t R_s^t B\Delta t)^2 - (Q^t)^2] \\ &\leq \frac{1}{2}[\sum_{k=s,r}(R_k^{max}\Delta t B)^2 + Q^t(l_s^t R_s^t \Delta t B - l_r^t R_r^t \Delta t B)]\end{aligned} \tag{15}$$

Adding up (14) and (15), taking the expectation conditioned on $\mathbf{Z^t}$, define

$$c = \frac{1}{2} \sum_{k=s,r} [(E_k^{max})^2 + (P_k^{max} \Delta t)^2 + (R_k^{max} \Delta t B)^2] \tag{16}$$

From that we can get the desired upper bound of drift-plus-penal function.

Minimizing the drift-plus-penalty function is difficult while we can minimizing its upper bound instead. The Lyapunov algorithm only need the current system state, therefore, we minimize the upper bound of the drift-plus-penalty function in a per-slot problem. Removing the expectation, $\mathbf{Z^t}$ and constant c, we have following equivalent problem:

$$\begin{aligned} P2: \min_{\mathbf{l^t},\mathbf{P^t},\mathbf{e^t}} \quad & \sum_{k=s,r} (B_k^t - \theta_k)(e_k^t - l_k^t P_k^t \Delta t) + Q^t(l_s^t R_s^t \Delta t B - l_r^t R_r^t \Delta t B) \\ & - V(l_s^t R_s^t + l_r^t R_r^t) \\ s.t \quad & (1), (2), (6), (7), (8) \end{aligned}$$

To solve $P2$, We can divide it into two subproblem. The first one is to obtain the optimal harvesting energy, which is a Linear programming problem:

$$\begin{aligned} \min_{\mathbf{e^t}} \quad & \sum_{k=s,r} (B_k^t - \theta_k) e_k^t \\ s.t \quad & (2) \end{aligned} \tag{17}$$

The optimal solution for harvesting energy is:

$$e_k^{t*} = \begin{cases} E^{max}, B_k^t - \theta_k \leqslant 0 \\ 0 \quad , B_k^t - \theta_k > 0 \end{cases} \quad k \in \{s, r\}, \tag{18}$$

The other subproblem is for optimal link selection and power control:

$$\min_{\mathbf{l^t},\mathbf{P^t}} \quad \sum_{k=s,r} (B_k^t - \theta_k)(-l_k^t P_k^t \Delta t) + Q^t(l_s^t R_s^t \Delta t B - l_r^t r_r^t \Delta t B) - V(l_s^t R_s^t + l_r^t R_r^t) \tag{19}$$

$$s.t \quad (1), (6), (7), (8)$$

The constraint (7) is non-convex. However, the alternation iterations method can be used for solving this problem. Given the link selection first, we can get the following problem. For source to relay link:

$$\begin{aligned} \min_{P_s^t} \quad & (B_s^t - \theta_s)(-P_s^t \Delta t) - (V - Q^t \Delta t B) R_s^t \\ s.t \quad & (6), (8) \end{aligned} \tag{20}$$

By differentiating the objective function, we can derive a pole solution as:

$$P_s^{t'} = \frac{Q^t \Delta t B - V}{ln2(B_s^t - \theta_s)\Delta t} - \frac{\sigma^2}{h_s^t} \tag{21}$$

Since the objective function and constraint are convex, the problem can be solved by CVX. The objective function is a decreasing function when $B_s^t - \theta_s \leqslant 0$, therefore the optimal value is $P_s^* = P_s^{max}$. When $B_s^t - \theta_s > 0$, the objective is not monotonic, when the pole $P_s^{'}$ meets $0 < P_s^{'} < P_s^{max}$, $P_s^* = P_s^{'}$, otherwise we can get the optimality at its boundary. Therefore, we can get the P_s^* as following:

$$P_s^{t*} = \begin{cases} P_s^{max}, & B_s^t - \theta_s > \dfrac{(Q^t \Delta t B - V)h_s^t}{ln2\Delta t(P_s^{max} + \sigma^2)} \\ P_s^{t'}, & \dfrac{(Q^t \Delta t B - V)h_k^t}{ln2\sigma^2 \Delta t} \leq B_s^t - \theta_s \leq \dfrac{(Q^t \Delta t B - V)h_s^t}{ln2\Delta t(P_s^{max} + \sigma^2)} \\ 0, & B_s^t - \theta_s < \dfrac{(Q^t \Delta t B - V)h_s^t}{ln2\sigma^2 \Delta t} \end{cases} \tag{22}$$

For relay to destination link:

$$\begin{aligned} \min_{P_r^t} \quad & (B_r^t - \theta_r)(-P_r^t \Delta t) - Q^t(R_r^t \Delta t B) - V R_r^t \\ s.t \quad & (6), (8) \end{aligned} \tag{23}$$

Considering the constraint (6), we have the optimal power allocation:

$$P_r^{t*} = \begin{cases} min(P_r^{max}, \dfrac{\sigma^2(2^{Q^t} - 1)}{h_r^t}), & B_r^t - \theta_r \in D1. \\ min(P_r^{t'}, \dfrac{\sigma^2(2^{Q^t} - 1)}{h_r^t}), & B_r^t - \theta_r \in D2. \\ 0, & B_r^t - \theta_r \in D3. \end{cases} \tag{24}$$

where the $P_r^{t'} = -\frac{Q^t \Delta t B + V}{ln2(B_r^t - \theta_r)\Delta t} - \frac{\sigma^2}{h_r^t}, D1 = \{B_r^t - \theta_r > -\frac{(Q^t \Delta t B + V)h_r^t}{ln2\Delta t(P_r^{max} + \sigma^2)}\}, D2 = \{\frac{-(Q^t \Delta t B + V)h_r^t}{ln2\sigma^2 \Delta t} \leq B_r^t - \theta_r \leq \frac{-(Q^t \Delta t B + V)h_r^t}{ln2\Delta t(P_r^{max} + \sigma^2)}\}, D3 = \{B_r^t - \theta_r < \frac{-(Q^t \Delta t B + V)h_s^t}{ln2\sigma^2 \Delta t}\}$.

Now, we consider the link selection solution, for the P_k^{t*} is obtained, we have:

$$\min_{\mathbf{l^t}} \quad \sum_{k=s,r} (B_k^t - \theta_k)(-l_k^t P_k^{t*} \Delta t) + Q^t \Delta t B(l_s^t R_s^{t*} - l_r^t R_r^{t*}) - V(l_s^t R_s^{t*} + l_r^t R_r^{t*}) \tag{25}$$

$$s.t \quad (1), (7)$$

Since the solution only has two probabilities, we compare the two situations of $l_s^t = 1$ and $l_r^t = 1$, then choose the one whose objective function more optimal. Thus, the link selection is finished.

We summarize the whole Lyapunov optimization algorithm as follow:

Algorithm 1. Lyapunov Optimization Algorithm.

1: At the beginning of each time slot, obtain the energy queue length B_k^t, the data queue length Q^t and the channel state h_k^t,$\forall k \in \{s, r\}$.
2: Decide e_k^t by (16) according to the energy queue length B_k^t and perturbation parameters θ_s, θ_r, $\forall k \in \{s, r\}$.
3: decide P_s^{t*} or P_r^{t*} by (19) and (21), choose the link by (22).
4: Update the energy queue length B_k^t and the data queue length Q^t.
5: Set t = t + 1.

4 Performance Analysis

In this section, we analyse the performance of our Lyapunov optimization algorithm. We first define the perturbation parameters. According to [8,9], we make the fading channel gain have upper-bounded. Assume the maximum channel gain is $\frac{|h_k^{max}|}{\sigma^2}$, which means that $|h_k^t| < |h_k^{max}|$. We have

$$\theta_s = P_s^{max}\Delta t + \frac{V|h_s^{max}|}{ln2\sigma^2}, \tag{26}$$

$$\theta_r = P_r^{max}\Delta t + \frac{V|h_r^{max}|}{ln2\sigma^2} \tag{27}$$

The energy queue level will be confined within a given interval with θ_s and θ_r.

Proposition 2. *The energy queue level have the upper bound of* $\theta_k + E_k^{max}$, *where* $k \in \{s, r\}$.

Proof: According to the solution, if $\theta_k \leq B_k^t \leq \theta_k + E_k^{max}$, we have $e_k^{t*} = 0$, therefore $B_k^{t+1} \leq B_k^t \leq \theta_k + E_k^{max}$. if $B_k^t \leq \theta_k$, we have $e_k^{t*} = E_k^{max}$, therefore $B_k^{t+1} \leq B_k^t + e_k^t \leq \theta_k + E_k^{max}$. The proposition proved.

The perturbation parameters shows that the energy causality constraint (3) will not be violated, and the power solutions are feasible to P1. To verify the algorithm's asymptotic optimality, we need to relax the original problem $P1$ to assist the algorithm design. We define the relax problem $P3$

$$\begin{aligned} P3: \max_{\mathbf{l^t},\mathbf{P^t},\mathbf{e^t}} \quad & \frac{1}{T}\lim_{T\to\infty}\sum_{t=0}^{T-1} E(l_s^t R_s^t + l_r^t R_r^t) \\ s.t. \quad & (1),(2),(7),(8) \\ & \frac{1}{T}\lim_{T\to\infty}\sum_{t=0}^{T-1} E(e_k^t - l_k^t P_k^t \Delta t) = 0, k \in \{s, r\}, \qquad (28) \\ & \frac{1}{T}\lim_{T\to\infty}\sum_{t=0}^{T-1} E(l_s^t R_s^t B\Delta t - l_r^t R_r^t B\Delta t) = 0, \qquad (29) \end{aligned}$$

In $P3$, the energy causality constraint (3) is replaced by constraint (28), the data queue constraint (6) is replaced by constraint (29). It indicates that in

the long-term, the average harvesting energy equals to the average energy for transmission, the average data transmission to relay equals to the average data for destination. Moreover, any feasible solution for $P1$ is also feasible for $P3$.

Proposition 3. *There exists a stationary and randomized solution for $P3$, which meet all the constraints except for (28), (29), will satisfied the following equalities:*

$$E(l_s^t R_s^t + l_r^t R_r^t) = P3^* + \varepsilon \quad \forall t \in T \tag{30}$$

$$E[\sum_{k=s,r} (e_k^t - l_k^t P_k^t \Delta t)] = \gamma, k \in \{s, r\}, \forall t \in T \tag{31}$$

where $\varepsilon \geq 0, \gamma \geq 0$,$P3^*$ means problem $P3's$ optimal solution.

Proof: The proof can be obtained by Theorem 4.5 in [9] which is omitted.

Proposition 4. *The lyapunov algorithm's asymptotic optimality can be achieved:*

$$P1^* \leq P2^* + \frac{c}{V}. \tag{32}$$

Proof: Since the lyapunov algorithm obtains the optimal solution of its drift-plus-penalty function, we have:

$$\begin{aligned} &E[L^{t+1} - L^t \mid \mathbf{Z^t}] - VP2^* \leqslant \Delta V^t = \ E[L^{t+1} - L^t \mid \mathbf{Z^t}] - VE[l_s^t R_s^t + l_r^t R_r^t] \\ &\leqslant E[\sum_{k=s,r} (B_k^t - \theta_k)(e_k^t - l_k^t P_k^t \Delta t) - Q^t \Delta t B(l_r^t R_r^t - l_s^t R_s^t) - V(l_s^t R_s^t + l_r^t R_r^t)] + c \\ &\leqslant E[\sum_{k=s,r} (B_k^t - \theta_k)\gamma - V(P3^* + \varepsilon)] + c, \end{aligned} \tag{33}$$

Since the B_k^t have upper bound, letting the $\gamma \to 0, \varepsilon \to 0$, and for problem $P3$ is a relaxation of $P1$, i.e., $P1^* \leq P3^*$, we obtain:

$$P1^* - P2^* \leq P3^* - P2^* \leq \frac{1}{V}(-E[L^{t+1} - L^t \mid \mathbf{Z^t}] + c) \leq \frac{c}{V}, \tag{34}$$

Summing up the equations for $t = 0, \cdots T-1$, dividing by T and letting $T \to \infty$, the asymptotic optimality is proved. when the $\frac{c}{V} \to 0$, the lyapunov algorithm asymptotically achieves the optimal performance of the original problem $P1$.

5 Simulation Results

In this section, we provide the simulation results of our algorithm. We consider the three-node relay with complex Gaussian random channel, in addition, $\Delta t = 1\,\text{ms}$, $B = 1\,\text{MHz}$, $\sigma = 1$, $P_k^{max} = 2\,\text{w}$, the energy e_k^t is uniformly distributed between 0 and E_k^{max}, $E_k^{max} = 1\,\text{w}$. The link selection with control parameter

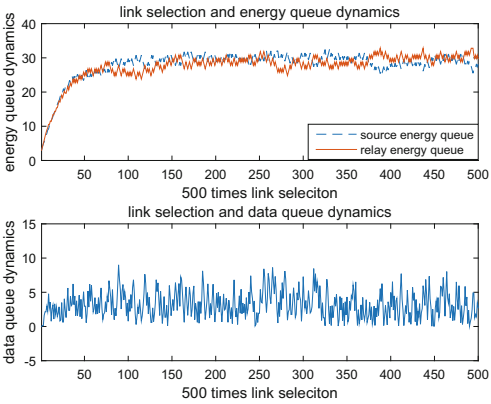


Fig. 2. Link selection with queue dynamics

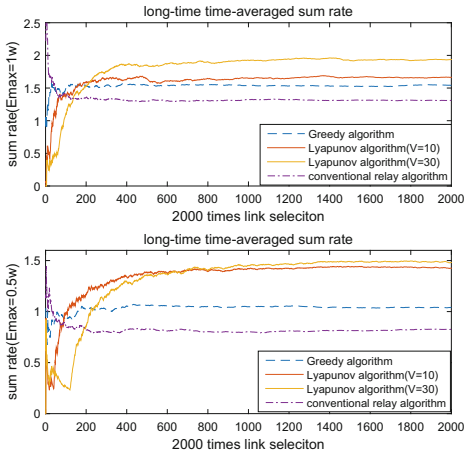


Fig. 3. Long-term average transmission rate comparison

V = 15 is shown in Fig. 2. As shown, the energy queue is fluctuant within a range of 30, the data queue fluctuate between 0 and 10, which means link selection is efficient, verifies the Lyapunov algorithm would stabilize the queue.

The long-time time-average sum rate (bps/Hz) with respect to time is shown in Fig. 3. As shown, the larger value of V, the better long-time sum-rate performance of Lyapunov algorithm. For comparison, we propose the greedy algorithm, which optimal the sum-rate for every slot based on current energy queue and data queue states [6]. We also consider the conventional relay network, with the first slot for source to relay link and the second slot for relay to destination link [3]. For the different amount of energy harvesting, the greedy algorithm always outperform the conventional relay algorithm, while the Lyapunov algorithm always outperform these two algorithms. That is because the conventional

relay algorithm will be disturbed by channel state, the greedy algorithm will experience the energy shortage, while the Lyapunov algorithm could choose the channel with better state and keep the queues' stability.

6 Conclusions

In this paper, we consider a three-node EH relay network with link selection. We design a Lyapunov optimization approach aiming at maximizing the long-term averaged transmission under the energy queue and data queue length constraints, and obtain the asymptotically optimality. Simulations show that the Lyapunov algorithm outperform the greedy algorithm and have low-complexity. It will be interesting to extend our algorithm to more EH network.

Acknowledgments. This work is supported by Guangzhou education bureau science and technology project 1201630400.

References

1. Van Der Meulen, E.C.: Three-terminal communication channels. In: Advances Applied Probability, vol. 3, pp. 120–154 (1971)
2. Zlatanov, N., Schober, R., Popovski, P.: Buffer-aided relaying with adaptive link selection. IEEE J. Sel. Areas Commun. **31**, 1530–1542 (2013)
3. Ahmed, I., Ikhlef, A., Schober, R., Mallik, R.K.: power allocation for conventional and buffer-aided link adaptive relaying systems with energy harvesting nodes. IEEE Trans. Wirel. Commun. **13**(3), 1182–1195 (2014)
4. Huang, L., Neely, M.J.: Utility optimal scheduling in energy-harvesting networks. IEEE/ACM Trans. Netw. **21**(4), 1117–1130 (2013)
5. Mao, Y., Zhang, J., Letaief, K.B.: A Lyapunov optimization approach for green cellular networks with hybrid energy supplies. IEEE J. Select. Areas Commun. **33**(12), 2463–2477 (2015)
6. Amirnavaei, F., Dong, M.: Online power control optimization for wireless transmission with energy harvesting and storage. IEEE Trans. Wirel. Commun. **15**(7), 4888–4901 (2016)
7. Choi, K.W., Kim, D.I.: Stochastic optimal control for wireless powered communication networks. IEEE Trans. Wirel. Commun. **15**(1), 686–698 (2016)
8. Neely, M.J., Huang, L.: Dynamic product assembly and inventory control for maximum profit. In: Proceedings of IEEE Conference Decision Control (CDC), pp. 2805–2812 (2010)
9. Neely, M.J.: Stochastic Network Optimization with Application to Communication and Queueing Systems. Morgan & Claypool, San Rafael (2010)

Parallel Symbol Timing Recovery Using FPGA for 600 Msps QPSK

Di Huang[1(✉)], Zhijie Wang[2], Jun Wang[2], and ZiYao Liu[1]

[1] Beijing Institute of Technology, Beijing 100081, China
huangdibit@163.com, liuziyaochn@163.com
[2] Institute of Telecommunication Satellite, China Academy of Space Technology, Xi'an 710000, China
1554264456@qq.com, bitwj@163.com

Abstract. This paper presents an efficient parallel symbol timing architecture for high data rate communications receivers. The presented architecture relies on a modified version of the classic Gardner loop, and it features a "multi-channel pipeline" interpolator that enables the symbol rate to be several times higher than the clock rate of the FPGA, hence maximize the achievable throughput. The presented timing recovery scheme is demonstrated on a Xilinx XC7VX690T FPGA at 150 MHz clock rate together with an ADC at 4.8 GHz sampling rate, for an QPSK data-stream at 600 Msps symbol rate. Also, it is observed the presented scheme occupies only 2% of the logic, storage and computational resources in the targeted FPGA. With minor modifications, our algorithm may be adapted for other Amplitude-Phase modulation constellations such as 8PSK, 16PSK or QAM.

Keywords: Symbol timing recovery · Interpolation · Multi-channel pipeline FPGA

1 Introduction

Symbol synchronization, a.ka. timing recovery is one of the key techniques for digital communications receivers. In principle, symbol synchronization is to find the best sampling locations for each symbol [1], from an input baseband digital waveform at a sampling rate of f_{smp}, and usually f_{smp} is chosen such as $f_{smp} = N \cdot R_s$, where R_s is the symbol rate and N is a positive integral. Classic timing recovery methods in their original forms, for example the Gardner loop [2, 3], assumes at the receiver may perform Digital Signal Processing operations at a clock rate, say f_{clk}, equal to or larger than f_{smp}, and that is the very start-point for many practical digital receivers. However, increasing symbol rate means increasing information transmission bandwidth, which is highly demanded in numerous application contexts, ranging from global satellite communication systems to UAVs' 4K video transmission, just to name a few. As R_s rises to a level that f_{smp} or even R_s is larger than f_{clk}, symbol synchronization itself becomes a bottleneck to the overall design of the digital receiver, as now it must be implemented in a *parallel* way [4–7], and be capable of handling very high speed data throughput.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 216–225, 2018.
https://doi.org/10.1007/978-3-319-78130-3_23

The main contribution of this paper is that we have designed an all-parallel symbol timing strategy for high symbol rate digital communications signals. Our strategy is based on an improved version of the classic Gardner loop, which comprises four principal components namely the Timing Error Detector (TED), the Loop Filter (LF), the Digital Controlled Oscillator (NCO) and the interpolator [8]. Among the above four, the parallel implementation of the interpolator poses the greatest challenge, and it is where we put our emphasis. In particular, we presented a "multi-channel pipeline" structure for the interpolator, where K identical channels all work at f_{clk}, and together they yield an overall throughput of $R_{\text{s}} = K \cdot f_{\text{clk}}$. On each channel, a Farrow interpolator is implemented by LUT, and the signal processing on the channel is designed as fully pipelined so as to achieve the maximum possible single-channel throughput f_{clk}. Low-complexity TED and LF are also discussed. The presented symbol timing recovery scheme is implemented on a Xilinx XC7VX690T FPGA working at $f_{\text{clk}} = 150\,\text{MHz}$. The FPGA accepts the output of a 4.8 GHz ADC, and performs symbol synchronization for a 600 Msps-QPSK signal at an Intermediate Frequency (IF) of $f_{\text{i}} = 1.2$ GHz. Experimental results are provided to testify the effectiveness of the proposed parallel scheme. It is observed that the proposed scheme only takes 2% of the logic, computational and storage resources of the targeted FPGA.

The remainder of the paper is organized as follow. We first describe a scheme of the "parallel high data rate" demodulator and illustrate the module of timing recovery in which. We dedicate a section to analyses the "multi-channel pipeline" interpolator which enables the symbol rate to be four times higher than the clock rate of the FPGA. Next, we describe each module' implementation details in the loop. The test result, such as constellation map and convergence curve are given followed by conclusion.

2 Parallel Demodulator Scheme

Figure 1 shows the main scheme of the parallel receiver. The Intermediate Frequency (IF) $f_{\text{i}} = 1.2\,\text{GHz}$, the 600 Msps IF signal is sampled at 4.8 GHz by EV10AQ190 One-channel Mode 10-bit A/D converter. The IF signal sampled by ADC is parallelized into 16 channels by a Xilinx FPGA built-in module, namely the Input Serialize & Reserialize (ISERDES), and then fed into 16 channel parallel FIFOs. All the subsequent operations are clocked at the a rate of $f_{\text{clk}} = 150\,\text{MHz}$.

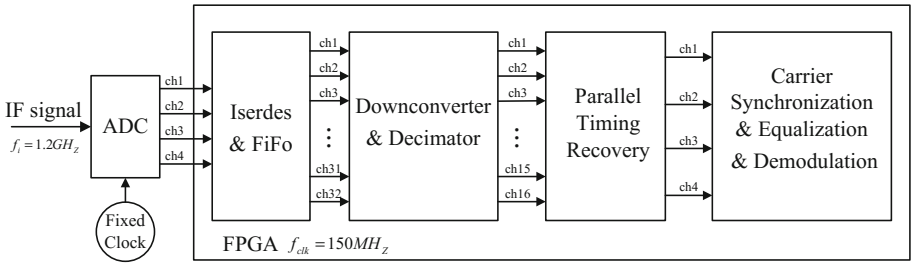


Fig. 1. The scheme of parallel high data rate receiver

Inside the FPGA, the 16-channel parallel Digital Downconverter (DDC) turns the IF sampled signal into digital I/Q baseband signal. A Decimator is inserted to reduce the sampling rate from 4.8 GHz to 2.4 GHz, so as to reduce the computational resource consumption. Even so, I/Q baseband signal at the rate of 2.4 GHz is still quite a big challenge for the subsequent timing recovery which has to work at a much lower clock rate of $f_{clk} = 150$ MHz. And the solution is the emphasis of this paper, namely a high throughput timing recovery featuring a "multi-channel pipeline" architecture. The other signal processing modules that follow the timing recovery, namely carrier synchronization, equalization and demodulation, as can be seen in Fig. 1, are out of the scope of this paper, and hence shall not be discussed.

2.1 Parallel Timing Recovery Architecture

As a key part of the receiver, Parallel Timing Recovery turns 16 channels QPSK baseband signal into 4 channels at the FPGA clock rate $f_{clk} = 150$ MHz. The block diagram of the symbol timing recovery is shown in Fig. 2.

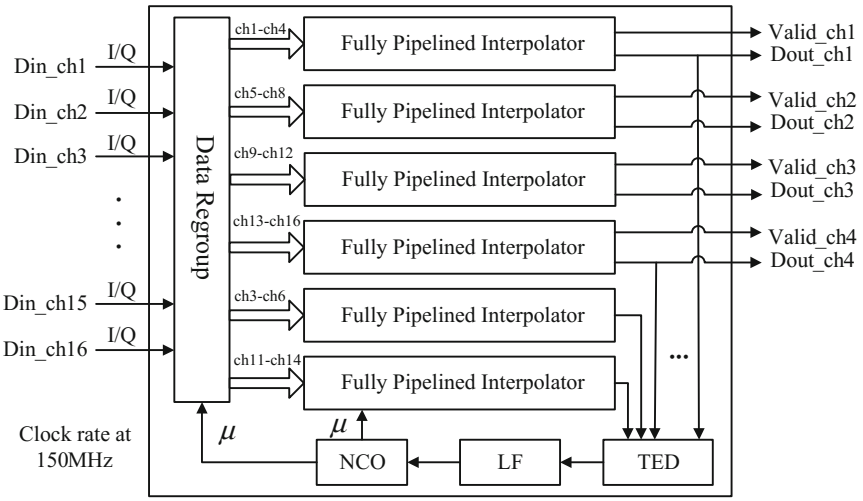


Fig. 2. Parallel timing recovery architecture

Data Regroup chooses 4 groups signal to Interpolators according to the fractional interval μ, and there are 4 channels in each group. Fully Pipelined Interpolator uses "multi-channel pipeline" structure to get 1 channel timing recovered signal from 4 channels signal, under the condition that the symbol rate R_s is 4 times higher than the clock rate f_{clk}. TED, LF and NCO are implemented with low complexity. Output data from interpolator also serve as the input of the TED. Loop Filter uses mean value of timing error estimation to compute control word, which afterwards would be sent to NCO. The fractional interval μ, which is generated from NCO to indicate the location, is sent to interpolator to correct the timing error.

2.2 Interpolator and Its Implementation

Gardner algorithm [9] is widely used in practical implementation for its simple structure. Polynomial interpolator can be efficiently implemented by Farrow Structure [10]. Farrow structure only need to transmit a fractional interval μ to compute an interpolation. This structure has been widely used in high speed digital system because it is fairly easy to implement. The calculation mathematical formulas of interpolation are as follows:

$$y(k) = \sum_{l=0}^{N} \mu_k^l v(l). \tag{1}$$

$$v(l) = \sum_{i=I_1}^{I_2} b_l(i) x(m_k i). \tag{2}$$

Where, $x(i)$ is the input of interpolator, and b(i) is the coefficient of Farrow structure calculated by Lagrange formula, which is a fixed value which is not related to the fractional interval μ. This structure pre-computes the coefficients b(i) and stores them in ROM [11, 12], which could be efficiently implemented by embedded Block RAM in FPGA. Figure 3 shows the "multi-channel pipeline" structure of interpolator.

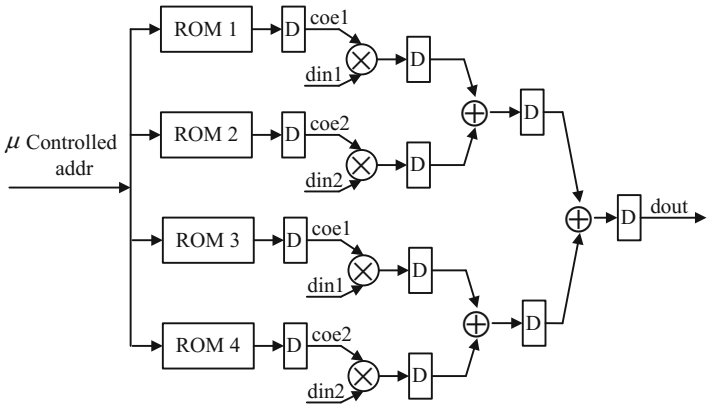


Fig. 3. Interpolator using ROM-based coefficients

Each channel has 4 sequential sub-modules. The first sub-module is implemented by LUT. Fractional interval μ would be used as the address to read the coefficients b(i) from ROM. The second sub-module is used to obtain the product of input signal and coefficient. The third and fourth sub-modules are used to get summation as the output of interpolator.

2.3 Timing Error Detector and Its Implementation

This paper selects the Gardner timing error estimation algorithm [13] to estimate the timing error. Each symbol needs only two sampling points, also the timing error need not to be cumulated. The step is obtained by

$$err(k) = y(k - 1/2) * (y(k) - y(k - 1)). \tag{3}$$

Timing Error Detector adopts the parallel pipeline structure, in which clock rate is $f_{clk} = 150\,\text{MHz}$. $y(k)$ is the interpolated value. If there is no timing error, $y(k - 1/2)$ should be zero. Otherwise, the difference between $y(k)$ and $y(k - 1)$ is proportional to the value of error and also indicates the error direction. The final timing error is obtained by the mean value of two channels timing error.

2.4 Loop Filter and Its Implementation

The loop filter adopts a proportional integral structure which is always used in phase-locked loop. Figure 4 shows its implementation structure.

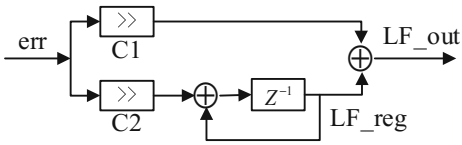


Fig. 4. The block diagram of loop filter

From Fig. 4, output of the loop filter Err_Lp is obtained by

$$LF_reg(k+1) = LF_reg(k) + C2 * err(k). \tag{4}$$

$$LF_out(k+1) = LF_reg(k+1) + C1 * err(k). \tag{5}$$

$err(k)$ is the output of TED. In order to not use any multipliers, C1 and C2 are realized by shift operation. Through unceasingly debug in the practical simulation, C1 and C2 finally obtain the appropriate value.

2.5 Timing Controller and Its Implementation

Timing controller computes the enable signal and the fractional interval μ through internal addition algorithm. An incrementing NCO controls the μ. NCO which increases a certain step W every sampling time is a module 1 counter. Figure 5 shows the implementation structure of the timing controller. η is the content of the NCO register at one clock before the strobe signal, ε_0 could be an approximate representation of T_i/T_s with little performance loss. T_s is the sampling interval, T_i is the bit interval to the filter. Using the shift operation to replace multiplier is very convenient.

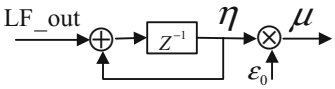


Fig. 5. Implementation structure of timing controller

2.6 Data Regroup and Its Implementation

According to Gardner algorithm, the data access location should move forward one more sampling data when $\mu > 1$, and it should move backward one sampling data when $\mu < 0$. Because of the parallel structure, the parallel input signal should be adjusted according to the fractional interval μ. This structure chooses 16 channels signal as interpolator inputs from 20 channels signal. Channel 1 to channel 4 are achieved from the previous sampling time and Channel 5 to channel 20 are achieved from this sampling time. Figures 6 and 7 show the implementation structure.

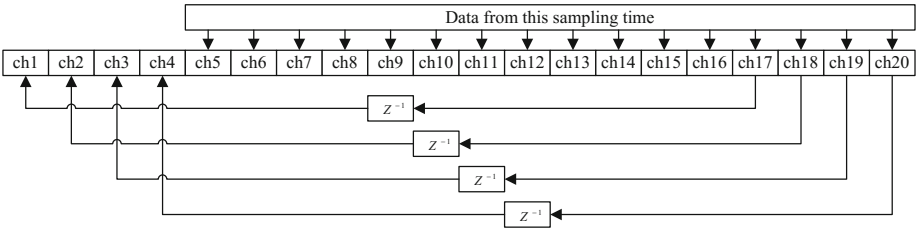


Fig. 6. Channel construction

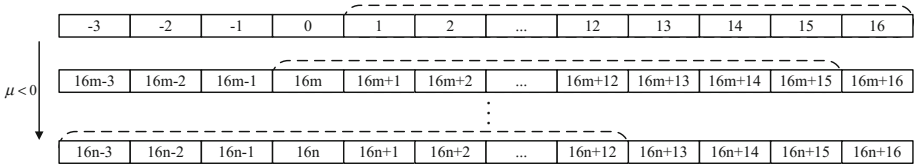


Fig. 7. Implementation structures of data regroup

The initial interpolation takes data(1) to data(16) as the interpolator inputs. Suppose $\mu < 0$ firstly occurred when interpolator has finished mth interpolation, data(16m) to data(16m + 15) should be inputted to interpolator. And so on, if the fourth $\mu < 0$ occurred after the nth interpolation, Data Regroup outputs data(16n − 3) to data(16n + 12) as the inputs of interpolator. If $\mu > 1$ *occurred*, channel 13 to channel 16 of interpolator's inputs are replaced by zero, and Dout_ch4 is not valid.

3 Simulation and Implementation Results

To verify the effectiveness of the symbol timing recovery scheme, we implement it on our own hardware platform, which is illustrated in Fig. 8. For a detailed list of the component in the used Software Defined Radio (SDR) hardware platform, the readers are referred to Table 1. As mentioned, the ADC samples the IF QPSK signal with a center frequency 1.2 GHz at a sampling rate of 4.8 GHz. The symbol rate of the IF signal is 600 Msps. Recalling Fig. 2, the timing recovery accepts a pair of down-converted IQ baseband signal at a sampling rate of 2.4 GHz (in the form of 16 channels, at the rate of 150 MHz each), and its outputs are 4 channel parallel optimally sampled symbol sequence at the rate of 150 MHz.

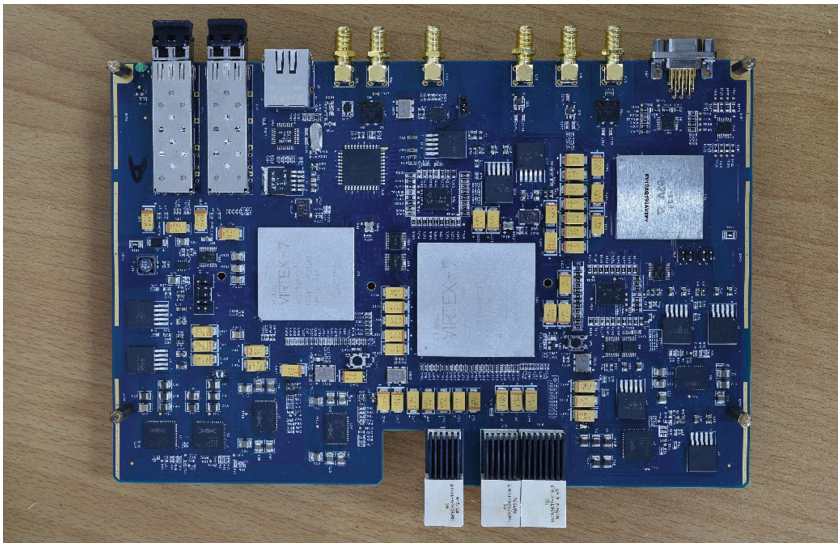


Fig. 8. Soft defined radio platform

Table 1 summarizes the resource utilization of the proposed symbol timing recovery scheme in the targeted FPGA model (Xilinx XC7VX690T FPGA). In particular, the overall resource consumption of the proposed timing recovery architecture is less than 2% of the total available amount in the targeted FPGA. The latest research [14] has been done for QPSK baseband signal with symbol rate of 45M. It's total implementation resources are 719 Slices and 18 embedded multipliers. By contrast, our proposed timing recovery schemes is not only suitable for high symbol rate but also very cost-efficient.

In Fig. 9 we give a typical convergence curve of the multi-channel pipeline timing recovery architecture, where the loop bandwidth is set as 0.001 of the symbol rate, i.e. 6 MHz, without loss of generality. The figure is produced by ChipScope, an integrated on-line diagnosis tool provided by Xilinx to monitor the real-time status of the FPGA's internal registers. In the figure the horizontal axis is the index of samples at the rate of

Table 1. Complexity comparison

Logic utilization	Xilinx XC7VX690T FPGA		
	Used	Available	Utilization
Number of slice	1860	75900	2%
Number of slice LUTS	3672	303600	1%
Number of block RAM/FIFO	4	2060	1%
Number of DSP48E1s	46	2800	1%

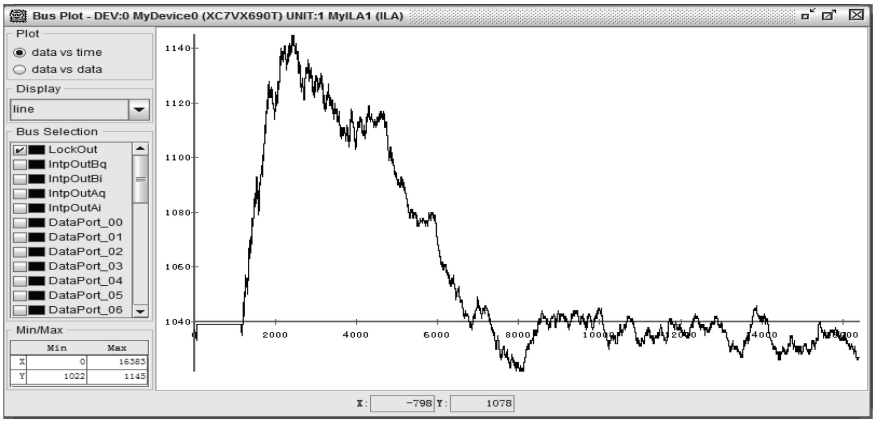


Fig. 9. Convergence curve of 600 Msps data rate

150 MHz, i.e., the FPGA's working clock. The vertical axis is the position of the optimal sampling epoch, which is discretized at a resolution of Tsmp/2^10. As observed from the figure, it takes about 6000 samples (at the rate of 150 MHz) or 24000 symbols (at the rate of 600 Msps) for the loop to enter a convergent state, beginning from a cold start. Moreover, the stable vibration of the loop is about ±20 according to the vertical axis in the figure, or ±0.02 Ts, translated by the relationship.

In Fig. 10 we plot the constellation with the output of the proposed timing recovery architecture. Figure 10(a) is generated by the raw data, i.e. the original output QPSK symbol samples after timing recovery. The constellation points are not very clear, like the input signal is noisy, but in fact we do not include noise in our test. The explanation for this phenomenon is when the symbol rate is as high as 600 Msps, even the coaxial cable between the test signal generator and the receiver will exhibit non-uniform amplitude & phase Response in the frequency domain, and thus cause Inter-Symbol-Interference (ISI). By comparison to Fig. 10(a), (b) is generated using the equalized version of the timing recovery outputs. The equalization in this example is carried out off-line by a Matlab routine, after the original raw data is collected from the FPGA by ChipScope. As can be seen in Fig. 10(b), the equalized constellation is very high-quality, indicating the proposed parallel timing recovery architecture works fine for 600 Msps QPSK digital communication signals.

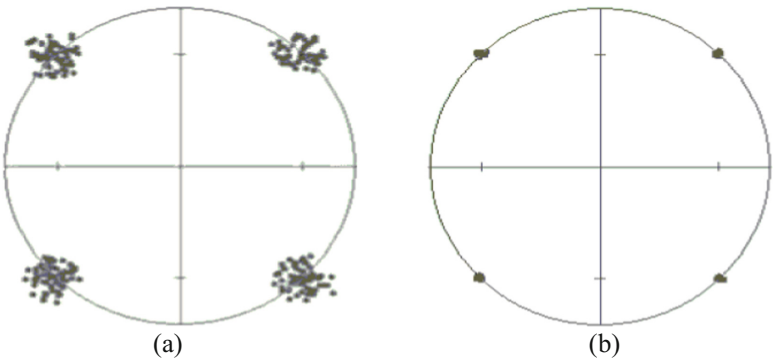


Fig. 10. Constellations of (a) output signal after timing recovery. (b) output signal after equalization.

4 Conclusion

In this paper we have designed a new all-digital "multi-channel pipeline" symbol timing recovery scheme, to bridge the gap between the extremely high symbol rate and the relatively low clock rate of the signal processors (usually FPGAs). The proposed scheme features a multichannel parallel pipelined architecture, and this allows it to be easily and efficiently implemented on FPGAs. As an example application of the proposed scheme, we have implemented it on a Software Defined Radio (SDR) platform for a 600 Msps QPSK IF signal. It is observed our timing recovery scheme could accept I/Q baseband signal at the rate of 2.4 GHz, based on a FPGA clocking rate of only 150 MHz. Also it is observed the proposed timing recovery algorithm only occupies about 2% of the logic, storage and computational resources in the targeted FPGA XC7VX690T. With some minor modifications, this scheme could be easily adapted to 8PSK, 16PSK or QAM and other higher data rata speed modem.

Acknowledgement. This work is supported by National Natural Science Foundation of China under contract No. 61471360

References

1. Ren, T., Petovello, M.: Collective bit synchronization for weak GNSS signals using multiple satellites. In: Position, Location and Navigation Symposium (2014)
2. Gardner, F.M.: Interpolation in digital modems. I. Fundamentals. J. IEEE Trans. Commun. **41**(3), 501–507 (1993)
3. Erup, L., Gardner, F.M., Harris, R.: Interpolation in digital modems. II. Implementation and performance. J. IEEE Trans. Commun. **41**(6), 998–1008 (1993)
4. Thompson, M., Jones, W.: An all digital 650-Mbps demonstration receiver for NASA's high data rate satellite applications. In: Military Communications Conference (1992)

5. Schmidt, D., Lankl, B.: Parallel architecture of an all digital timing recovery scheme for high speed receivers. In: International Symposium on Communication Systems Networks and Digital Signal Processing (2010)
6. Lin, C., Shao, B., Zhang, J.: A high data rate parallel demodulator suited to FPGA implementation. In: International Symposium on Intelligent Signal Processing and Communication Systems (2011)
7. Lin, C., Zhang, J., Shao, B.: A high speed parallel timing recovery algorithm and its FPGA implementation. In: International Symposium on Intelligence Information Processing and Trusted Computing (2011)
8. Sheng, K., Anjum, M.R., Dida, M.A.: Principle of bit-synchronization loop. In: Multi-Topic Conference (2014)
9. Vesma, J., Saramaki, T.: Interpolation filters with arbitrary frequency response for all-digital receivers. In: International Symposium on Circuits and Systems (1996)
10. Farrow, C.W.: A continuously variable digital delay element. In: International Symposium on Circuits and Systems (1988)
11. Cardells-Tormo, F., Perez-Pascual, A., Torres-Carot, V.: Design of a DVB-S receiver in FPGA. In: Signal Processing Systems (2003)
12. Srinivasan, M., Chen, C.C., Grebowsky, G.: An all-digital high data-rate parallel receiver. J. Surf. Sci. **604**(15–16), 1294–1299 (1997)
13. Gardner, F.M.: A BPSK/QPSK timing-error detector for sampled receivers. J. Trans. Commun. **34**(5), 423–429 (1986)
14. Jian, Z., Nan, W., Jingming, K.: High speed all digital symbol timing recovery based on FPGA. In: International Conference on Wireless Communications, NETWORKING and Mobile Computing (2005)



Joint Fingerprinting and Curve Fitting Scheme with NLOS Identification for Indoor Localization

Ruiyang Yuan[1,2], Xiaohui Li[1,2(✉)], Yanbin Zhao[3], Bin Zhou[4], and Si Huang[1]

[1] State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, Shaanxi, China
xhli@mail.xidian.edu.cn
[2] Collaborative Innovation Center of Information Sensing and Understanding, Xidian University, Xi'an 710071, Shaanxi, China
[3] Geographic Information Center, Yangquan, Shanxi, China
[4] AVIC Computing Technique Research Institute, Xi'an, Shaanxi, China

Abstract. Due to the wide deployment of wireless local area networks (WLANs) and the easy acquirement of received signal strength (RSS), indoor localization based on RSS has attracted considerable attention in both academia and industry. In this paper, we extract the scatter factor from RSS and then propose a novel indoor localization scheme based on location fingerprinting and curve fitting techniques. The scheme is set up in two phases. In the offline phase, we create a fingerprint and apply the curve fitting technique to construct fitted RSS-distance functions in LOS (Line Of Sight) and NLOS (None Line Of Sight) condition, respectively. In online positioning phase, the proposed optimized fingerprinting-based localization algorithm (OFPL) uses trilateration technique to assist the fingerprinting localization in choosing the proper location of the receiver. Through conducting field and extensive experiments, we can draw the conclusion that our proposed algorithm can obtain some improvement in localization accuracy compared with the classical fingerprinting-based localization.

Keywords: Scatter factor · NLOS identification · Curve fitting
Indoor localization · Location fingerprint

1 Introduction

Indoor localization is now increasing in popularity among academia and industry. There is a wide range of applications, e.g., location detection of exhibits in an exhibition hall, patient tracking inside the building of the hospital, and emergency personnel positioning in a disaster area. Although global positioning system (GPS) performs extremely well for outdoor localizations, it does not work effectively for indoor environments due to the disability of GPS signals to penetrate in-building materials. Therefore, navigating in indoor environments is beyond the coverage of GPS.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 226–239, 2018.
https://doi.org/10.1007/978-3-319-78130-3_24

The majority of the indoor locating systems and techniques can be summarized in two major classes: (1) fingerprints based and (2) range finding. The first class includes all those systems that use Received Signal Strength Indicator (RSSI) or Channel State Information (CSI) fingerprint maps built during a calibration phase and used as a real time information comparison reference to estimate the user's position. Systems that use fingerprint mapping include the Zhou [1], the IMLours [2], the FIFS [3], the Kim [4], etc. The second kind of systems uses RSSI or CSI propagation model to obtain the distance between a transmitter and a receiver, and then the trilateration is applied to calculate the transmitter's or receiver's position. The systems using the distance involve the FILA [5], the BLE [6], the Wang [7], the Kodippili [8] and the GPS.

However, most of the above systems have a common drawback, for instance, most systems using fingerprint mapping require precise and complex measurements before the real time location, and some even need using specialized hardware and software on the client devices. As for the range finding based systems, the CSI is usually explored to mitigate the effects of multipath, while the special 802.11 NICs(Network Interface Cards) are required [5]. In theory, it is possible to establish a model to estimate the separation distance using the received power. In practice, however, the propagation of a RF wave is attenuated by reflection, diffusion, deflection and transmission. In addition to the line-of-sight (LOS) signal, there are possibly multiple signals arriving at the receiver through different paths [9]. This multipath effect is even more severe in indoor environments where a ceiling, floor and walls are present. Consequently, it is not easy to establish a simple relationship between received power and separating distance if the environment factor is not taken into consideration.

Through extensive experiments, we argue that a better approach to improve the accuracy of indoor localization, is finding an environment factor which is stable and provides the capability to minimize the multipath effect. Based on RSSI, in this paper, we present the design and implementation of optimized fingerprint location (OFPL), a novel approach based on the environment factor and optimized fingerprint for indoor localization using WLANs. To achieve the target of fast and accurate indoor localization, signal processing techniques are leveraged to mitigate the multipath effect.

In summary, the main contributions of this paper are as follows:

- Different from applying the mean values and standard deviation information of RSS to diverse environment sensing scenarios, we extracted scatter factor from raw RSS. The factor has demonstrated its high sensitivity to NLOS identification.
- We construct a fitted RSS-distance functions with the scatter factor δ in LOS and NLOS conditions with curve fitting technology, which have been proved that they can greatly improve the distance estimation accuracy.
- Our field experiments demonstrate that OFPL consistently outperform previous approaches on localization accuracy and significantly reduce the latency, as compared with the classical fingerprinting-based localization.

The rest of this paper is organized as follows. In Sect. 1.1, we introduce some preliminaries. This is followed by the system architecture design in Sect. 2. In Sect. 3, we demonstrate the methodology. The implementation of OFPL and experimental evaluations are presented in Sect. 4. In Sect. 5, conclusions are presented.

1.1 Preliminaries

In this subsection, we first introduce the background information of the NLOS identification and then design the localization fingerprinting which is the foundation of OFPL. Finally, indoor propagation model is introduced. Indoor localization can be generally divided into two categories: propagation model based and fingerprinting based. In both categories, the localization process consists of two phases: the offline phase and the online phase. In the offline phase, propagation model or fingerprints are first measured, whereas in the online phase, a mobile device can achieve localization according to the RSS received at that time.

1.2 The NLOS Identification and Mitigation Problems

The complex indoor environment often creates perplexing radio signals, where they have to propagate via multiple intricate NLOS paths. It is common for the LOS path to be mixed with multiple NLOS paths, or too harshly attenuated to be perceivable against the noise floor. Hence the NLOS identification and mitigation problems are very necessary to be dealt to get an accurate multipath propagation model. Mathematically, the identification procedure is formulated as a binary hypothesis test [10]. For instance, received signals strength under LOS condition normally experience smaller fluctuation than NLOS condition, and this forms the basis for most PhaseU [11] based on LOS identification schemes.

1.3 Location Fingerprinting Problem

In general, when the referenced points are close in the Euclidean space, they often experience similar RSSs. This is because the radio propagation environments and distance between these points and the transmitter are similar. Furthermore, in indoor environments, concrete walls have a serious attenuation effect on radio transmission, which often cause several dBm path loss in the RSS. So it is necessary to choose reliable factors as location fingerprinting.

Firstly, a mobile phone is needed to measure RSS values (in dBm) and then be used in the calibration stage. Such data are collected from several surround APs at the reference points. After getting the RSS, some factors should be proposed as fingerprinting. Through observation of the RSS samples from LOS and NLOS conditions, some typical features can be extracted from the collected RSS measurements to identify unique location, including the mean, the standard deviation, and the Rician K factor [12].

1.4 Formulation Problem

Through reading a large number of references and doing various analysis, we can demonstrate that whether the LOS path of the wireless signals are blocked or not makes a very significant difference on the determination of the parameter in the propagation model, so the formulation of propagation in LOS and NLOS conditions should be different. In the traditional method, the distance measures can be estimated based on, for example, a logarithm-distance path-loss model [13] as shown in

$$PL\,(d)\,[dB] = PL\,(d_0)\,[dB] + 10n \log\left(\frac{d}{d_0}\right) + w \tag{1}$$

where the power $PL\,(d)$ is given in a logarithmic scale, d_0 is a reference distance, n is a path loss exponent, which is different between LOS and NLOS condition, and $w \sim N(0, \sigma^2)$ is a normally distributed random variable which models the slow shadowing phenomenon. While in indoor environment, we usually define $d_0 = 1$ m, so the $PL(d_0) = A$, and then the final path lose model [14] can be written as

$$PL\,(d)\,[dB] = A\,[dB] + 10n \log\left(\frac{d}{d_0}\right) + w \tag{2}$$

2 Feature Extruction from RSSI

In this section, we conduct numerous observations of multiple RSS samples in LOS and NLOS conditions in different scenarios. We first extract scatter factor from the RSS samples, and then remove the outliers and mitigate random noise. Finally, the candidate features are extracted from the processed RSS.

2.1 RSS Noise Filtering

Taking vulnerability of RSS into consideration, we propose to use a refined Hampel Filter [15] to reduce the noise effect and remove the outliers. The refined Hampel Filter can deal with impulse and Gaussian noise, while mean and median filters can handle only one of them [16]. In addition, it is simple to implement: the refined Hampel Filter firstly sorts the input vector, s, and then computes the median of a window composed of the vector and its sixty surrounding samples which is thirty per side. It also estimates the standard deviation of each sample about its window median using the median absolute deviation (MAD). If a sample differs from the median by more than three standard deviations, it is replaced with the mean of the sixty surrounding samples. The filtered RSS can be defined as Eq. (3),

$$s_i = \begin{cases} s_i & |s_i| \leq 3 * MAD \\ S_Mean & others \end{cases} \quad i = 1, \cdots, n \tag{3}$$

where s_i is one of the RSS sample at $i-th$ sampling time, n is the number of the sampling, and MAD stands the median absolute deviation of the input vector. The S_Mean represents the mean of the sixty surrounding samples.

2.2 The Descriptions of RSS Features

Through observations of multiple RSS in various scenarios, several typical features are extracted from a set of RSS samples, and then these features are applied to the NLOS identification and mitigation and fingerprinting.

(1) **The Scatter Factor** δ
In the LOS condition, there always exists a straight line between the receiver and the transmitter, but in the NLOS condition, the received signal experiences the reflection, absorption, refraction, diffusion and scattering all the time. So in the LOS condition, the absolute value of the difference between the sampling value and the mean value can be larger than the NLOS's. Taking the difference between the signal in the LOS and NLOS conditions into consideration, we formulate an Eq. (4) to calculate the feature. In case of LOS, it is $\delta < 0$, otherwise it is $\delta \approx 0$ (NLOS condition). Therefore, estimation of δ can be used as a criteria of NLOS condition. We should pay attention to that the scatter factor must be extracted from RSS measurements before preprocessing, because the scatter factor reflects the influence degree of environmental factors. The scatter factor is primarily used to discern LOS and NLOS.

$$\delta = \frac{\sum_{i=1}^{N} |s_i - \bar{s}|}{N \cdot \bar{s}} \tag{4}$$

where N is the number of samplings, s_i is one of the RSS samples at $i-th$ sampling time and the $\bar{s} = \frac{1}{N} \sum_{i=1}^{N} s_i$ is the mean of RSS samples.

(2) **The Mean and the Standard Deviation** (μ, σ)
The mean μ and standard deviation σ can be well suitable to identify a unique location in LOS or NLOS conditions with feature discussed below.

(3) **The Rician *K* Factor** (Kr)
Existing theoretical and empirical studies have shown that the Rician K factor is defined as the ratio between the power in the direct path and the power in other scattered paths, $Kr = \frac{\mu^2}{2\sigma^2}$. Therefore, in NLOS condition where no direct path exists, the Rician K factor is expected to be zero. The features extracted from RSS are shown in Fig. 1.

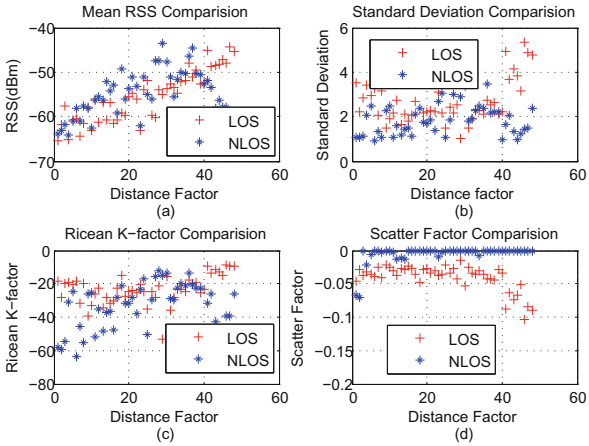


Fig. 1. Features extracted from RSS under NLOS/LOS conditions

3 NLOS Identification and Mitigation

According to the great NLOS identification method proposed in the PhaseU [11] and LiFi [10], and some experiments we have conducted, we find a simple but efficient method to achieve NLOS identification. Although we have advanced superior features, tough challenges reside delivering the features to enable a practical system.

3.1 NLOS Identification with Scatter Factor

In this subsection, we present a decision test in order to decide for the presence of LOS or NLOS components through a moment-based estimation of the scatter factor. In the following, we use the definition previously described in Sect. 1.1 in order to implement a decision test to discern the LOS and NLOS propagation: if the scatter factor δ assumes a value less than one, the propagation between transmitter and receiver can be defined as LOS. Otherwise, the transmission identifies NLOS connections.

Given a set of RSS samples, we calculate the scatter factor as in Sect. 2 and then formulate the following typical binary hypothesis test with LOS condition H_0 and NLOS condition H_1 [17,18].

$$\begin{cases} H_0 : \delta < \delta_{th} \\ H_1 : \delta > \delta_{th} \end{cases} \tag{5}$$

where δ_{th} denotes a pre-defined threshold, which is obtained through a lot of experiments. Through a large of measurements test in various indoor environments, a pre-defined and calibrated unified threshold can be applied to most scenarios very well.

3.2 NLOS Mitigation Using Curve Fitting Technique

NLOS mitigation is achieved by using the curve fitting technique in LOS and NLOS, respectively. For LOS and NLOS conditions the parameters in (5) can vary greatly, which will produce significant errors in distance estimation when inappropriate parameters are used. Therefore, we first identify the LOS/NLOS conditions using the scatter factor δ and then estimate the transmitter-receiver distances with different propagation models for LOS and NLOS conditions.

The purpose of NLOS mitigation is to reduce the ranging errors in NLOS conditions. Due to the complicated indoor environment and vulnerable RSS, it is quite difficult to apply the curve fitting technique to establish the RSS-distance relation in indoor environment. The objective of curve fitting is to construct a curve (a mathematical function) that has the best fit to a series of data [19]. In this paper, we use curve fitting technique to build an RSS-distance fitting function to describe the relation between the RSS and distance to the transmitter. Instead of the traditional log-distance propagation model, we propose to build a fitting function by two main parameter, RSS and δ, in LOS and NLOS conditions, respectively.

$$d = \begin{cases} \alpha_L \cdot s^2 + \beta_L \cdot s + \gamma_L \cdot \delta + c_L & LOS \\ \alpha_N \cdot s^2 + \beta_N \cdot s + \gamma_n \cdot \delta + c_N & NLOS \end{cases} \tag{6}$$

where s is received power in dBm, and α, β, γ and c indicate three unknown parameters, which are relative to the environment. The d denotes the distance between the receiver and the transmitter.

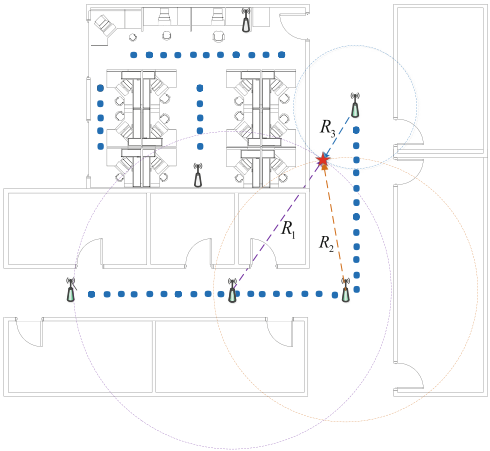


Fig. 2. The typical experiment environment. The dots represent the Reference Point locations and the access points are ceiling type AP (Color figure online)

3.3 Location Fingerprinting

The fingerprinting requires the full training database information for all the N fingerprints: $(x_i, y_i, \mu_i, \sigma_i, Kr_i)$,$i = 1, \cdots, N$, obtained during the offline phase. Figure 2 illustrates the typical experiment indoor environment. To reduce the workload of the previous offline phase without impacting on the location accuracy, we choose the reference points as shown in the Fig. 2.

4 Experiments and Evaluation

In this section, we introduce the experiment setup and methodology firstly, and then interpret the performance evaluation of proposed NLOS identification approach and curve fitting formulas, and accuracy of localization. Figure 3 shows the flow chart of the experiments. In the offline phase, we firstly get the RSS of the every reference points, and secondly extract the scatter factor and feature, and finally figure out the RSS-distance fitting function in LOS and NLOS conditions. In the online phase, first, we use the mobile phone to get the RSS. Second, we extract the scatter factor from the raw RSS and get the feature from the RSS which is processed by the Hampel Filter. Third, through the Eq. (6) we can calculate the distances between the mobile phones and the APs and then using the Least Square Method, we figure out a position through the set of Eq. (7), which is marked with a red star as shown in the Fig. 2. Generally, because of the noise and signal fluctuation, the red star is usually not the position of the mobile. Finally, we use the red star's position and the reference points to calculate the mobile phone's position.

$$(x_i - x)^2 + (y_i - y)^2 = d_i, i = 1, 2, \ldots, n \tag{7}$$

where (x, y) is the mobile phone's location, and (x_i, y_i) is the $i - th$ AP's location, d_i denotes the distance between the $i - th$ AP and the mobile phone, and n is the number of APs.

4.1 Experiment Setup

We conduct experiments in a typical laboratory environment, the first floor of one of the laboratory building in our university, which includes laboratory, office rooms and corridors, as shown in Fig. 2. The majority of RSS samples were collected along the corridors surrounded by rooms which are surrounded by soil and rocks. The corridors' width is about 3 m and the APs in corridors are all ceiling type, with operating frequencies of 2.4 GHz and 5 GHz. In our experiments, we deploy seven APs. The Reference points are created according to the normal fingerprint technique by a grid approach.

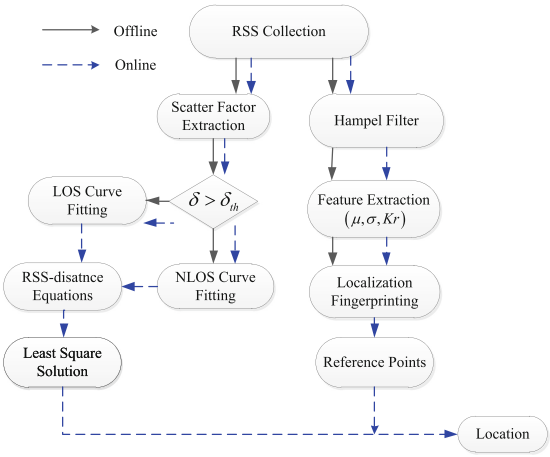


Fig. 3. Flow chart of the experiments

4.2 Data Sets and Training

Data collection: RSS is collected along the corridors and different rooms. The measurements are extracted from about 100 reference spots with different environment settings, i.e., TX-RX distances vary from 1 m to 30 m in different scenarios. 240 samples at each given Reference points are collected. Data are collected from diverse routers with multiple antennas operating in IEEE802.11n AP mode at 2.4 GHz. A Huawei smart phone is used to scan the RSS of the transmitters, which runs a specific app. During our experiment, we do not limit the movement of the people. Generally, more people move around and in and out the rooms at the office time.
Data process: First, we extract scatter factor from a set of RSS measurements. Second, we filter the RSS measurements with the refined Hemple Filter. Third, we extract the other features mentioned in Sect. 1.1. Finally, curve fitting formulas and fingerprints are worked out.

4.3 Experiment Results

We compare our proposed OFPL algorithm with the traditional fingerprint-based $K - th$ nearest neighbor algorithm. In OFPL, we first locate a mobile phone to an area according to the AP which has the strongest signals. Then we conduct the experiment as the online phase describes in the Sect. 4. Thirdly, we sort the m in ascending order which is calculated using the Eq. (8) and get the first three reference points. Finally, we calculates the distances between the three reference points and the red star which represents the location of the mobile calculated by the Eq. (7) and give different weights to the three reference points according to the distances. The smaller the distance, the greater the weight, and we figure out the mobile phone's position.

$$m_j = \sqrt{\sum_{i=1}^{n} \left[(\mu_i - \mu_{ij})^2 + (\sigma_i - \sigma_{ij})^2 + (Kr_i - Kr_{ij})^2 \right]} \tag{8}$$

where m_j denotes the matching degree between the mobile phone position and the $j-th$ fingerprints, μ_i is the mean of the RSS of $j-th$ reference point from the $i-th$ AP in the received signal, μ_{ij} is the mean of the RSS from the $i-th$ AP in the signal of the fingerprint, and $j = 1, 2, \ldots, m$, m is the number of the fingerprints in the database. Similarly, we can understand the meaning of the σ_{ij} and the Kr_{ij}.

NLOS Identification Performance. In this subsection, we mainly analysis the RSS obtained in the corridor and laboratory. We have conducted extensive experiments to test the proposed method using the extracted factor from RSS. Figure 4 shows the result of NLOS identification. As we can see, the rate of NLOS identification can reach a higher level than the PhaseU [11] and LiFi [10], introduced in the Sect. 3. Due to the movement of people and the existence of others APs, the RSS got in the corridors changes a great deal. So the rate of LOS Identification is lower than the laboratory's LOS condition.

Distance Estimation Performance. We conduct the distance estimation for each AP over all reference points. It is seen that the curve fitting method with two variables can achieve more accurate distance estimations. In particular, the average distance estimation error for an AP in LOS condition is about 2 m by using the curve fitting, whereas it is more than 2.5 m by using curve fitting proposed in [19] and 5 m by using the propagation model method. While in the NLOS condition, the average distance estimation error for an AP is about

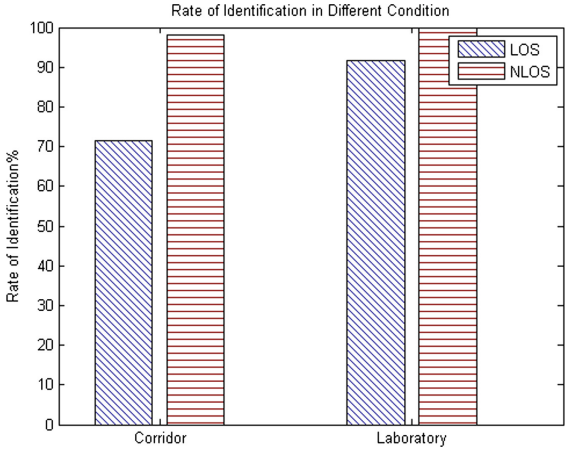


Fig. 4. Rate of identification

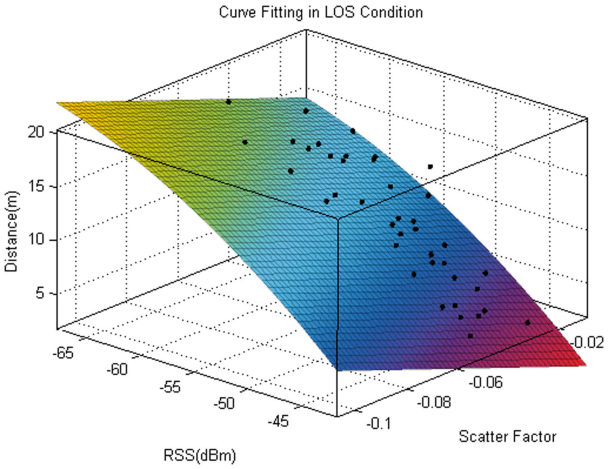


Fig. 5. Curve fitting result in LOS condition

0.79 m by using the proposed curve fitting, whereas it is more than 0.8 m by using curve fitting proposed in [19] and 4 m by using the propagation method. Figures 5 and 6 show the RSS fitting result in LOS and NLOS condition with two variables, RSS and scatter factor δ. We can see that the fitting surface can well demonstrate the relationship of the RSS-distance in LOS and NLOS condition, respectively. Figure 7 plots the cumulative distribution function (CDF) of the distance estimation error for one of APs. In this case, the proposed curve fitting method has a great improvement, about 30%, compared with the curve fitting method in terms of the distance estimation accuracy.

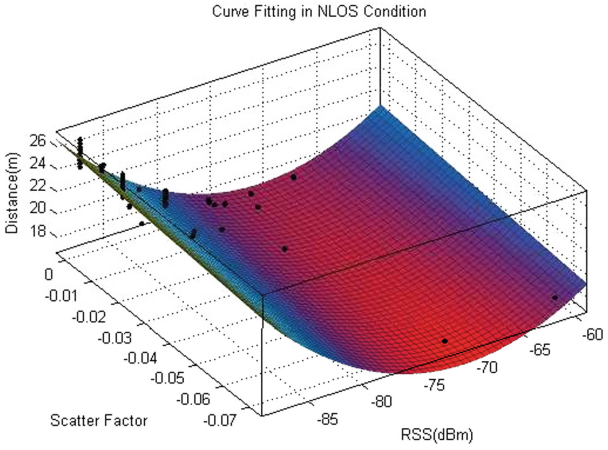


Fig. 6. Curve fitting result in NLOS condition

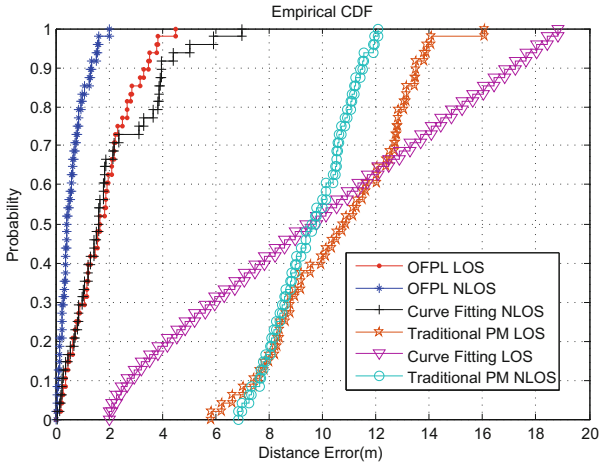


Fig. 7. The CDF of localization error comparison

Localization Performance. Fig. 8 plots the localization performance when different numbers of APs are used. We conduct extensive experiments in various indoor environments using the proposed OFPL algorithm and calculate the localization error with statistical methods. We can come up with the conclusion from the result that it is not always good for us to deploy many APs to complete the localization, because the more APs are used, the less difference among the reference points' fingerprints and the more interference among the APs' signals. So it will be difficult to localize the receiver. In addition, it can be observed that the minimum localization error is 1.97 m in our field experiments, and the average improvement of OFPL compared with the algorithm in [19] is about 35%.

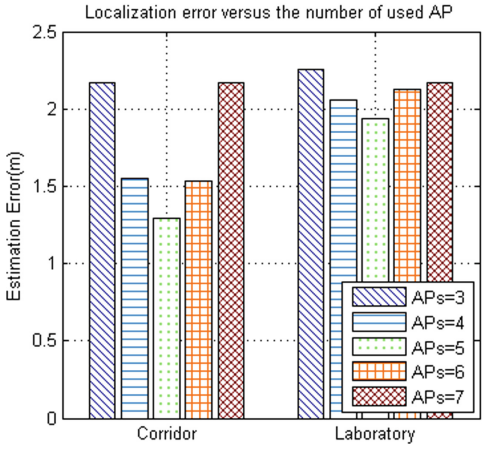


Fig. 8. The localization error versus the number of used AP

5 Conclusion

In this paper, we extract the scatter factor, which depicts the severity of multi-path effect, and then utilize the factor to complete a simple but efficient NLOS identification in various indoor environments. What's more, we use curve fitting technology to construct a fitted RSS-distance functions in LOS and NLOS conditions, respectively, which have been proved that they can improve the distance estimation accuracy very well. At last, we propose the OFPL algorithm and demonstrate that our algorithm can obtain some improvement in localization accuracy compared with the classical fingerprinting-based localization through various field experiments.

Acknowledgment. This work was supported by State Major Science and Technique Project (MJ-2014-S-37), National Natural Science Foundation of China (61201134) and 111 Project (B08038).

References

1. Zhou, M., Pu, Q., Xu, K., et al.: Location fingerprint discrimination maximization for indoor WLAN access point optimization using fast discrete water-filling. In: 2015 IEEE Global Communications Conference (GLOBECOM), 6–10 December 2015
2. Zhou, M., Zhang, Q., Tian, Z.: IMLours: indoor mapping and localization using time-stamped WLAN received signal strength. In: 2015 IEEE Wireless Communications and Networking Conference (WCNC), 9–12, March 2015
3. Xiao, J., Wu, K., Yi, Y., Ni, L.: FIFS: fine-grained indoor fingerprinting system. In: 21st International Conference on Computer Communications and Networks, pp. 1–7 (2012)
4. Kim, J., Ji, M., Jeon, J.-I., et al.: K-NN based positioning performance estimation for fingerprinting localization. In: 2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN), 5–8 July 2016
5. Wu, K., Xiao, J., Yi, Y., et al.: FILA: fine-grained indoor localization. In: IEEE International Conference on Computer Communications (2012)
6. Lohan, E.S., Talvitie, J., et al.: Received signal strength models for WLAN and BLE-based indoor positioning in multi-floor buildings. 2015 International Conference on Location and GNSS (ICL-GNSS), 22–24 June 2015, pp. 1–6 (2015)
7. Wang, K., Nirmalathas, A., et al.: Optical wireless-based indoor localization system employing a single-channel imaging receiver. J. Lightwave Technol. **34**(4), 1141–1149 (2015)
8. Kodippili, N.S., Dias, D.: Integration of fingerprinting and trilateration techniques for improved indoor localization. In: 2010 Seventh International Conference on Wireless and Optical Communications Networks (WOCN), 6–8 September 2010
9. Gaber, A., Omar, A.: Utilization of multiple-antenna multicarrier systems and NLOS mitigation for accurate wireless indoor positioning. IEEE Trans. Wirel. Commun. **15**, 6570–6584 (2016)
10. Zhou, Z., Yang, Z., Wu, C., et al.: LiFi: line-of-sight identification with WiFi. In: IEEE International Conference on Computer Communications (2014)

11. Wu, C., Yang, Z., Zhou, Z., et al.: PhaseU: real-time LOS identification with WiFi. In: IEEE International Conference on Computer Communications (2015)
12. Xiao, Z., Wen, H., Markham, A., et al.: Identification and mitigation of non-line-of-sight conditions using received signal strength. In: IEEE 9th International Conference on Wireless and Mobile Computing, Networking and Communications Lyon, 7–9 Oct 2013, pp. 667–674 (2013)
13. Lim, S.Y., Yun, Z., Iskander, M.F.: Propagation measurement and modeling for indoor stairwells at 2.4 and 5.8 GHz. IEEE Trans. Antennas Propag. **62**, 4754–4761 (2014)
14. Solahuddin, Y.F., Mardeni, R.: Indoor empirical path loss prediction model for 2.4 GHz 802.11n network. In: 2011 IEEE International Conference on Control System, Computing and Engineering (ICCSCE), 25–27 Nov 2011, pp. 12–17 (2011)
15. Liu, X., Cao, J., Tang, S., et al.: Wi-sleep: contactless sleep monitoring via WiFi signals. In: IEEE Real-Time Systems Symposium, pp. 346–355 (2014)
16. Sabek, I., Youssef, M., Vasilakos, A.V.: ACE: an accurate and efficient multi-entity device-free WLAN localization system. IEEE Trans. Mobile Comput. **14**, 261–273 (2015)
17. Zhang, J., Salmi, J., Lohan, E.-S.: Analysis of kurtosis-based LOS/NLOS identification using indoor MIMO channel measurement. IEEE Trans. Veh. Technol. **62**, 2871–2874 (2013)
18. Zhou, Z., Yang, Z., Wu, C., et al.: WiFi-based indoor line-of-sight identification. IEEE Trans. Wirel. Commun. **14**(11), 6125–6136 (2015)
19. Wang, B., Zhou, S., Liu, W., et al.: Localization based on curve fitting and location search using received signal strength. IEEE Trans. Ind. Electron. **62**, 572–578 (2014)



Fairness-Aware Scheduling for Millimeter Wave Backhaul Networks

Danfeng Meng[1], Xiaohui Li[1(✉)], Yanbin Zhao[2], Bin Zhou[3], and Dantao Li[1]

[1] Xidian University, Xi'an 710071, Shaanxi, China
xhli@mail.xidian.edu.cn
[2] Geographic Information Center, Yangquan, Shanxi Province, China
[3] AVIC Computing Technique Research Institute, Xi'an, Shaanxi Province, China

Abstract. Millimeter wave (mmWave) backhaul networks are emerging as a promising candidate for the 5G mobile network. By exploiting highly directional antennas, the concurrent transmission technology which can greatly improve the network throughput is enabled in the mmWave backhaul communication. However, the existing concurrent transmission scheduling scheme ignores fairness, leading to the situation that some links in good state are always dispatched while some in poor state suffering starvation. In this paper we propose a Fairness-aware Global Scheduling (FAGS) algorithm in which a utility function and a scheduling threshold are designed to maximize the system throughput with the fairness requirements satisfied. And to quantify the impact of the scheduling algorithm on fairness a Jain Index is proposed. Simulations conducted in the 60 GHz band demonstrate the superior performance of our algorithm compared with other existing schemes.

Keywords: 5G · Millimeter wave · Backhaul networks
Concurrent transmission · Fairness

1 Introduction

The next generation ubiquitous and ultra-high bandwidth communication system, known as 5G, is planned to support up to 1000-fold increase in wireless data traffic within the next 20 years [1]. The trend for dense deployment in future 5G mobile communication networks makes current wired backhaul infeasible owing to the high cost. Millimeter wave backhaul communication, a promising technique with the capability of providing a multi-gigabit transmission rate, offers a flexible and cost-effective candidate for 5G communication system.

However, one fundamental distinguishing feature of mmWave communications is the high propagation loss. A directional antenna with high directivity gains which allows more efficient concurrent transmissions is utilized to combat the severe path loss. Although concurrent transmission is an effective method to enlarge throughput in the backhaul network, serious unfairness may arise because some links get more opportunities to be concurrently transmitted than others.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 240–253, 2018.
https://doi.org/10.1007/978-3-319-78130-3_25

In a scenario where lots of available links are densely deployed, an effective and fairness-aware backhaul scheduling schemes needs to be designed.

The scheduling schemes of the backhaul network have been investigated in the literature [2–4]. The Exclusive Region (ER) based scheduling algorithm is proposed and derived in [5]. It makes sure that concurrent transmissions always outperform the serial TDMA (Time Division Multiple Access) by co-scheduling flows in the exclusive region. A concurrent transmission scheduling algorithm is introduced in [6] with rate adaptation, however, it does not consider the unique features of mmWave systems, such as high propagation loss and the use of directional antenna. In [7] multiple communication links are scheduled in the same time slot if the interference in this slot is below a specific threshold. In [8] the STDMA (Space/Time Division Multiple Access) algorithm takes full use of the concurrent transmission to achieve as much throughput as possible, but the fairness is ignored. To the best of our knowledge, none of the previous works are devoted to address the balance between the fairness requirement and the throughput between links in the mmWave backhaul network. And there are two aspects of challenges for the scheduling problem. In the first aspect, concurrent transmissions need to be fully exploited to maximize the system throughput. In the second aspect, the scheduling scheme should guarantee the fairness for links in the backhaul network.

In this paper we propose a heuristic algorithm for the mmWave backhaul network to achieve a good compromise between throughput and fairness. The main contribution of the paper is three-fold. First, we formulate the problem of optimal scheduling in the mmWave backhaul network as a nonlinear integer programming. Concurrent transmission is fully taken into account in this problem. Second, a heuristic scheduling algorithm, namely, the fairness-aware global scheduling (FAGS) is proposed to solve the formulated problem with low complexity. In the algorithm a utility function is defined to select the suitable links and a scheduling threshold is designed to eliminating the unfairness between links. Besides, a Jain Index is proposed to indicate the fairness level. Finally, simulation results of the FAGS algorithm in the 60 GHz band are presented in the end of the paper.

The rest of this paper is organized as follows. Section 2 gives an overview of the system model and assumption. Section 3 presents the problem formulation and analysis. The FAGS algorithm is described in Sect. 4. Section 5 gives the simulation and analysis of our scheduling scheme, followed by concluding remarks in Sect. 6.

2 System Model and Assumption

2.1 Single-Hop Link Scheduling

A typical mmWave backhaul network is shown in Fig. 1, where the small cells BSs are densely disposed. Mobile users are related with base stations (BSs) of the small cells, and BSs are connected by mesh backhaul network in the mmWave band. There are one or more BSs connected to the core network via

the microcells, which are called gateways. Besides, there is a backhaul network controller (BNC) residing on one of the gateways.

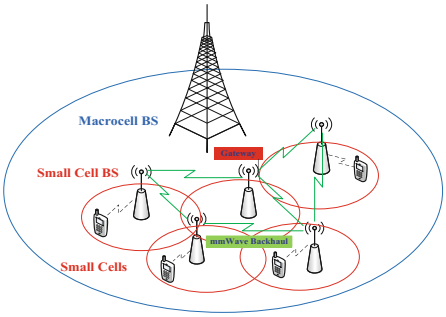


Fig. 1. The mmWave backhaul network for the small cells densely deployed

In the backhaul network, an end-to-end flow may go through multiple hops with a proper routing protocol in place. Once the routing path is fixed, all the single-hop links along the path will share the same constraint as the end-to-end flow. For this reason, the scheduling can only concentrate on the single-hop links, and the designing of routing protocol will be left for our future work. In this paper, considering there is a proper routing strategy in place, the only problem to be solved is the single-hop link scheduling scheme.

2.2 Frame Architecture

In the backhaul network, time is partitioned into superframes [9], each of which is further divided into a beacon period (BP) and a channel time allocation period (CTAP), as shown in Fig. 2. The beacon period is controlled by the backhaul network controller to provide timing and global information, meanwhile broadcasting scheduling decision for the CTAP. The channel time allocation period is used for data transmission among BSs in a peer-to-peer fashion. There are at most M channel time slots in CTAP of each superframe. The BNC can arrange the length of CTAP adaptively according to the total occupied number of time slots if it does not outpace M.

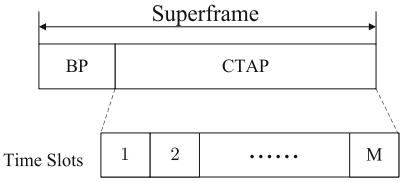


Fig. 2. The superframe structure for mmWave backhaul network

2.3 Physical Model

Since non-line-of-sight (NLOS) transmissions in 60 GHz channels suffer from significant attenuation and a shortage of multipath [10], mmWave backhaul networks mainly rely on line-of-sight (LOS) propagation to achieve high data rate. It's assumed that there are N flows requesting transmission time slots in the superframe, and each flow represents one backhaul link. For flow i, its sender and receiver is denoted by s_i and r_i, respectively. And the distance between the sender s_i and the receiver r_j is l_{ij}. The antenna gain of s_i in the direction of $s_i \rightarrow r_j$ is indicated by $A_t(i,j)$, in the same way, the antenna gain of r_i in the direction of $s_j \rightarrow r_i$ by $A_r(j,i)$. Then considering the path loss and signal dispersion over distance, the received power at r_i from s_i can be calculated as [11]

$$P_r(i,i) = k_0 A_t(i,i) A_r(i,i) l_{ii}^{-n} P_t \tag{1}$$

where k_0 is a constant coefficient and proportional to $(\frac{\lambda}{4\pi})^2$(λ denotes the wavelength), n denotes the path loss exponent, and P_t denotes the transmission power. Considering the simultaneous transmission of multiple links in the system, the received interference at r_i from s_j can be calculated as

$$P_r(j,i) = \rho k_0 A_t(j,i) A_r(j,i) l_{ji}^{-n} P_t \tag{2}$$

where ρ is the multi-user interference (MUI) factor related to the cross correlation of signals from different flows [12]. According to the Shannon's channel capacity, the achievable data rate of flow i can be estimated as

$$R_i = \eta W log_2(1 + \frac{P_r(i,i)}{N_0 W + \sum\limits_{j \neq i} P_r(j,i)}) \tag{3}$$

where $\eta \in (0,1)$ describes the efficiency of the transceiver design, W is the bandwidth, and N_0 is the one sided power spectral density of white Gaussian noise.

3 Problem Formulation and Analysis

It's assumed that there are N flows in the M time slots. For the time slot $m \in \{1,\ldots,M\}$, $\mathbf{R}^m = \{R_1^m,\ldots,R_N^m\}$ denotes the transmission rate vector, and $\boldsymbol{\alpha}^m = \{\alpha_1^m,\ldots,\alpha_n^m,\ldots,\alpha_N^m\} \in \{0,1\}^N$ indicates the link status ("1" means that the link is active, and "0" means that the link keeps silent). Thus, the optimization problem can be formulated to maximize the system throughput.

$$\begin{aligned} &\underset{\{\alpha_1^m,\ldots\alpha_n^m,\ldots\alpha_N^m\}}{\text{Maximize}} \sum_{m=1}^{M}\sum_{n=1}^{N} R_n^m \cdot t_{slot}/(T_{BP} + M \cdot t_{slot}) \\ &\text{where } R_n^m = \alpha_n^m \cdot \eta W log_2(1 + \frac{P_r(n,n)}{N_0 W + \sum\limits_{j \neq n} \alpha_j^m P_r(j,n)}) \end{aligned} \tag{4}$$

R_n^m represents the transmission rate of link n in the time slot m, t_{slot} is the time duration of each time slot for CTAP period, and T_{BP} is the time duration for beacon period.

The complexity of solving this optimization problem is extremely high. In addition, at the optimal point, only a few links achieve high throughput while others will be starved to death. Therefore, a fairness-aware suboptimal solution algorithm is raised to realize the compromise between link fairness and system throughput, beginning with a *utility function* definition.

3.1 Utility Function

We assume Ω is the set of links that are being scheduled at the present time. A utility function is put forward to determine whether link i can join the current set. When flow i is chosen into the set, the gain obtained by flow i is expressed as its transmission rate, calculated by

$$G_i = \eta W log_2(1 + \frac{P_r(i,i)}{N_0 W + \sum\limits_{j \in \Omega} P_r(j,i)}) \tag{5}$$

At the same time, the cost (due to the increased interference from flow i to flow $k \in \Omega$) is illustrated as the reduced transmission rate for flow k, given by

$$\begin{aligned} G_{ik} = &\eta W log_2(1 + \frac{P_r(k,k)}{N_0 W + \sum\limits_{j \in \Omega, j \neq k} P_r(j,k)}) \\ &- \eta W log_2(1 + \frac{P_r(k,k)}{N_0 W + \sum\limits_{j \in \Omega, j \neq k} P_r(j,k) + P_r(i,k)}) \end{aligned} \tag{6}$$

Thus, a *utility function* is designed to represent the "system gain" for flow i:

$$U_i = G_i - \sum_{k \in \Omega} G_{ik} \tag{7}$$

If $U_i > 0$, that means flow i will bring a positive transmission rate gain to the network and can be scheduled in the present time slot; otherwise, it should be silent. For each scheduling time slot, the link with the positive value of utility function will be chosen. In this way, the system throughput can be the maximal but it is easy to cause unfairness among the links. Thus, a fairness weight factor is presented to ensure fairness among the links.

3.2 Fairness Weight Factor

For the mmWave backhaul network, it's very challenging to achieve fairness. First, the notion of fairness is quite different from that in traditional networks, where fairness can be defined for a specific link with a fixed capacity.

In a mmWave backhaul network, the links interfere with each other. The direct or indirect interfering relationship among links determines that the fairness in mmWave backhaul network should have a global definition instead of being limited to a specific link. Second, spatial reuse may conflict with fairness. Hence, a feasible tradeoff between throughput and fairness should be considered. The well-known proportional fair scheduling [13,14] obtains a good compromise between throughput and fairness for code-division multiple access (CDMA) cellular systems. For each time slot, the scheduler schedules the link with the highest priority value defined as the ratio of the signal-to-interference-plus-noise ratio (SINR) to the average throughput. If a link has a low average throughput, its chance to be selected to transmit is relatively large, so as to achieve a certain level of fairness.

Based on the proportional fair scheduling, a fairness weight factor is introduced for mmWave backhaul network to ensure the fairness of scheduling. The basic idea is that links in good condition[1] are given priority while those in bad condition can get an acceptable scheduling share. Specifically, the fairness weight factor is written as

$$\omega_i(t) = \frac{\mu_i}{T_i(t-1)+\gamma}, t = 1, \ldots, M \tag{8}$$

where μ_i is the pre-specified service priority factor of link i, $T_i(t-1)$ represents the throughput of the link i obtained from the first $t-1$ time slots and γ is a constant to balance the throughput. $T_i(t-1)$ is written as

$$T_i(t-1) = \frac{\sum_{m=1}^{t-1} R_i^m \cdot t_{slot}}{T_{BP} + (t-1) \cdot t_{slot}} \tag{9}$$

Specially, $T_i(0) = 0 (t = 1)$ meaning that the throughput of each link is zero before the first time slot. Then the utility function is modified based on the fairness weight factor. The modified *utility function* is

$$U_i^t = \omega_i(t) G_i - \sum_{k \in \Omega} \omega_i(t) G_{ik}, t = 1, \ldots, M \tag{10}$$

The modified utility function gives the link selection criteria at the time slot t directly. From Eq. (8) and (10) it can be seen that if the link state of link i is better, then the number of times it may be scheduled in a short time is relatively large. So link i can get relatively high throughput. However, if the throughput is too large, the value of the utility function will be reduced under the fairness weight factor. Thus, each flow can get a fairness-aware scheduling.

3.3 Jain Index

For mmWave backhaul networks, a fair resource allocation is not necessarily the case when each link receives the same throughput and service level. Consider a

[1] In this paper, a link is said in good (bad) link condition if it is in an advantageous (disadvantageous) position, such as with a short (long) link distance and/or in a low (high)-interference neighborhood.

network with 3 links, where links 1 and 2 are close to each other while link 3 is far away from them. Links 1 and 2 generate large interference to each other. Hence, it's not appropriate to allow them transmitting simultaneously i.e. the probability that link 1 and link 2 are scheduled at the current time is 50%. As there is no interference link in link 3's neighborhood, it is better to allow link 3 to be active at the present time. This simple example shows that, to evaluate fairness, the link condition should be taken into account. However, it is challenging to evaluate the link condition in the mmWave backhaul network. It should be determined by its path loss and the interference level. Here we use a heuristic approach. To begin with, a function that measures the link condition is given as

$$\nu_i = \frac{1}{1 + S(I_i)} \tag{11}$$

where $S(I_i)$ is the size of relative interference set for link i. And the relative interference set of link i is expressed as RI_i.

$$RI_i = \{j | j \neq i, \max\{\frac{P_r(j,i)}{P_r(i,i)}, \frac{P_r(i,j)}{P_r(j,j)}\} > \sigma\} \tag{12}$$

where σ is the interference threshold to measure relative interference between link i and j. From (11) and (12) it can be seen that if link i is less interfered by other links, a larger value for ν_i can be obtained,otherwise, the value is small. In this way, the link condition can be judged by the value of ν_i.

On the performance analysis of the scheduling algorithm in Sect. 5, fairness is a very important indicator. The impact of scheduling strategy on the system fairness needs to be quantified. Based on the link condition function, Jain Index is proposed for evaluating the fairness and is defined as follows

$$\text{Jain Index} = \frac{(\sum_{i=1}^{N} \frac{T_{i,total}}{\mu_i \nu_i})^2}{N \cdot \sum_{i=1}^{N} (\frac{T_{i,total}}{\mu_i \nu_i})^2} \tag{13}$$

where $T_{i,total}$ represents the total throughput achieved by link i during the whole scheduling time. And we have

$$T_{i,total} = \frac{\sum_{m=1}^{M} R_i^m \cdot t_{slot}}{T_{BP} + M \cdot t_{slot}} \tag{14}$$

In the definition of Jain Index, the greater of the Jain Index value is, the more fairness of the allocation scheme achieves. If Jain Index = 1, the scheduling scheme is perfectly fair, that is, each component is exactly equal.

4 Scheduling Algorithm Design

In this section, we propose the fairness-aware global scheduling (FAGS) algorithm. The algorithm combines the utility function defined in Sect. 3 and a scheduling threshold given in the following to put forward a suitable scheme.

4.1 Scheduling Threshold

In order to further ensure the fairness among links, a scheduling threshold for the scheduled times of each link is proposed as follows

$$TH_{schedule} = \imath \cdot M \tag{15}$$

where $\imath$ is the scheduling threshold factor, namely, a constant coefficient to adjust the value of $TH_{schedule}$ which represents the maximum scheduled times for each link. The threshold value $TH_{schedule}$ is used to assure link i will not be selected too frequently. We use $Ca(i)$ to record the number of times that link i was scheduled. If $Ca(i) > TH_{schedule}$, then link i can not be scheduled anymore. In addition, since $\imath$ determines the value of $TH_{schedule}$, the choice of the scheduling threshold factor $\imath$ has an important effect on system performance, which will be demonstrated in the simulation section.

4.2 Fairness-Aware Global Scheduling (FAGS) Algorithm

Combining the utility function and the scheduling threshold, we bring forward the FAGS algorithm. The scheduling algorithm starts in the first time slot and uses the utility function to select as many links as possible until all the links that meet the scheduling threshold conditions are traversed. And then the next time slot begins scheduling until the M time slots are completed. The decomposed method can solve the problem in a manner by reducing the searching space from $O(M^N)$ to $O(N \cdot M)$. For detail, the scheduling algorithm in each time slot can be concluded in three steps.

Step 1: Remove the links that do not meet the scheduling threshold. For link i, the number of scheduled times must be less than the scheduling threshold, namely, $Ca(i) < TH_{schedule}$. If this condition is not satisfied, the link can't join into the current transmission set in the present time slot.

Step 2: Select the first scheduled link. The first link should not only have the largest value of utility function but also have never been scheduled in the past slots. This way of choice makes sure that the link in bad condition avoids suffering starvation and each link is properly scheduled.

Step 3: Chose the other links to take part in the present scheduling set in a greedy way. For each time we select the link with the largest value of utility function from the remaining links set and remove it from the set. The scheduling does not end until the utility function of the selected link is negative or the remaining links set is empty, then turning to the next time slot.

The pseudo-code of the fairness-aware global scheduling (FAGS) algorithm is presented in Algorithm 1. The initialization is completed by lines 1–3. The scheduling matrix $\mathbf{B}$ indicates whether the link i is scheduled in the m-th time slot and it's initial value is $\mathbf{0}$; The decision vector $\boldsymbol{\alpha}^m_{\text{FAGS}}$ indicates which flows are scheduled in the k-th slot. A is the set of all links; A^* is the set of links that meeting the scheduling threshold and is initialized to empty; Ω is the set of

Algorithm 1. Fairness-Aware Global Schedule

1.**Initialize** $N * M$ scheduling matrix $\mathbf{B} = \mathbf{0}$
2.**Initialize** scheduling vector $\boldsymbol{\alpha}^m_{\mathrm{FAGS}} = [\alpha^m_1, \ldots \alpha^m_l, \ldots \alpha^m_N] = \mathbf{0}$
3.**Initialize** $A = \{l = 1, \ldots, N\}, A^* = \Phi, \Omega = \{l = 1, \ldots, N\}$
4.**For** slot $m(1 \leq m \leq M)$
5. $A^* = A \backslash X_{l \in A}\{l : Ca(l) > TH_{schedule}\}$
6. **If**$(\Omega \neq \Phi)$
7. Compute the *Utility Function* U^m_l of flows in Ω
8. $l^* = \mathrm{argmax}_{l \in \Omega} U^m_l$, $\alpha^m_{l^*} = 1$
9. $Ca(l^*) + +, \Omega = \Omega \backslash \{l^*\}, A^* = A^* \backslash \{l^*\}$
10. **Else**
11. Turn to 13
12. **End if**
13. **While** $(A^* \neq \Phi)$
14. Compute *Utility Function* U^m_l for flows in A^*
15. $l^* = \mathrm{argmax}_{l \in A^*} U^m_l$
16. **If** $U^m_{l^*} \leq 0$
17. Turn to line 21
18. **Else**
19. $\alpha^m_{l^*} = 1, A^* = A^* \backslash \{l^*\}, Ca(l^*) + +, \Omega = \Omega \backslash \{l^*\}$
20. **End if**
21. **End while**
22. $\mathbf{B}(:, m) = {\boldsymbol{\alpha}^m_{\mathrm{FAGS}}}^T$
23.**End for**
24.**Return B**

links that hasn't been scheduled and is initialized to all links. Line 4 ensures that the number of scheduling times can't exceed the scheduling threshold. The first scheduled link is chosen by line 5–11. In line 12–20, we use the utility function to select as many links joining the current transmission set, in order to get a larger throughput. Line 21–23 updates the scheduling vector $\boldsymbol{\alpha}^m_{\mathrm{FAGS}}$ and scheduling matrix $\mathbf{B}$, then the scheduling is completed.

5 Performance Evaluation

In this section, we evaluate the performance of the proposed scheduling algorithm in the 60 GHz band, and investigate the impact of the scheduling threshold factor on the performance of our scheme. Besides, we also compare the proposed scheduling algorithm with the existing schemes in terms of Jain Index and system throughput.

5.1 Simulation Setup

The simulation is conducted in a backhaul network with 10 base stations which has at most 90 flows. For the scheduling performance is dependent on the location of stations, the position for each BS is randomly generated within a square area of $100 \times 100\,\mathrm{m}^2$. Meanwhile, for every flow its source and destination is randomly chosen. And we use the channel model in Ref. [15] for the path loss. Besides, we adopt the widely used realistic directional antenna model in Ref. [16]. All the BSs in the system use the same transmission power level. Some other parameters are shown in Table 1.

Table 1. Simulation parameters

Symbol	Value	Symbol	Value
P_t	0.5 mW	N_0	$4 * 10^{-20}$ W/Hz
W	2160 MHz	σ	10^{-4}
μ_i	1	n	-2
M	500	η	0.8
ρ	0.01	γ	10^8
t_{slot}	18 us	T_{BP}	850 us

5.2 Choice of the Scheduling Threshold Factor

Since the choice of the scheduling threshold factor has an important impact on the performance of our scheme, we now investigate it under different typical number of flows, that is, $N = 10, N = 30, N = 50, N = 70$.

The system throughput of our scheme under different values of the scheduling threshold factor is displayed in Fig. 3. From the figure we can see that with the increase of the scheduling threshold factor, the system throughput increases significantly. In the cases $N = 10, N = 30, N = 50, N = 70$, the system throughputs are increased by 27%, 34%, 47%, 44% when τ changes from to 0.1 to 1. Owing to the improving of τ, the value of $TH_{schedule}$ increases proportionally and the number of times for each link being scheduled are gradually increased. Thereby, the system throughput increases significantly. However, when the threshold is large enough, the main factor limiting the system throughput is the interference between the links. Thus, when τ is greater than 0.5, the system throughput increases slowly.

In Fig. 4, we plot the Jain Index of our scheme under different values of the scheduling threshold factor. It can be seen that with the increase of the scheduling threshold factor, the value of Jain Index decreases. When the scheduling threshold factor is small, some of the links in good condition can't be scheduled

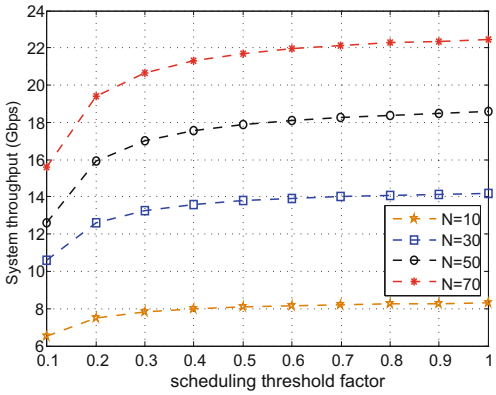


Fig. 3. The system throughput under different scheduling threshold factors

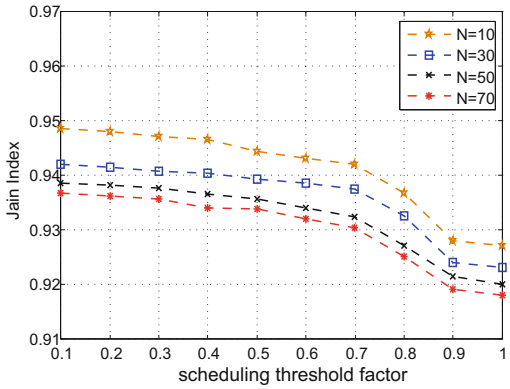


Fig. 4. The Jain Index under different scheduling threshold factors

too frequency than other links, thus ensuring the fairness of the system; otherwise, the fairness reduces. In addition, when τ is in the range from 0.7 to 0.9, the slope of the curve becomes larger, significantly reducing the Jain Index.

Choosing the appropriate scheduling threshold factor is important for both the system throughput and fairness. According to the simulation results, when the threshold factor is controlled between [0.5–0.7], the fairness of the system can be guaranteed without losing the system throughput. And we will set $\tau = 0.6$ in the following simulation.

5.3 Comparison with Other Schemes

In this case, we vary the number of flows in the backhaul network from 10 to 90. We adopted the serial TDMA scheme, and the MQIS scheme [11] for comparison. Under the increasing number of demanding flows, the system throughput and

Jain Index of the proposed scheme are compared with the other two schemes and the results are shown in Figs. 5 and 6 correspondingly.

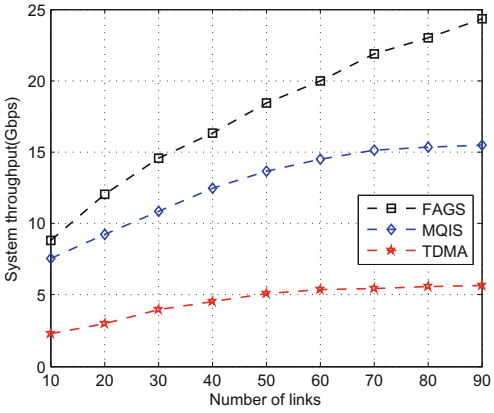


Fig. 5. The system throughput under different number of links

It can be seen from Fig. 5 that the system throughput increases when the number of links becomes larger. And compared with the MQIS and TDMA algorithms, the FAGS algorithm has obvious advantages. When $N = 90$, the system throughput of FAGS algorithm is 3.28 times higher than that of TDMA algorithm and 57% higher than that of MQIS algorithm. Since the TDMA scheme doesn't use the spatial multiplexing technology, the throughput curve is relatively gentle. The MQIS algorithm and the FAGS algorithm all utilize the spatial multiplexing technique, but after $N = 70$, the MQIS algorithm is no longer increased, and the latter continues to increase. In the MQIS algorithm, only those links in good state have completed the scheduling, the remaining links can begin to be scheduled, which resulting in some links can't be timely scheduled. In this paper, the proposed FAGS algorithm is scheduled in each time slot, which ensuring as many links as possible obtaining timely scheduling. Thus, compared with the MQIS algorithm, the system throughput of our proposed FAGS algorithm increases greatly.

Figure 6 compares the Jain Index of the three algorithms under different number of links. Because of the absolute fairness, the Jain Index of the TDMA algorithm is always 1. The Jain Index of the FAGS algorithm decreases with the increase of N but gradually stabilized. And the Jain Index is always above 0.9. The MQIS algorithm decreases sharply with the increase of N, and the fairness is the worst. With the increase of interference between links, there are more links in the poor condition. The MQIS algorithm prefers the links with good condition, resulting in lower fairness. The FAGS algorithm makes use of the fairness weight factor to ensure that the links in different states being scheduled fairly. Thereby, the fairness is improved compared with the MQIS algorithm.

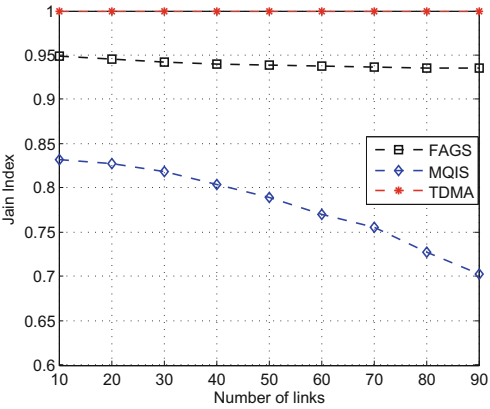


Fig. 6. The Jain Index under different number of links

To summarize, the TDMA algorithm can guarantee the absolute fairness, but its system throughput is too small; the MQIS algorithm has high system throughput but the fairness is too low. The FAGS algorithm achieves the balance between throughput and fairness.

6 Conclusions

In this paper, we investigate the problem of maximizing the system throughput with the fairness requirements satisfied for the mmWave backhaul network of small cells densely deployed. We propose the Fairness-aware Global Scheduling (FAGS) algorithm in which a utility function and a scheduling threshold are designed for improving the system performance. To quantify the impact of scheduling algorithm on the fairness Jain Index is proposed. And the choice of the scheduling threshold factor is also analyzed to provide guidelines for parameter selection in practice. Extensive simulations demonstrate the superior performance of our algorithm in system throughput and fairness compared with the TDMA algorithm and the MQIS algorithm.

Acknowledgment. This work was supported by the State Major Science and Technique Project (MJ-2014-S-37), the National Natural Science Foundation of China (61201134), and the 111 Project (B08038).

References

1. Seppanen, K., Kilpi, J., Paananen, J., Suihko, T., Wainio, P., Kapanen, J.: Multipath routing for mmWave WMN backhaul. In: 2016 IEEE International Conference on Communications Workshops (ICC), pp. 246–253 (2016)
2. Baldemair, R., Irnich, T., Balachandran, K., Dahlman, E., Mildh, G., Seln, Y., Parkvall, S., Meyer, M., Osseiran, A.: Ultra-dense networks in millimeter-wave frequencies. IEEE Commun. Mag. **53**(1), 202–208 (2015)

3. Taori, R., Sridharan, A.: Point-to-multipoint in-band mmWave backhaul for 5G networks. Commun. Mag. **53**(1), 195–201 (2015)
4. Niu, Y., et al.: Exploiting device-to-device communications in joint scheduling of access and backhaul for mmWave small cells
5. Cai, L.X., Cai, L., Shen, X., Mark, J.W.: REX: a randomized exclusive region based scheduling scheme for mmWave WPANs with directional antenna. IEEE Trans. Wirel. Commun. **9**(1), 113–121 (2010)
6. Yang, Z., Cai, L., Lu, W.: Practical concurrent transmission scheduling algorithms for rate-adaptive wireless networks. In: Proceedings of IEEE INFOCOM (2010)
7. Sum, C.S., Lan, Z., Rahman, M.A., et. al.: A multi-GBPS millimeter wave WPAN system based on STDMA with heuristic scheduling. In: Proceedings of IEEE GLOBECOM, June 2009
8. Qiao, J., Cai, L.X., Shen, X., Mark, J.W.: STDMA-based scheduling algorithm for concurrent transmissions in directional millimeter wave networks. In: 2012 IEEE International Conference on Communications (ICC), pp. 5221–5225 (2012)
9. Qiao, J., Cai, L.X., Shen, X.: Multi-hop concurrent transmission in millimeter wave WPANs with directional antenna. In: 2010 IEEE International Conference on Communications, pp. 1–5 (2010)
10. Guo, N., Qiu, R.C., Mo, S.S., Takahashi, K.: 60-GHz millimeter wave radio: principle, technology, and new results. EURASIP J. Wirel. Commun. Netw. 8 p. (2006). Article ID. 68253
11. Zhu, Y., Niu, Y., Li, J., Wu, D.O., Li, Y., Jin, D.: QoS-aware scheduling for small cell millimeter wave mesh backhaul. In: 2016 IEEE International Conference on Communications (ICC), pp. 1–6 (2016)
12. Niu, Y., Gao, C., Li, Y., Su, L., Jin, D., Zhu, Y., Wu, D.O.: Energy efficient scheduling for mmWave backhauling of small cells in heterogeneous cellular networks. IEEE Trans. Veh. Technol. **PP**(99), 1 (2016)
13. Holtzman, J.M.: Asymptotic analysis of proportional fair algorithm. In: Proceedings of IEEE PIMRC 2001, pp. F-33–F-37, (2001)
14. Jalali, A., Padovani, R., Pankaj, R.: Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system. In: Proceedings of IEEE VTC 2000, pp. 1854–1858 (2000)
15. Toyoda, I., Seki, T., Iiguse, K.: Reference antenna model with side lobe for TG3c evaluation. IEEE document, IEEE 802.15-06/474r0 (2006)
16. Rezaei, M.J., Sabahi, M.F., Shahtalebi, K., Zaeem, R.M., Sadeghi, R.: A new fairness index and novel approach for QoS-aware resource allocation in LTE networks based on utility functions. In: 2015 Signal Processing and Intelligent Systems Conference (SPIS), Tehran, pp. 124–127 (2015)



Lyapunov Optimization for NOMA Wireless Transmission with Energy Harvesting and Storage

Qinbo Chen(✉), Qianyun Gong, Maoxin Tian, Hui Cai, and Xiancai Chen

School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, Guangdong, China
{chenqb6,gongqy3,tianmx,caih8,chenxc6}@mail2.sysu.edu.cn

Abstract. In this letter, we study optimal power allocation for a Non-Orthogonal Multiple Access (NOMA) communication system, where the transmitter transmits superimposed information symbols to two users over fading channel with energy harvesting technology. Our objective is to maximize the long-term time-averaged transmission rate subject to the power constraint of the transmitter. By exploiting Lyapunov optimization scheme, we equivalently transform the non-convex optimization problem into a single time slot optimal problem and solve it by alternative optimization scheme with the help of convex optimization. At the end, simulation results have shown that the proposed optimal scheme outperforms other two compared algorithms and can achieve Lyapunov stability.

Keywords: Energy harvesting (EH) · Power allocation
Lyapunov optimization · Non-Orthogonal Multiple Access (NOMA)
Throughput maximization · Convex optimization

1 Introduction

The energy harvesting (EH) techniques [1], which allow battery operated transmitter to harvest energy from solar radiation or artificial radio signals, can solve energy scarcity problem and provide other sources of energy for electronic devices. Non-Orthogonal Multiple Access (NOMA) [2] technology can achieve a higher user rate compared to traditional Orthogonal Multiple Access (OMA) Scheme. NOMA system lets transmitter transmit the superposition of multiple user signals to multiple destinations. By enforcing Successive Interference Cancellatio (SIC) orderly, every user may decode its own signal that depends on channel condition. Lyapunov optimization is an effective way to solve stochastic and technically difficult problems while it can stabilize the battery status.

However, the aforementioned works use Lyapunov optimization for destination-to-destination communication based on conventional OMA scheme [3], which is not only very sophisticated and also difficult to enforce in reality

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 254–263, 2018.
https://doi.org/10.1007/978-3-319-78130-3_26

[4]. Along with the development of 5G, there are many people want to apply NOMA scheme to every aspect of communication. NOMA based on Lyapunov optimization is a promising one. To the best of our knowledge, the research on the MIMO-NOMA communication with energy harvesting is missing.

In this letter, our objective is to maximize the long-term time-averaged transmission rate subject to the transmitter power constraint based on a two user coma system. By exploiting Lyapunov optimization scheme, we equivalently replace the non-convex optimization problem with a single time slot optimal problem and solve it by alternative optimization scheme with the help of convex optimization. At the end, simulation results have shown that the proposed optimal scheme outperforms other two compared algorithms and can achieve Lyapunov stability.

2 System Model

As shown in Fig. 1, we consider a NOMA communication system consisting of one transmitter, two destinations. All devices are single antenna. Transmitter is equipped with energy harvesting and storage devices and harvests energy from environment in its vicinity (e.g., solar, radio wave). The system operates in discrete slotted time $t = \{0, 1, 2, ...\}$ with duration Δt, and all operations are performed per time slot. We define the destination D1 as strong user and D2 as weak user based on NOMA scheme, so the communication channel between transmitter and D1, $h_1(t)$, is stronger than the communication channel between transmitter and D2, $h_2(t)$. We assume a slow block fading scenario, where two channels, are assumed to be constant during slot t and change over time slot.

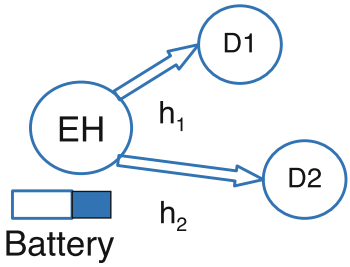


Fig. 1. System model

For the transmitter, we use $E(t)$ to denote the amount of energy battery harvests during time slot t. Let $E_s(t)$ denote the energy state of battery (SOB) at the beginning of time slot t. We constrain it by

$$0 \leq E_s(t) \leq E_{\max}, \ \forall t \tag{1}$$

where $E_{\max}$ means the maximum capacity of battery that can be used, different battery varies.

$P(t)$ donates the transmit power used by transmitter at time slot t, and transmitter determines it at the beginning of time slot t and it is unchanged until the end of time slot t. It is also constrained by

$$0 \leq P(t) \leq P_{\max}, \ \forall t. \tag{2}$$

where $\Delta t P_{\max} \leq E_{\max}$, due to energy causal constrain. We also assume battery exists maximum charging amount within every slot called E_{cm}, so we define $E_{cm} \leq \Delta t P_{\max}$. This assumption is satisfied under today's real circumstance.

Considering energy causal constrain, the dynamics of $E_s(t)$ should satisfy

$$E_s(t+1) = E_s(t) - \Delta t P(t) + E(t) \tag{3}$$

By combining all constrain listed above, we can also get

$$\Delta t P(t) \leq E_s(t), \ \forall t. \tag{4}$$

and the harvested energy within every time slot also should be limited by

$$E(t) = \min\{E_{\max} - (E_s(t) - \Delta t P(t)), E_{cm}\}. \tag{5}$$

the first formula at the right of (5) means available battery capacity. In this letter, we assume the charging efficiency and discharging efficiency equal to 1, so we don't consider energy loss.

For two destinations, in a given time slot t, the signal x transmitted by the transmitter based on NOMA scheme can be presented through superimposed information symbols

$$x(t) = \sqrt{\alpha P(t)} x_1(t) + \sqrt{(1-\alpha) P(t)} x_2(t) \tag{6}$$

where $x_1(t)$ and $x_2(t)$ denote information symbols destined for D1 and D2, respectively. The expectation $E\left[x_1^2(t)\right] = 1$ and $E\left[x_2^2(t)\right] = 1$ and $0 \leq \alpha \leq 1$ means the power allocation coefficient.

The received signal at destination $i, i = \{1, 2\}$ is

$$y_i(t) = h_i(t)\sqrt{\alpha P(t)} x_1(t) + h_i(t)\sqrt{(1-\alpha) P(t)} x_2(t) + n_i \tag{7}$$

where receiver noise n_i is additive white Gaussian noise with zero mean and variance σ^2 and transmitter totally knows $h_1(t)$ and $h_2(t)$ at each time slot t.

The instantaneous signal-to-interference-and-noise ratio (SINR) of the transmitter-to-D1 is

$$\gamma_1(t) = \frac{\alpha P(t) |h_1(t)|^2}{\sigma^2} \tag{8}$$

The another SINR is

$$\gamma_2(t) = \frac{(1-\alpha) P(t) |h_2(t)|^2}{\sigma^2 + \alpha P(t) |h_2(t)|^2} \tag{9}$$

Therefore the received rate for D1 is $R_1(t) = \frac{1}{2}\log\{1 + \gamma_1(t)\}$ and the received rate for D2 is $R_2(t) = \frac{1}{2}\log\{1 + \gamma_2(t)\}$. The sum rate $R(t) = \omega_1 R_1(t) + \omega_2 R_2(t)$($\omega_1$ and ω_2 are parameters used for adjusting the weight between two users) [5].

3 Problem Formulation

By using $h_1(t)$, $h_2(t)$ and $E(t)$, we want to determine the optimal parameter α and $P(t)$ at each time slot t. The ultimate goal of this letter is to design an algorithm to maximize the long-term time-averaged expected rate $E[R(t)]$, while satisfying all constrains listed above. We do not consider duplicate constrains, so this optimization problem is

$$\mathbf{P1}: \quad \max_{\{P(t),\alpha\}} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[R(t)]$$
$$\text{subject to} \quad (2), (3), (4)$$

where expectation is taken with respect to $E(t)$, $\gamma_1(t)$, and $\gamma_2(t)$, battery operational constrain (1) will be met at the end of solution.

Obviously, this problem is not a convex problem, and we intend to use Lyapunov optimization to design a solution to $P1$ based on the statistical knowledge. Based on the discipline of Lyapunov, we replace the per time slot constrain with a long-term time-averaged relation among these variables.

We define $\bar{E} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}(t)$ and $\bar{P} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} P(t)$. Based on the constrain (3), we can get

$$\mathbb{E}[E_s(T)] - \mathbb{E}[E_s(0)] = \sum_{t=0}^{T-1} \mathbb{E}[E(t) - \Delta t P(t)]. \tag{10}$$

Because of constrain (1), the left side is bounded. Dividing both side of (10) by T, and take $T \to \infty$, we can have

$$\bar{E} - \Delta t \bar{P} = 0. \tag{11}$$

We remove the constrain (4) and replace per time slot constrain (3) with long-term time-averaged constrain (11), so we get a relaxed problem of $P1$

$$\mathbf{P2}: \quad \max_{\{P(t),\alpha\}} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[R(t)]$$
$$\text{subject to} \quad (2), (11)$$

and we finally can use standard Lyapunov Optimization to get a solution [6].

4 Solutions

Based on Lyapunov optimization [7], we introduce a virtual queue $V(t)$ for $E_s(t)$

$$V(t) = E_s(t) - X \tag{12}$$

where X is a time-independent constant. Obviously, $V(t)$ also satisfies the constrain (11), and we can design X to ensure the solution is feasible to $P1$ at the end. Replacing the $E_s(t)$ in constrain (3) with $V(t)$, we can get a new equation

$$V(t+1) = V(t) - \Delta t P(t) + E(t). \tag{13}$$

Therefore, we can define some vital Lyapunov variables. The quadratic Lyapunov function is $L(V(t)) \triangleq V^2(t)/2$, and the per-slot Lyapunov drift conditioned on $V(t)$ at time slot t is

$$\Delta(V(t)) \triangleq \mathbb{E}\left[L(V(t+1)) - L(V(t))|V(t)\right] \tag{14}$$

where the expectation is taken with respect to $E(t)$, $\gamma_1(t)$, and $\gamma_2(t)$, given the queue length $V(t)$. Within the standard Lyapunov optimization framework, we do not solve the problem $P2$ directly, we solve it by minimizing the $drift - plus - cost$ metric that can stabilize queue $V(t)$ and optimize the time-averaged objection function. We can define this function as

$$\Delta(V(t)) + C\mathbb{E}[-R(t)|V(t)]$$

where C is a constant as a weight parameter. Based on [8], we can get the $drift - plus - cost$ function's upper bound

$$\begin{aligned} &\Delta(V(t)) - C\mathbb{E}[R(t)|V(t)] \leq \\ &B + V(t)\mathbb{E}[E(t) - \Delta t P(t)|V(t)] - C\mathbb{E}[R(t)|V(t)] \end{aligned} \tag{15}$$

where $B = max\{E_{cm}, \Delta t P_{max}\}^2/2$. Due to the difficulty of solving $drift - plus - cost$ function directly, we minimize its upper bound. This is a usual method to solve this problem based on Lyapunov. Finally, we design an algorithm to determine $P(t)$ and α by minimizing the upper bound of the $drift - plus - cost$ function per time slot. Given $h_1(t)$, $h_2(t)$, and $V(t)$, we can convert this problem to an equivalent per-slot optimization problem by removing the expectation $\mathbb{E}[-]$ and the constant B.

$$\begin{aligned} \mathbf{P3}: \min_{P(t),\alpha} \quad & V(t)[E(t) - \Delta t P(t)] \\ & - C\{\omega_1 \frac{1}{2}\log\{1 + \frac{\alpha P(t)\left|h_1(t)\right|^2}{\sigma^2}\} \\ & + \omega_2 \frac{1}{2}\log\{1 + \frac{(1-\alpha)P(t)\left|h_2(t)\right|^2}{\sigma^2 + \alpha P(t)\left|h_2(t)\right|^2}\}\} \\ \text{subject to} \quad & 0 \leq \alpha \leq 1 \\ & 0 \leq P(t) \leq P_{\max} \end{aligned}$$

Obviously, this problem is not a convex problem. In this letter, we use alternative optimization to solve it.

A. Fix α, determine $P(t)$

When α has been fixed, this problem can be rewritten as

$$\begin{aligned}\mathbf{P4}: \min_{P(t)} \quad & V(t)[E(t) - \Delta t P(t)] \\ & - C\frac{\omega_1}{2}\log\{1 + \frac{\alpha P(t)\left|h_1(t)\right|^2}{\sigma^2}\} \\ & - C\frac{\omega_2}{2}\log\{\sigma^2 + P(t)\left|h_2(t)\right|^2\} \\ & + C\frac{\omega_2}{2}\log\{\sigma^2 + \alpha P(t)\left|h_2(t)\right|^2\} \\ & \text{subject to} \quad 0 \le P(t) \le P_{\max}\end{aligned}$$

The first to the third equation are all convex, and we can use Taylor expansion to reconstruct the fourth. Finally, we can use convex optimization [9] to get the optimal $P^*(t)$.

B. Fix $P(t)$, determine α

When $P(t)$ has been fixed, we discard these equations without parameter α, this problem can be rewritten as

$$\begin{aligned}\mathbf{P5}: \min_{\alpha} \quad & - C\frac{\omega_1}{2}\log\{1 + \frac{\alpha P(t)\left|h_1(t)\right|^2}{\sigma^2}\} \\ & + C\frac{\omega_2}{2}\log\{\sigma^2 + \alpha P(t)\left|h_2(t)\right|^2\} \\ & \text{subject to} \quad 0 \le \alpha \le 1.\end{aligned}$$

We can use convex optimization to find optimal α directly. Finally, we can get the optimal α^*.

C. Apply the solution to P1

P2 and P3 are the same relaxed problems, their solution may be not feasible to P1. We can use the method mentioned by [8] to ensure it. We design proper X and C to achieve this goal based on bounded channel scenario. Because the destination D1 is a strong user, only assuming that it has a bound is enough. we define $\Gamma_1(t) = \frac{|h_1(t)|^2}{\sigma^2}$ and $\Gamma_1(t) \le \Gamma_{1max}$, and set $X = \Delta t P_{max} + C\Gamma_{1max}/\Delta t$, while $C \in (0, C_{max}]$ with

$$C_{max} = \frac{E_{max} - E_{cm} - \Delta t P_{max}}{\Gamma_{1max}/\Delta t}.$$

D. Performance Analysis

We define S_{alg} as the Algorithm 1's solution of P1 and S_{opt} as the optimal solution of P1. We can prove the performance bound between the two solutions by Theorem 1.

Algorithm 1. The total solution algorithm

1: Given Γ_{1max} and C_{max}
2: **Initialization**: All system parameters
3: **Repeat**:
Obtain $P(t)$ by solving (P4);
Obtain α by solving (P5);
4: **Until**: Convergence.

Theorem 1. *We only concern bounded channel scenario. Assume the $\gamma_1(t)$, $\gamma_2(t)$, and E(t) are all i.i.d over time. Therefore the performance gap between theory and practice is bounded and can be presented as*

$$S_{opt} - S_{alg} \leq \frac{B}{C} \tag{16}$$

where B is defined below (15).

Proof. Follow the same proof processes in [10] Theorem 1.

5 Simulation Results

In this section, we present the computer simulation results of Algorithm 1's performance in this NOMA system. We assume that $E(t)$ is uniformly distributed between $[0, 0.3]$ J, $\delta^2 = 1$, and $E_{max} = 50$ J. The maximum charging amount per slot is $E_{cm} = 0.3$ J and the maximum transmission power is $P_{max} = 0.5$ W. We set time slot duration $\Delta t = 1$ sec, $\omega_1 = 1$ and $\omega_2 = 2$. We only consider single antenna and generate channel $h_1(t)$ and $h_2(t)$ as *i.i.d.* complex Gaussian random variable over time t. The mean normalized channel gain of $h_1(t)$ is 10 dB and $h_2(t)$'s mean normalized channel gain is 0 dB. This results in $\Gamma_{1max} = 16.6$ dB and we use C_{max} to construct X.

We choose some representative algorithms to compare with Algorithm 1:

(*a*) *Halving algorithm*: At each time slot, the transmitter uses half of maximum possible transmission power that based on greedy algorithm. It can save energy for future use.
(*b*) *Conventional NOMA algorithm*: At each time slot, the transmitter always tries to maximize the transmission rate based on NOMA alternative optimization and considers Energy causality.

$$\begin{aligned} \max_{P(t),\alpha} \quad & R(t) \\ \text{subject to} \quad & 0 \leq \alpha \leq 1 \\ & 0 \leq P(t) \leq P_{\max} \end{aligned}$$

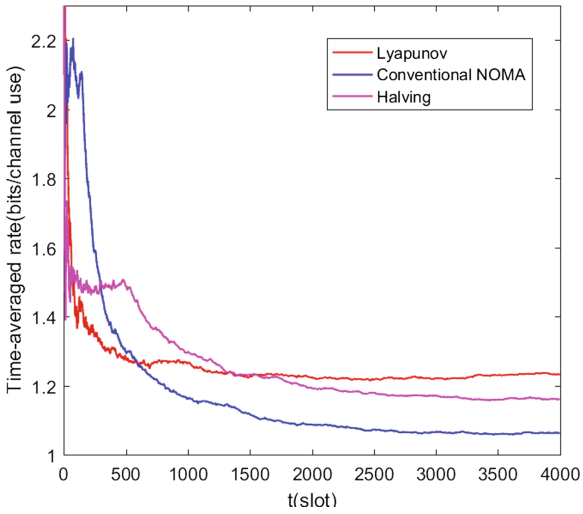


Fig. 2. Time-averaged sum rate vs. time slot t.

Figure 2 shows the sum rate of proposed optimal solutions compared with Halving solution and Conventional NOMA solution by using time-averaged rate $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[R(t)]$. Our proposed optimal algorithm provides better performance over all other algorithms. The achieved average rate of strong user by our algorithm is about 8% and 17.7% higher than Halving algorithm and Conventional NOMA algorithm. The performance gain of our algorithm over other algorithms comes from strategic energy conservation and opportunistic transmission.

Figure 3 shows that the relation among the channel gain $\gamma_i(t), i = \{1, 2\}$, the SOB $E_s(t)$ and user power. User power changes according to channel gain so algorithm 1 is an opportunistic transmission algorithm depends on channel condition. When channel condition is good, both strong user and weak user can get more power allocation compared to bad channel condition. The battery status roughly sustains at a certain level so that Lyapunov Optimization works well and can stabilize the battery status while maximizing user rate because Lyapunov stability means battery contain sustains at a certain level in this case.

Figure 4 shows that battery capacity E_{max} can affect the performance of sum rate of the Algorithm 1. In order to achieve a higher long-term averaged rate, we need to increase battery capacity. According to Fig. 4, when we constantly increase battery capacity, long-term averaged rate also increases at the beginning. When it approaches a point, the rate converges to a constant that still fluctuates and limited by the maximum power P_{max} and charging rate E_{cm}. Therefore, we can use small battery capacity to achieve the maximum user sum rate.

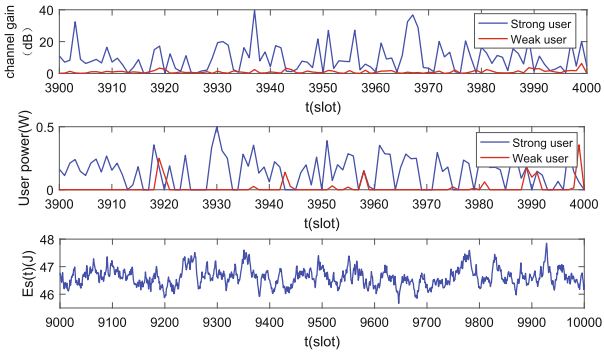


Fig. 3. Time trajectory of system parameters: Top: channel gain; Middle: user power; Bottom: $E_s(t)$.

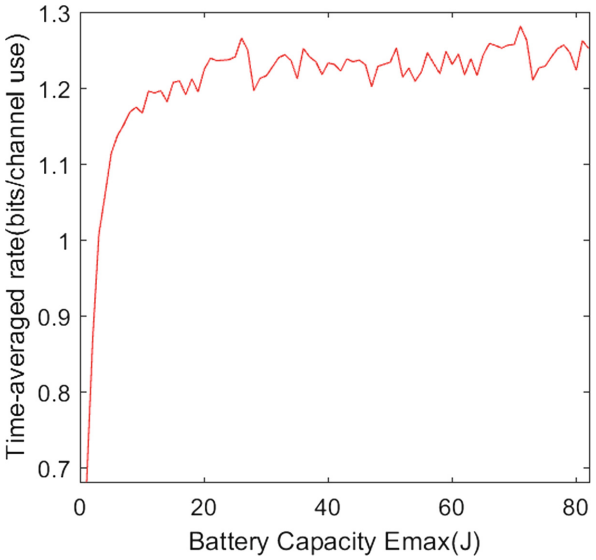


Fig. 4. The long-term averaged rate vs. battery size E_{max}.

6 Conclusions

In this letter, considering the NOMA scheme, our Lyapunov optimization algorithm over fading channel with energy harvesting scheme transforms the long-term time-averaged transmission rate to a single time slot optimal problem and uses alternative optimization to solve it. Simulation results have shown that the proposed optimal scheme outperforms other two compared algorithms and can achieve Lyapunov stability.

Acknowledgments. The research work was supported by the Guangdong science and technology project (No. 2016A010101032), Guangzhou science and technology project (No. 2014A030311032) and Guangzhou college and university science and technology project (No. 61472458).

References

1. Lee, S., Liu, L., Zhang, R.: Collaborative wireless energy and information transfer in interference channel. IEEE Trans. Wirel. Commun. **14**(1), 545–557 (2015)
2. Zhang, Q., Li, Q., Qin, J.: Robust beamforming for nonorthogonal multiple-access systems in MISO channels. IEEE Trans. Veh. Technol. **65**(12), 10231–10236 (2016)
3. Yuan, F., Zhang, Q.T., Jin, S., Zhu, H.: Optimal harvest-use-store strategy for energy harvesting wireless systems. IEEE Trans. Wirel. Commun. **14**(2), 698–710 (2015)
4. Xu, J., Zhang, R.: Throughput optimal policies for energy harvesting wireless transmitters with non-ideal circuit power. IEEE J. Sel. Areas Commun. **32**(2), 322–332 (2014)
5. Neely, M.J., Huang, L.: Dynamic product assembly and inventory control for maximum profit. In: 49th IEEE Conference on Decision and Control (CDC), Atlanta, GA, pp. 2805–2812 (2010)
6. Choi, K.W., Kim, D.I.: Stochastic optimal control for wireless powered communication networks. IEEE Trans. Wirel. Commun. **15**(1), 686–698 (2016)
7. Neely, M.J.: Stochastic Network Optimization With Application to Communication and Queueing Systems. Morgan and Claypool, San Rafael (2010)
8. Amirnavaei, F., Dong, M.: Online power control optimization for wireless transmission with energy harvesting and storage. IEEE Trans. Wirel. Commun. **15**(7), 4888–4901 (2016)
9. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
10. Mao, Y., Zhang, J., Letaief, K.B.: A Lyapunov optimization approach for green cellular networks with hybrid energy supplies. IEEE J. Sel. Areas Commun. **33**(12), 2463–2477 (2015)



SEA: SDN-Based Evolution Architecture for 5G Network

Chen He(✉) and Xiaofeng Qiu

Beijing Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing, China
{hchen, qiuxiaofeng}@bupt.edu.cn

Abstract. Software Defined Network (SDN) provides a solution to design 5G network architectures. The existing 5G network architectures based on SDN have various problems, such as the difficulty in implementing and deploying, the poor scalability, etc. In this paper, we propose SEA (SDN-Based Evolution Architecture for 5G Network). SEA reforms the LTE/EPC system and retains the main service logic of LTE, so that it has excellent service performance. Besides, SEA uses the GTP-enabled SDN switch, which has only 4 MB code, to forward traffic, making it easier to deploy on a large scale. Finally, we implement the SEA demo and LTE contrast systems. Experiments show that the SEA demo and LTE system take the same time to build the bearing. Meanwhile, both of them can respond to the data request of the user equipment (UE) within 30–60 ms. In addition, SEA can recover data path within 6 ms.

Keywords: Software Defined Network · 5G network architecture
LTE

1 Introduction

The Software Defined Networking (SDN) brings new opportunities for network innovation. SDN is aimed at decoupling the control plane and data plane of the network and seeking centralized control and programmability of the network state. In typical SDN architecture, the SDN controller is responsible for centralizing control of the SDN switch, which is the generic forwarding device on the data plane. OpenFlow protocol is a mainstream application protocol in SDN. The SDN controller communicates with the SDN switch using OpenFlow protocol and opens the North API to facilitate deployment of network applications.

With the rapid expansion of LTE network commercial deployment scale, the mobile data traffic has experienced explosive growth [1]. A study [2] shows that the total network cost will exceed the total revenue very soon. One of the main reasons is that the current mobile network has the tightly coupled architecture. In order to meet the increasing demand for bandwidth, 5G mobile communication system research and development was quickly put on the agenda. Designing the 5G network architecture has also become an important part of the study. At present, multiple international standardization organizations are doing research work on the 5G network architecture. SDN, which decouples the control and data planes, can solve the existing problems in

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 264–273, 2018.
https://doi.org/10.1007/978-3-319-78130-3_27

LTE's tight coupling network architecture. Therefore, it has become a mainstream approach for designing 5G network architecture [3, 4].

A number of SDN-based 5G network architectures have been proposed in the papers recent years. However, there exist large differences in the various architectures, and the foci are also different. In order to provide reference for future research by summarizing and comparing the various architectures, we propose the following three aspects to analyze the pros and cons of architectures:

(1) Difficulty of Implementation: Considering the reality, the difficulty of implementing 5G architecture should be considered as an aspect of architecture evaluation. We believe that the difficulty of deploying new architectures is positively correlated with the number of newly designed network elements (NEs) in the architecture. The larger the number of newly designed NEs, the higher the cost of new device development and deployment, and the greater the difficulty of implementing the architecture. Therefore, the 5G architecture with lower implementation difficulty should be considered more superior in the assessment.
(2) Architectural Performance: Compared with the LTE core network architecture, the 5G architecture, which makes some changes in network elements and communication methods, will affect the performance of the system's communication. We mainly consider service performance on the control plane and forwarding performance on the data plane. Communication between network elements is the main cause of the delay in service processing. So we think that in the service process, the less the steps of communication between the elements, the more excellent service performance. Similarly, on the data plane, we believe that the fewer devices the traffic must to pass, the shorter the forwarding delays.
(3) Scalability: Mobile core network is a huge system, which requires 5G architecture to have good extendibility in two aspects: network node scalability and controller extendibility. Network node scalability mainly refers to the difficulty of large-scale deployment of device. The 5G architecture, which contains fewer types of device, is easier to deploy on a large scale. Controller extendibility mainly considers whether the network controller can be flexible to expand other functions, such as security features, etc.

Among the existing SDN-based 5G architectures, some of the SDN-based redesigned architecture is difficult to implement, while others reduce the implementation difficulty by reforming the LTE/EPC architecture. But the latter use a variety of forwarding devices on the data plane, which is not enough to simplify and increase the difficulty of implementation. Besides, some of these architectures introduce a larger service delay that adversely affects architectural performance.

In this paper, we propose an SDN-based 5G network evolution architecture called SEA. It uses only the SDN switch to forward traffic on the data plane, which is easy to implement and deploy in large numbers. In addition, SEA retains most of the LTE service processes, so that it does not introduce significant delay compared to LTE, which means the system performance is well. Finally, SEA has good scalability, for instance, it can enhance the security capabilities through functional expansion.

SEA reforms the LTE/EPC architecture based on SDN. It uses the SDN switch that supports the POF protocol [5] as the only forwarding device on the data plane. This switch can support GTP and IP/TCP and other protocols, so that it can complete the data plane service in mobile network and is easy to implement and deploy a lot. The SDN controller in SEA centrally controls the data plane forwarding device, which facilitates the security extension of the architecture or the integration of heterogeneous networks. SEA has made minor changes to LTE's service processes but has less impact on architectural performance.

At the same time, we implement a SEA demo and measure the performance of the three main aspects of SEA: establishing default bearer, forwarding on data plane and using PACKET_IN.

The remainder of this paper is organized as follows. In the second section, we will introduce SDN-based 5G network architectures which have been proposed and evaluate them according to the three points proposed above. In Sect. 3, we introduce SEA, describe its typical service processes, and show its advantages. In Sect. 4, we perform performance testing, analysis and evaluation of SEA's typical services. Finally, a summary and outlook is given in Sect. 5.

2 Related Works

Based on the characteristics of SDN, researchers have designed some future network architectures in order to replace the current inefficient networks. Some researchers believe that the data plane and the control plane should be decoupled completely in the future mobile network [3]. The UE data is connected to the Public Data Network (PDN) through the data plane, and the control plane is responsible for managing the services such as user authentication and forwarding control. The network element should be reasonably optimized and reallocated. Also in [4], wireless network architecture with small cells, massive MIMO, NFV, and SDN technology is proposed and the data plane/control plane is separated, and the plurality of that can be combined into a network cloud.

However, a complete redesign of the 5G network architecture means that traditional network infrastructures will no longer be applicable, so both of the architecture designs described above will cost a huge in the actual network upgrade [6] and are hard to implement.

Taking into account the deployment of existing network equipment, some researchers have proposed SDTN (software defined transitional networking) to seek a phased implementation [7]. SDTN still uses the existing traditional network forwarding equipment on the data plane, which reduces the cost of network upgrade to a certain extent. But this architecture still introduces a large number of new network elements, increasing the cost of implementation. At the same time, SDTN significantly increases the processing steps of the main service (such as the establishment of bearer), increasing delay, and affecting system performance.

Unlike the redesign of 5G network architecture, some researchers believe that reforming LTE/EPC architecture will cost lower and be easy to implement. Reference [8] decouples the data plane function and the control plane function of SGW and use

SGW-C (SGW Control Plane) and Mobility Management Entity (MME) as the northbound application (APP) of the SDN controller. However, this architecture increases the service steps of establishing bearer, introduces a significant system delay. Besides, the controller in [8] does not have the centralized control of the data plane, making it difficult to expand the function.

Two ways for reforming LTE are envisioned by [9]: decoupling SGW and PGW functions into control and data planes, and coordinating communication between the control plane and the data plane by SDN controller; or using the SDN controller to communicate with the MME, direct access to mobility events, and timely development of routing strategies. The former design is excellent, but it has a variety of forwarding devices in the data plane, so still not simple enough.

In [10], all network elements (including eNodeB, SGW, PGW) with data forwarding services in LTE are decoupled. Then it combines the separated control plane function and the MME function. The network SDN controller centrally manages and controls the data forwarding of the data plane. However, since the control functions all concentrated in one network element, it increase the difficulty of implementation and more likely to lead to a single point of failure.

Table 1 summarizes the shortcomings of the 5G architecture proposed in the existing literature from the three aspects of architecture evaluation.

Table 1. Summary of deficiencies in 5G architecture by existing research

Research work	Difficulty of implementation	Architecture performance	Extendibility
[3, 4]	Hard		
[7]	Hard	Increase the delay	
[8]	Not easy	Increase the delay	Weak
[9]	Not easy		
[10]	Little hard		

3 Proposed Evolution Architecture

3.1 Architecture Description

In order to reduce the implementation difficulty and improve the scalability of the 5G network architecture, we propose the SDN-Based Evolution Architecture for 5G Network called SEA, as shown in Fig. 1. In SEA, we have retained some network elements (such as eNodeB, MME, etc.) of LTE. We also separate the control function from both SGW and PGW as independent NEs (SGW-C and PGW-C). On the data plane, we use the SDN switch to forward traffic. Besides, the SDN controller is responsible for managing communication between the control plane and the data plane. SEA consists of the following entities:

SGW control plane (SGW-C): represents the SGW's intelligence part. It is responsible for GTP tunnel establishment including TEIDs allocation. The SGW-C communicates

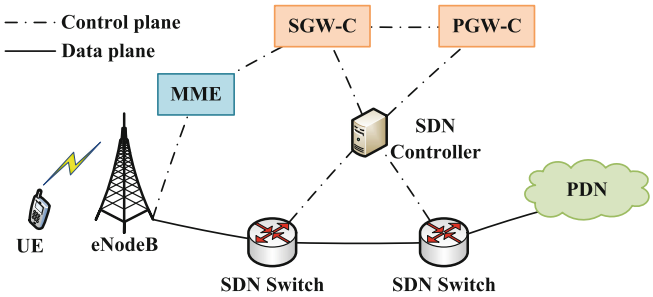


Fig. 1. SEA: SDN-based evolution architecture for 5G Network

with the SDN controller through the northbound API and sends commands to establish a GTP tunnel. The communication between the SGW-C and the MME is the same as that in the LTE network. The communication between the SGW-C and the PGW-C is also retained.

PGW control plane (PGW-C): represents the PGW's intelligence part. It is responsible for establishing the default bearer and assigning IP addresses to the user equipment (UE) while the user is attached. The PGW-C also communicates with the SDN controller through the northbound API and sends commands to establish the connection between the access data of the UE and PDN.

SDN Controller (SDN-ctr): is the main part of SEA. It is responsible for obtaining data plane management policies (such as establishing GTP tunnel) and related parameters from SGW-C and PGW-C. At the same time, it manages the SDN switch and sends the corresponding flow table to the most appropriate SDN switch to implement the management policy. The SDN controller uses a protocol called POF, which will be described in detail below.

SDN Switch (SDN-sw): is the only forwarding device on the data plane and is responsible for the data forwarding of the entire data plane. The SDN switch in SEA also uses the POF protocol, which is the same as the SDN controller. The SDN switch supports GTP (GPRS Tunneling Protocol), and can complete the establishment of GTP tunnel, GTP Message encapsulation/decapsulation and other operations.

eNodeB (eNB): has the same function as in the 3GPP standards.

MME: is a NE from LTE and is retained and used by SEA.

Protocol Oblivious Forwarding (POF) is an extension of the OpenFlow protocol proposed by Huawei [5]. The POF protocol still retains OpenFlow's processing logic based on multistage flow tables while increasing support for multiple network protocols such as GPRS Tunneling Protocol (GTP). Both SDN-ctr and SDN-sw in SEA use the POF protocol, so they can support protocols such as GTP and IP, which enable them to handle both GTP messages and IP/TCP messages.

3.2 Operation of SEA

In this section, we describe the changes to LTE major service processes in SEA. We introduce two major service scenarios of SEA about the changes: scenario 1 is how to establish the default bearer. It contains two sub-scenes: a: how to establish the GTP tunnel after decoupling SGW. b: how to connect the GTP tunnel andPDN after decoupling PGW. Scenario 2 is SEA's exception handling and recovery on the data plane, which is about the use of PACKET_IN.

3.2.1 Establishment of the Default Bearer

Figure 2 shows the process of establishing the default bearer for the UE in SEA. Since the eNodeB and MME in SEA are still the same as those in the 3GPP standard, the eNodeB still forwards the request to the MME for authentication when the UE sends the access request. Then the MME requests the SGW-C and the PGW-C to establish a default bearer for the UE.

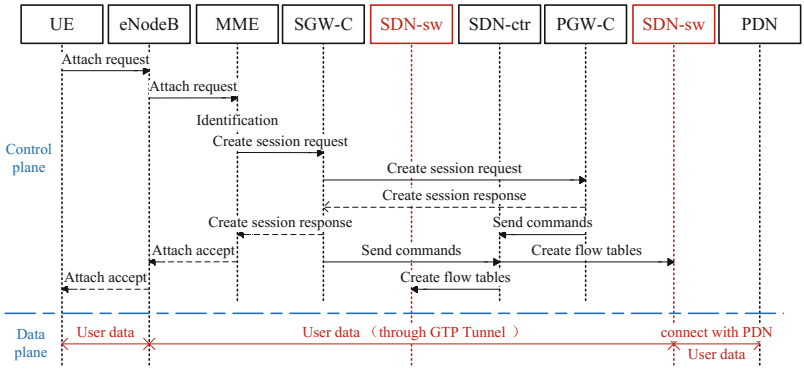


Fig. 2. Establish the default bearer in SEA, including establishing the GTP tunnel and exchanging messages between the GTP tunnel and PDN

The establishment of the default bearer is divided into two parts. One is to establish the GTP tunnel by the SGW-C, and the other part is to connect the GTP tunnel and PDN by the PGW-C.

Establish the GTP Tunnel: The SGW-C is responsible for establishing the GTP tunnel. The SGW-C assigns the TEID, a unique identifier, to the uplink and downlink of the GTP tunnel and sends commands to SDN Controller to establish GTP tunnels.

Connect the GTP Tunnel and PDN: The PGW-C is responsible for connecting the GTP tunnel and PDN. It establishes the default bearer and assigns an IP address for the UE during UE attachment. On the data plane, connecting GTP tunnel and PDN requires the SDN switch to convert IP/TCP messages and GTP messages, as determined by the commands issued by the PGW-C.

After parsing these commands, which from the SGW-C and the PGW-C, the SDN-ctr selects the two most suitable SDN-sws to replace SGW and PGW functions

respectively on the data plane. (For convenience, we will refer to these two SDN-sws as SGW-sw and PGW-sw, respectively). Then the SDN-ctr sends the flow tables to these two SDN-sws to establish a GTP tunnel between the eNodeB and the PGW-sw through the SGW-sw and connect the GTP tunnel to PDN through the PGW-sw.

3.2.2 Exception Handling and Recovery on Data Plane

Generally, in SEA, when a UE requests access to the network, the commands are sent directly from SGW-C and PGW-C to SDN-ctr. The SDN-ctr parses the commands and sends the matching flow tables to the appropriate SDN Switch to establish the channel on the data plane.

However, some exceptions can cause the data channel to be interrupted. For example, the SDN-sw does not receive the flow table due to network congestion, the SDN-sw is emptied, which means it does not have any flow table, due to device restart, and so on. In these cases, SEA will use the PACKET_IN mechanism to recover the data channel.

In SEA, the SDN-sw and the SDN-ctr that support the POF protocol still use OFPT_PACKET_IN, which conforms to the OpenFlow protocol, to handle unmatched packets. When the data channel is interrupted, SDN-sw lacks the forwarding rules of the packet. Then the SDN-sw will use OFPT_PACKET_IN to ask SDN-ctr how to process the data messages and cache these messages at the same time. After analyzing OFPT_PACKET_IN, the SDN-ctr requests SGW-C or PGW-C to process it and return the answer to the SDN switch. After the SDN-sw gets the processing command and creates flow tables, the data channel will be restored. The processing of PACKET_IN in SEA is shown in Fig. 3.

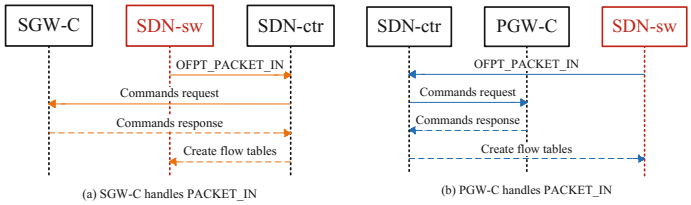


Fig. 3. PACKET_IN processing

3.3 Superiority of SEA

The SDN-based 5G architecture proposed in the existing literature can be divided into two kinds of ideas. One idea is to redesign the 5G architecture based on the SDN [3, 4, 7]. These architectures are costly due to the design of a large number of new network elements. Another idea is to reform the LTE/EPC architecture based on the SDN concept [8–10]. We think that this idea significantly reduces the difficulty of implementation, and therefore should be considered.

Based on the latter, we proposed SEA, which has reformed the LTE/EPC architecture. Compared to [8–10] that have complex forwarding devices, SEA is easier to implement. SEA uses only the SDN switch to forward traffic on the data plane,

resulting in lower development costs and easier large-scale deployment. SEA retains the basic service processes of LTE, which does not affect performance.

In addition, SDN-ctr has an excellent global perspective so that it can sense traffic in a timely manner. It enables SEA to have superior security and strong scalability. For example, SDN-ctr can be connected to the Security Controller (SC) through the open API to enhance its own security.

4 Performance Evaluation

In order to test the performance of SEA, we set up SEA experiment system. The system use the following software: the SDN-ctr developed based on Floodlight, the SDN-sw developed on the basis of Openvswitch, and both the SDN-ctr and the SDN-sw support the POF protocol, the small cell, the MME used in the LTE system, and the SGW-C and the PGW-C decoupled from the LTE system. Meanwhile, we use the LTE system mentioned above to do contrast experiment. We try to deploy the NEs that perform similar functions in these two systems on the same device.

We measured service delay in three main services of SEA experiment system, including establishment of default bearer, transmission on data plane and PACKET_IN processing. At the same time, we use the LTE system for the same test. The data comparison proves that SEA has excellent architectural performance.

4.1 Establishment of Default Bearer

This part tests the delay of the main steps for establishing a default bearer in SEA. As shown in Fig. 2, when SGW-C sends the create session response to MME, it immediately sends commands to SDN-ctr (PGW-C is the same). However, the default bearer has not yet been established at this time, since MME and eNodeB also need to do some follow-up work. Therefore, we can ensure that SDN-ctr has already sent commands to SDN-sws when eNodeB acknowledges the bearer establishment, and SDN-sws has created flow tables, in other words, the SEA system does not increase the delay of the establishment of default bearer.

Also we tested the process for establishing the default bearer in the LTE system. Figure 4 is the timing diagram of the main steps for SEA and LTE systems to establish default bearers (Some steps in SEA do not exist in LTE). We can see that compared with LTE, SEA and LTE essentially take the same time when establishing default bearer. It proves that SEA does not introduce new system delay, and has the same excellent service performance as LTE.

4.2 Transmission on Data Plane

We measured the delay between sending a DNS request from UE and receiving the DNS response from PDN to the UE. Also we do the same test in LTE. Figure 5 shows the delay of both. Obviously, SEA in forwarding performance is comparable to LTE.

In the SEA, since SDN-sw is the only forwarding device on the data plane, the transmission performance depends primarily on the forwarding performance of

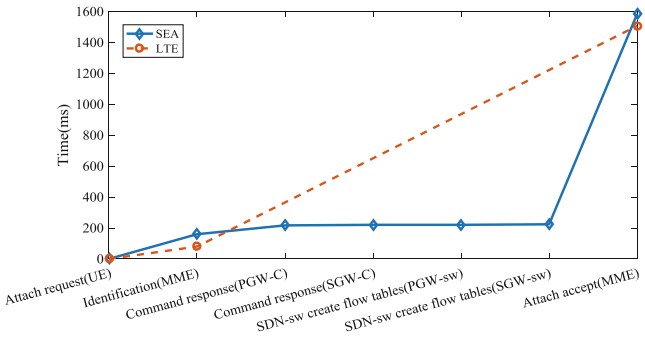


Fig. 4. Time analysis for establishing a default bearer

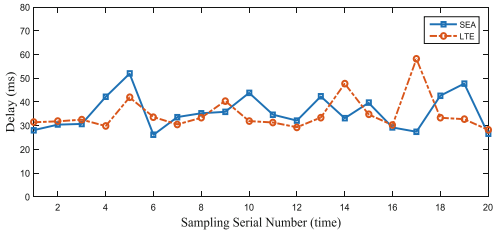


Fig. 5. Delay by UE between sending DNS request and receiving response

SDN-sws and the hardware devices that deploy the SDN-sws. Taking into account the SDN-sw structure is simple, easier to improve than LTE, we believe that the data transmission of SEA will be better than LTE in the future.

4.3 PACKET_IN Processing

PACKET_IN, which does not exist in the LTE, is a method used by SEA to handle exceptions. In this section, we perform the following test: Disconnect the SDN-sw abnormally to empty the flow table, reconnect it. We measured the delay of the SDN-sw recovery flow tables through PACKET_IN.

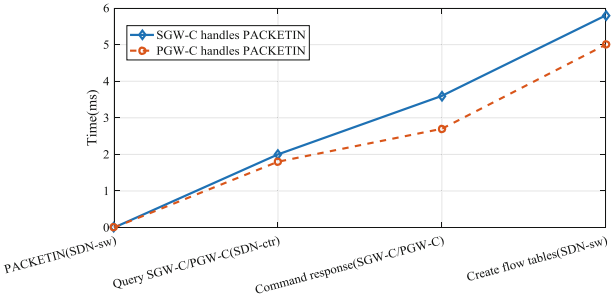


Fig. 6. Time analysis for PACKET_IN processing

Figure 6 is the timing diagrams for the main steps of SGW-C and PGW-C processing PACKET_IN. Obviously, SDN-sw takes less than 6 ms to recover the data transmission through PACKET_IN. As a result of the rapid recovery, we almost do not feel that there is an exception in the actual use.

5 Conclusion and Perspectives

This paper proposes a 5G network architecture based on SDN, called SEA, which has good centralized control capability and extendibility and is easy to implement. In particular, SEA uses the SDN-sw, which is no difference among each other, to forward data on data plane, making it easy to deploy in large numbers. Since the service process of SEA is similar to LTE, the performance is similar too. Eventually, we implemented SEA system and tested the main service performance as a reference. We demonstrate that SEA has excellent performance compared to LTE.

Our next work is to achieve abnormal traffic detection based on SEA, which is an extension of security.

Acknowledgment. This research was sponsored by National High Technology Research and Development Program of China (863 Program) (No. 2015AA016003) and Beijing Laboratory of Advanced Information Networks.

References

1. Cisco visual networking index: Global mobile data trac forecast update 2012–2017. White Paper, 6 February 2013
2. Tellabs: Tellabs “end of profit” study executive summary. Technical report, January 2011
3. Pentikousis, K., Wang, Y., Hu, W.: MobileFlow: toward software-defined mobile networks. IEEE Commun. Mag. **51**(7), 44–53 (2013)
4. Agyapong, P.K., Iwamura, M., Staehle, D., Kiess, W., Benjebbour, A.: Design considerations for a 5G network architecture. IEEE Commun. Mag. **52**(11), 65–75 (2014)
5. Song, H.: Protocol-oblivious forwarding: unleash the power of SDN through a future-proof forwarding plane. In: Proceedings of the 2013 ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking, pp. 127–132 (2013)
6. Levin, D., Canini, M., Schmid, S., Feldmann, A.: Incremental SDN deployment in enterprise networks. Comput. Commun. Rev. **43**(4), 473–474 (2013). Proceedings of the SIGCOMM 2013 and Best Papers of the Co-Located Workshops
7. Kyung, Y., Nguyen, T.M., Hong, K., Park, J., Park, J.: Software defined service migration through legacy service integration into 4G networks and future evolutions. IEEE Commun. Mag. **53**(9), 108–114 (2015)
8. Said, S.B.H., Sama, M.R., Guillouard, K., Suciu, L., Simon, G., Lagrange, X., Bonnin, J.-M.: New control plane in 3GPP LTE/EPC architecture for on-demand connectivity service. In: Proceedings of the 2013 IEEE 2nd International Conference on Cloud Networking, CloudNet 2013, pp. 205–209 (2013)
9. Costa-Requena, J.: SDN integration in LTE mobile backhaul networks. In: International Conference on Information Networking, ICOIN 2014, pp. 264–269 (2014)
10. Haw, R., Hong, C.S., Lee, S.: An efficient content delivery framework for SDN based LTE network. In: Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication, ICUIMC 2014 (2014)



A Method of Link Margin Analysis for Aeronautical Communication Systems in LOS Environment

Guofeng Jiang[1,2(✉)], Yangyu Fan[1], Hongbo Yuan[2], and Gang Zhou[3]

[1] School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China
kai6921@126.com
[2] The First Aeronautic College, Xinyang 464000, China
[3] 95965 PLA Troops, Gucheng 253800, China

Abstract. Link margin (LM) analysis has a critical importance on designing and using aeronautical communication (AC) systems. Antenna gains are vital to analyze LM, while gains of airborne antenna are affected by position or attitude of aircrafts. To study the relationship between LM and position or attitude of aircrafts, a method of LM analysis is presented based on coordinate transformation. The coordinate is used to be calculated for elevation and azimuth angles, which are associated with antenna gains. The method is validated in typical scenarios, and the result indicates that the LM calculating model is accuracy and practicable.

Keywords: Aeronautical communication · Link margin
Antenna gains · Coordinate transformation

1 Introduction

The airborne antenna is fixed on the fuselage, whose radiation characteristic varies according to the position and attitude of aircraft. So the position and attitude of aircraft must be taken into account when calculating LM. The common way to study LM is to calculate the positional angles relative to the antenna, and then the antenna's radiation properties data can be obtained according to the positional angles. There are some studies as discussed elsewhere (see [1–3]) on LM analysis, while the study of those literatures generally assume that the earth is an ideal sphere and the positional angles of the terminal are based on the one coordinate system. So the results from those literatures can be used in the near-range communication only, the calculation error will be increase in the far-range communication due to the error of positional angles relative to antenna. The calculation error of positional angles results from two reasons, one reason is that the earth is not an ideal sphere but an ellipsoid, the other reason is that the positional angles of the antenna should be calculated in the coordinate frame relative to their own aircraft-body separately. This paper presents a

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 274–283, 2018.
https://doi.org/10.1007/978-3-319-78130-3_28

mathematical framework to calculate elevation angle and azimuth angle based on the above reasons, which ensure the correct of calculation for LM even in the far-range.

The rest of this paper is organized as follows: In Sect. 2, LM model is built according to the functional block diagram of AC system. In Sect. 3, the relative positions (elevation and azimuth) are calculated based on the positions (longitude, latitude and altitude) and attitudes (yaw, pitch, and roll). In Sect. 4, the effectiveness of the method is validated in typical scenarios, and the results are discussed. In Sect. 5, the conclusion is followed.

2 LM Modeling of Aeronautical Communication System

2.1 The Calculation Equation of LM

LM measured in dB, is the difference between the receiver's sensitivity and the actual received power, and can be written as equation (see [1]):

$$L_m = P_t + G_t + G_r - S_r - L_t - L_f - L_r - L_p \tag{1}$$

Where L_m represents, P_t is output power of transmitter, L_t is losses in RF cable between transmitter and antenna, G_t is transmit antenna gain, L_f is free space path loss, L_p is polarization losses between transmitter antenna and receiver antenna, G_r is receiver antenna gain, L_r is losses in RF cable between receiver and antenna, and S_r is sensitivity of receiver.

2.2 The Calculation of Radio Propagation Loss in Free Space

There are many losses due to multipath environment during the radio wave propagation, while this paper mainly study the effect of the aircraft's position and attitude on link margin. So the space path losses are simplified and the free-space path loss (FSPL) is considered only. The FSPL can be calculated as (see [4]):

$$L_f = 32.4 + 20 \lg d + 20 \lg f \tag{2}$$

Where L_f is FSPL (in dB), d is the distance from the transmitter (in km), and f is the signal frequency (in MHz).

2.3 The Calculation of Antenna Polarization Attenuation

The polarization loss is used to quantitatively describe the coupling between the transmitter antenna and receiver antenna (see [5,6]). L_p represents the polarization loss and can be given as:

$$L_p = -20lg \left[\frac{1}{2} \pm \frac{2\left|A_T\right|\left|A_R\right|}{\left(1+\left|A_T\right|^2\right)\left(1+\left|A_R\right|^2\right)} + \frac{\left(1-\left|A_T\right|^2\right)\left(1-\left|A_R\right|^2\right)}{2\left(1+\left|A_T\right|^2\right)\left(1+\left|A_R\right|^2\right)} cos2\alpha \right] \tag{3}$$

Where A_T is axis ratio of transmitter antenna, A_R is axis ratio of receiver antenna, and *alpha* is the difference of polarization angle between the transmitter antenna and receiver antenna. The plus or minus would be selected plus when the polarization directions of the transmitter antenna and receiver antenna are same, otherwise it would be selected minus.

3 Calculation of the Relative Position Between Transmitter and Receiver

3.1 Calculation and Definition of the Relative Position in the Three-Dimensional Space

There is a Cartesian coordinates named as $o_b x_b y_b z_b$ coordinates system in the space, o_b is origin, $o_b x_b$ is x-axis, $o_b y_b$ is y-axis, and $o_b z_b$ is z-axis; The P is a some point in the space, and (x_{b1}, y_{b1}, z_{b1}) is coordinate of the P in the $o_b x_b y_b z_b$ coordinates system (see Fig. 1). The position of point P relative to the origin o_b can be specified by elevation (EL) angle and azimuth (AZ) angle.

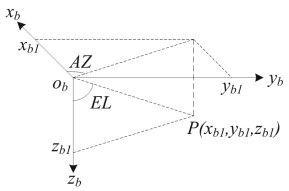


Fig. 1. The sketch map of the relative position in space

According to geometrical relationship (see [1]), AZ and EL are made:

$$AZ = \begin{cases} arccos\frac{x_{b1}}{\sqrt{x_{b1}^2+y_{b1}^2}}, & \text{if } y_{b1} \geq 0 \\ 2\pi - arccos\frac{x_{b1}}{\sqrt{x_{b1}^2+y_{b1}^2}}, & \text{if } y_{b1} < 0 \end{cases} \tag{4}$$

$$EL = arccos\frac{z_{b1}}{\sqrt{x_{b1}^2 + y_{b1}^2 + z_{b1}^2}} \tag{5}$$

3.2 Aircraft's Body Coordinates System and Attitude Angles

The aircraft body axis system is used to be the reference coordinate frame for calculating the relative position, which is defined as Fig. 2, the origin is in the center of gravity in the aircraft, x-axis points to nodes in longitudinal direction along the symmetrical axis, z-axis is perpendicular to x-axis pointing to belly in the symmetrical plane, and z-axis is defined from right-hand rule.

Figure 2 presents ground coordinate system $o_g x_g y_g z_g$ and Euler angles of aircraft. The origin o_g is just below aircraft, and o_g and o' and earth center

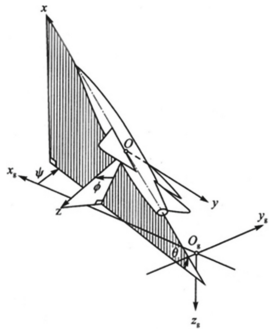


Fig. 2. Coordinate system and attitude angle of aircraft

(o', see Fig. 3) are aligned, x_g-axis adopt the direction from o_g to North Pole along the surface of the earth, z_g-axis is perpendicular to x-axis and pointed to earth center, and y_g-axis is determined by right-hand rule. Pitch angle *theta* is the angle between ox-axis and $o_g x_g y_g$ plane, which is positive when aircraft heads up, the angle *theta* between $o_g x_g$-axis and the projection of ox-axis on the $o_g x_g y_g$ plane is defined as yaw angle, which is positive when aircraft yaws right, and roll angle *varphi* is the angle between oz-axis and vertical plane including ox-axis, which is positive when aircraft rolls right.

3.3 Coordinate Transformation Used to Calculate Relative Position

The airborne sensors can provide aircraft's geographical coordinates (longitude, latitude, and altitude), but aircrafts coordinates in the other aircrafts body coordinate system are needed to calculate the relative position. The change of coordinates can be transformed through moving and rotating the coordinate system (see Fig. 3), the geographical coordinates of aircraft P and aircraft b are (L^p, B^p, H^p) and (L^b, B^b, H^b) separately, while (L^p, B^p, H^p) should be transformed from geographical coordinates to the other aircrafts body coordinate system labeled $o_b x_b y_b z_b$.

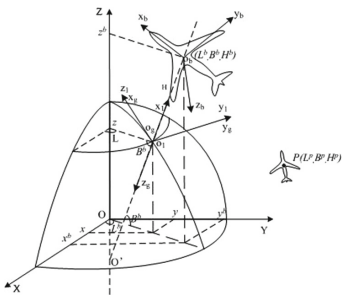


Fig. 3. Coordinate system and attitude angle of aircraft

(x_b^p, y_b^p, z_b^p) represents the coordinates of aircraft P in the $o_b x_b y_b z_b$, which is used to calculate the relative position according to (4) and (5), and it can be calculated as:

$$\begin{bmatrix} x_b^p \\ y_b^p \\ z_b^p \end{bmatrix} = T_{at} \cdot \left\{ \begin{bmatrix} x_g \\ y_g \\ z_g \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ H^b \end{bmatrix} \right\} \tag{6}$$

Where T_{at} represents attitude transformation matrix of aircraft used to transform the coordinates in the body coordinates system when attitudes of aircraft changed, and T_{at} can be calculated as (see [7]):

$$T_{at} = \begin{bmatrix} \cos\theta\cos\psi & \cos\theta\sin\psi & -\sin\theta \\ (\sin\phi\sin\theta\cos\psi - \cos\phi\sin\psi) & (\sin\phi\sin\theta\sin\psi + \cos\phi\cos\psi) & \sin\phi\cos\theta \\ (\cos\phi\sin\theta\cos\psi + \sin\phi\sin\psi) & (\cos\phi\sin\theta\sin\psi - \sin\phi\cos\psi) & \cos\phi\cos\theta \end{bmatrix} \tag{7}$$

Where (x_g, y_g, z_g) represents the coordinates of the aircraft P in the ground coordinates system $o_g x_g y_g z_g$, and can be calculated as:

$$\begin{bmatrix} x_g \\ y_g \\ z_g \end{bmatrix} = T_{oy1} \times \left(T_{Yb} \times \left(T_{Zb} \times \begin{bmatrix} X^p \\ Y^p \\ Z^p \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ OO' \end{bmatrix} \right) - \begin{bmatrix} N \\ 0 \\ 0 \end{bmatrix} \right) \tag{8}$$

Where T_{oy1} is rotation matrix used to transform $o_g x_g y_g z_g$ to $o_1 x_1 y_1 z_1$ coordinates system, and is shown as:

$$T_{oy1} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{bmatrix} \tag{9}$$

Where OO' is the range from the earth's core to the cross between curvature in prime vertical and OZ-axis, and can be calculated as:

$$OO' = N \times (1 - b^2/a^2) \times \sin B^b \tag{10}$$

Where N is radius of curvature in prime vertical, (X^P, Y^P, Z^P) represents Greenwich coordinates of aircraft P, and it can be changed from geographical coordinates (L^p, B^p, H^p) according to (11) (See [8]):

$$\begin{cases} X^p = N + H^p cos L^p cos B^p \\ Y^p = N + H^p \cos B^p \sin L^p \\ Z^p = \dfrac{b^2}{a^2} N + H^p \sin B^p \\ N = \dfrac{a^2}{\sqrt{a^2 \cos^2 B^p + b^2 \sin^2 B^p}} \end{cases} \tag{11}$$

Where a represents earth equatorial radius, and b represents earth polar radius. Where T_{yb} is rotation matrix used to rotate $OXYZ$ coordinates system B^b degree around OY-axis, and can written as:

$$T_{yb} = \begin{bmatrix} \cos B^b & 0 & \sin B^b \\ 0 & 1 & 0 \\ -\sin B^b & 0 & \cos B^b \end{bmatrix} \tag{12}$$

Where T_{zb} is rotation matrix used to rotate $OXYZ$ coordinates system L^b degree around OZ-axis, and can written as:

$$T_{zb} = \begin{bmatrix} \cos L^b & \sin L^b & 0 \\ -\sin L^b & \cos L^b & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{13}$$

4 Results and Discussion

4.1 The Typical Application System

The method of analyzing LM is now applied to an AC in LOS environment; for example, there are aircraft T and aircraft R in the air, aircraft T is regarded as the transmitter terminal and aircraft R is regarded as the receiver terminal, and the typical parameters of the AC are shown in Table 1.

Table 1. Typical parameters of the aeronautical communication

Para	f	P_t	S_r	L_r *or* L_t
Val	108 MHz	40 dBm	−110 dBm	3 dB

The coordinates of aircrafts are presented in Table 2, the NO.1 position shows that aircraft R and aircraft T have the same longitude, aircraft R is in the due south of aircraft T, and the NO.2 position shows that the difference between aircraft R and aircraft T is altitude only. The FSPL are calculated according to Eq. (2), and all these results are presented in Table 2 too.

Table 2. Typical parameters of the aeronautical communication

Number of position	Coordinates of aircraft T	Coordinates of aircraft R	Spatial distance	FSPL
1	E113.50.25	E113.50.25	414 km	125 dB
	N034.31.09	N030.47.01		
	3000 m	3000 m		
2	E113.50.25	E113.50.25	7 km	90 dB
	N034.31.09	N034.31.09		
	3000 m	10000 m		

To calculate LM, P_t and S_r and L_r and L_t and FSPL can be obtained from Tables 1 and 2, these parameters have nothing to do with the aircraft's attitude, while the parameters such as G_t and G_r and L_p are related to attitude and position of aircraft, which can be simulated using the electromagnet software.

4.2 The Typical Application Scenarios

The Variation of Link Margin When the Aircraft Rolls. The roll angle of aircraft T (*Phi-tx*) is supposed to be initiates from 0° and increases up to 360° in increments of 10°, and the other attitude angles of aircraft R and aircraft T are supposed to be zero. Just as the NO.1 position shown in Table 2, the angles variation on elevation (*EL-tx*) and azimuth (*Az-tx*) of aircraft R almost have no change when *Phi-tx* changes, so G_t doesn't change. Because the size of aircraft is much smaller than the distance between aircraft T and aircraft R, the aircraft T can be regarded as a point when calculating the position angles of aircraft T relative to aircraft R, and the change of attitude for aircraft T doesn't affect the position angles. That is to say, the angles of elevation (*EL-rx*) and azimuth (*Az-rx*) and antenna gain (G_r) of aircraft R don't change when attitude of aircraft T changes. So LM has no change when *Phi-tx* changes (See Fig. 4).

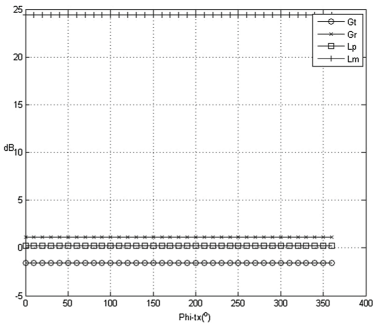


Fig. 4. Roll angle changes in NO.1 position

In the NO.2 position, the variations on antenna gain and polarization loss and LM are presented as Fig. 5. G_r has no change and G_t changes in a small range, while the polarization loss changes a lot, so the LM changes mainly depends on polarization loss, but LM changes in opposite direction of polarization loss.

The Variation of Link Margin When the Aircraft Yaws. In the NO.1 position shown in Table 2, when the yaw angle of aircraft T changes, aircraft T and aircraft R are in the same longitude, the variations on antenna gains and polarization loss and LM are presented as Fig. 6, both G_t and $G_r$change in a small range, and the polarization loss changes a lot which cause the LM to change a same value in opposite direction.

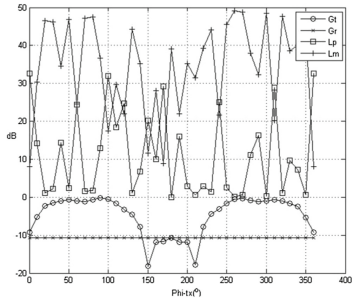


Fig. 5. Roll angle changes in NO.2 position

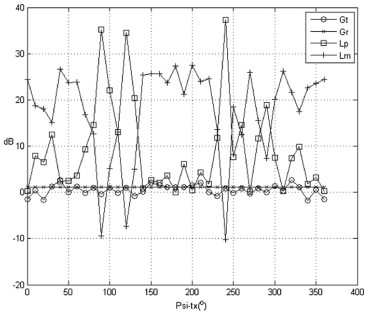


Fig. 6. Yaw angle changes in NO.1 position

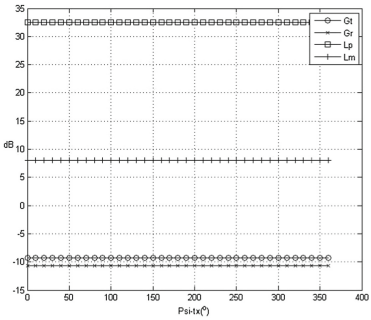


Fig. 7. Yaw angle changes in NO.2 position

In the NO.2 position, when the yaw angle of aircraft T changes, G_t and G_r don't change, and polarization loss don't change either, so LM keeps constant (see Fig. 7). In fact, just think that aircraft R is right above aircraft T, the relative angles of position between aircraft T and aircraft R don't change when aircraft T yaws, so antenna gains don't change either and LM don't change accordingly, which consists with the calculation and proves that the method of LM calculation is right.

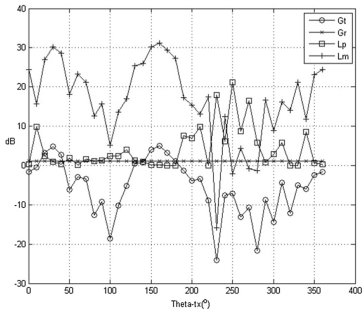


Fig. 8. Pitch angle changes in NO.1 position

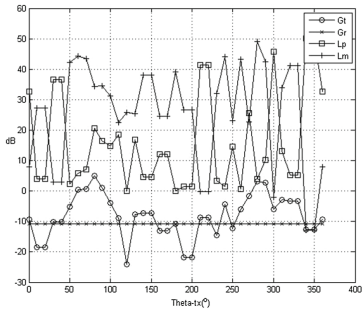


Fig. 9. Pitch angle changes in NO.2 position

The Variation of LM When the Aircraft Pitches. When the pitch angle of aircraft T changes, the variations are presented as Figs. 8 and 9. In NO.1 position, G_t changes a lot, and polarization loss changes a little when the pitch angle changes from 0° to 180°, while it changes a lot when the pitch angle changes from 180° to 360°. In NO.2 position polarization loss changes a lot when the pitch angle changes from 0° to 360°.

5 Conclusion

LM calculation model for AC system is set based on position and attitude of aircrafts. The relationship between LM and aircraft's position and attitude is discussed, and the method can be used to analysis real-time LM analysis for AC system, which is very important for designer or user to evaluate the radio link and ensure the quality of AC.

References

1. Liu, Y.: System level link budget method. J. Beijing Univ. Aeronaut. Astronaut. **39**, 47–51 (2013)
2. Erturk, M.C., Aksan, Y.: A tool for beamforming and real-time link budget analysis in aeronautical communications using kinematics. Int. J. Aerosp. Eng. **5**, 1–11 (2016)
3. Marzban, S., Mohamedpour, K.: Calculation of the communication angles between an air vehicle and receiver antenna in aeronautical telemetry. IET Microw. Antennas Propag. **12**, 2187–2195 (2010)
4. Qiong, C., Quanxing, J.: The manual of electromagnetic compatibility engineering design. National Defense Industry Press, Beijing (1993)
5. Nie, Z.: Antenna Engineering Handbook. Publishing House of Electronics Industry, Beijing (2002)
6. Zhang, J., Zhou, X., Jiang, P.: Analysis of polarization mismatching and radiation pattern of aircraft antenna. Chin. J. Radio Sci. **22**, 497–501 (2007)
7. Zhenglong, D.: Inertial Technology. Harbin Institute of Technology Press, Harbin (2006)
8. Gui, Y., Zhang, F., Jao, L.: A new method of trace and parameter measurements for flying objects. Acta Electronica Sinica **31**, 1894–1896 (2003)



Study on Handover of High Speed Railway by Rate-Splitting and Interference Cancellation

Bo Zhang[1], Hui Liu[2,3(✉)], Hai Zhu[1], and Hengzhou Xu[1]

[1] School of Network Engineering, Zhoukou Normal University, Zhoukou 466000, China
{zhangbo,hzxu}@zknu.edu.cn, zhuhai@xjtu.edu.cn
[2] School of Computer and Information Engineering, Henan University, Kaifeng 475001, China
liuhui16@rpc.edu.cn
[3] Department of Information Engineering, Henan Finance College, Zhengzhou 451464, China

Abstract. The down link communication environment of high-speed railway (HSR) is modeled by multi-user Gaussian Z interference channel (GZIC) model in the paper. The sum-capacity and the optimum transmission solution of multi-user GZIC are obtained in the strong interference regime. Based on the optimum solutions in all interference regimes, the multi-dimensional communication scheme is proposed and applied to HSR. Finally, three numerical simulations are done to show that the coverage of base station and the rate of user can increase and the frequency of handover can decrease using multi-dimensional communication scheme.

Keywords: High-speed railway · Interference channel · Handover Communication scheme · Rate-spitting · Interference cancellation

1 Introduction

In recent years, with the rapid development of high-speed railway (HSR), the operating mileage of HSR has more than 30 thousand kilometers in the world. Meanwhile, with the further development of HSR, the test speed has reached 575 km/h [1]. In the fast mobile communications environment of HSR, the present global system for mobile communications - railway (GSM-R) system has no ability to support future intelligent dispatching system and meet the growing demand of passengers for online game and multimedia service.

Many scholars have devoted their efforts to study the subjects of rapidly time-varying channel modeling [2], channel estimation [3], Doppler diversity [4,5], network architecture [6], interference [7] and handover of HSR's communication environment to estimate and track the fast time-varying channel of HSR, decrease handover failure probability, mitigate the impacts of interference and improve system performance. Various new networks architectures and power

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 284–293, 2018.
https://doi.org/10.1007/978-3-319-78130-3_29

allocation [8–11], such as distributed antenna system (DAS), cloud radio access network (C-RAN), coordinate multiple point transmissions (CoMP) and heterogeneous networks (HetNet), are used to improve the handover performance. All methods try to reduce the frequency of handover, make the coverage flexible, enlarge capacity, increase spectrum efficiency and better handover performance. However, new problems come simultaneously into being. There is a major challenge to deploy all antennas of DAS; CoMP increases the data communication among co-base stations; the seamless among different multiple radio access technologies remain a challenging problem and co-channel interference is still an open problem among the non cooperation base stations for high mobility communications. As the limited radio spectrum resource becomes more and more scarce, interference will be the dominant factor deteriorating the performance of communication. Therefore, how to restrain the effect of interference, reduce the frequency of handover efficiently and increase the rate of transmission at the edge of cell are major tasks for high mobility communications.

The main contributions of the paper are representing the communication environment of HSR by multi-user Gaussian Z interference channel (GZIC) and proposing the multi-dimensional communication scheme which can increase the rate of use and the coverage of base station.

2 System Model

Base stations are linearly deployed along the railway as shown in Fig. 1, where each user is covered by one serving base station and mainly interfered by the nearest neighbor base station. Focusing on the down link, the physical interference channel model can be expressed by multi-user GZIC as shown in Fig. 2. The input-output relationship can be described by

$$\begin{aligned} y_1^* &= a_{1,1}x_1^* + a_{2,1}x_2^* + z_1^*, \\ y_2^* &= a_{2,2}x_2^* + a_{3,2}x_3^* + z_2^*, \\ &\vdots \\ y_n^* &= a_{n,n}x_n^* + z_n^*. \end{aligned} \tag{1}$$

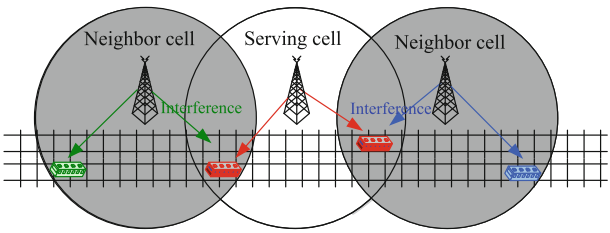


Fig. 1. The environment of railway communication.

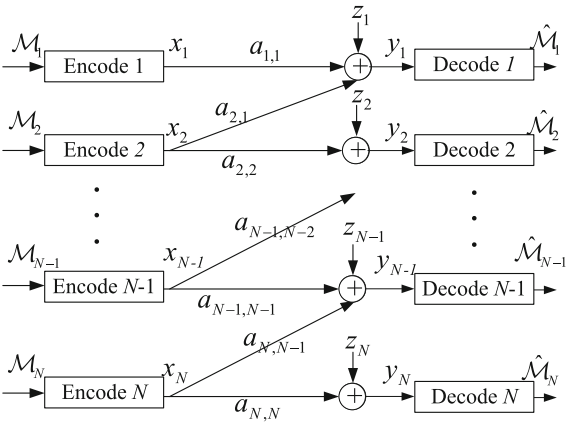


Fig. 2. The model of N-user GZIC.

where z_n^* is the i.i.d. real Gaussian noise with zero mean and N_n variance, i.e. $z_n^* \sim \mathcal{CN}(0, N_n)$, $n \in \{1, 2, \cdots, N\}$, x_n^* and y_n^* denote the input and output of user n respectively. The channel input is subject to an average power constraint $\mathbb{E}[|x_n^*|^2] \leq P_n$. The symbol $a_{n,k}$ denotes the channel gains from transmitter n to receiver k.

We assume that each receiver knows the codebooks of all users. In order to introduce the main ideas and results in the simplest possible setting, we do our analysis of the N-user GZIC by considering a symmetric interference channel, i.e. $\left|a_{(n,n)}\right| = \sqrt{A}, n \in \{1, 2, \cdots, N\}$ and $\left|a_{(n+1,n)}\right| = \sqrt{B}, n \in \{1, 2, \cdots, N-1\}$. Each transmitter has the same power $P_n = P, n \in \{1, 2, \cdots, N\}$. Each noise has the same power $N_n = N_0, n \in \{1, 2, \cdots, N\}$. Then the N-user GZIC can be represented by standard form

$$\begin{aligned} y_1 =& x_1 + hx_2 + z_1, \\ y_2 =& x_2 + hx_3 + z_2, \\ &\vdots \\ y_n =& x_n + z_n. \end{aligned} \tag{2}$$

where z_n is the i.i.d. real Gaussian noise with zero mean and unit variance, $\left|h^2\right|$ is B/A. x_n denotes the input of user n and satisfies $\mathbb{E}[|x_n|^2] \leq P^* = \frac{AP}{N_0}$. We define $\gamma(x) \triangleq \log(1 + x)$.

3 The Optimum Multi-dimensional Communication Scheme of N-User GZIC

In this section, the capacity of multi-user GZIC will be given in strong interference regime. Then we will construct the optimum multi-dimensional scheme

by capacity or achievable rate of multi-user GZIC in very weak, weak, moderate and very strong interference regimes.

3.1 The Capacity of Multi-user GZIC in Strong Interference Regime

Theorem 1. *In N-user Gaussian Z strong interference channel, the sum-capacity is*

$$C_{sum} = \begin{cases} \dfrac{N}{2}\gamma(h^2P^* + P^*), \text{when } N \text{ is even} \\ \gamma(P^*) + \dfrac{N-1}{2}\gamma(h^2P^* + P^*), \text{when } N \text{ is odd} \end{cases} \tag{3}$$

where h satisfies $1 \leq h^2 < P^* + 1$.

Proof. Firstly, an inner bound of sum-rate will be given. Let c_n denotes the reference interference factor

$$c_n = \sqrt{\frac{h^2 - c_{n+1}^2 + 1}{1 + c_{n+1}^2 P^*}}, \tag{4}$$

which $n \in \{1, 2, \cdots, N\}$ and $c_N = 1$. So far, the best sum-rate of N-user GZIC by a class of simple HK schemes [12] was given in Theorem 1 of citation [13] as follow

$$R_{sum-inner} = \sum_{n=1}^{N} \gamma(c_n^2 P^*) \tag{5}$$

Substituting Eqs. (4) into (5), we can obtain the inner bound of sum rate

$$R_{sum-inner} = \begin{cases} \dfrac{N}{2}\gamma(h^2P^* + P^*),\ N \text{ is even} \\ \gamma(P^*) + \dfrac{N-1}{2}\gamma(h^2P^* + P^*),\ N \text{ is odd} \end{cases} \tag{6}$$

Secondly, two outer bounds of sum-rate will be given based on the two genies-aid methods as follow.

When N is even, the genie provides the side information $s_n = x_{n+1}, (n = 2i, 1 \leq i < N/2)$ to receiver n. So the receiver $n(n = 2i, 1 \leq i < N/2)$ can be free of interference by cancelling the interference from side information. Then the N-user GZIC evolves into a hybrid channel which is composed of $N/2$ Z interference channels. The sum-capacity of Z interference channel is $C_{sum-Z} = \gamma(h^2P^* + P^*)$. Therefore, one outer bound of sum-rate is

$$R_{sum-outer(even)} = \frac{N}{2}\gamma(h^2P^* + P^*) \tag{7}$$

When N is odd, the genie provides the side information $s_n = x_{n+1}(n = 2i, 1 \leq i < (N-1)/2)$ to receiver n. Since receiver $N-1$ eliminates the interference from the side information, the rate of user N is not limited by interference

link. Then the N-user GZIC evolves into a hybrid channel which is composed of $(N-1)/2$ Z interference channels and one point to point AWGN channel. Therefore, the second outer bound of sum-rate is

$$R_{sum-outer(odd)} = \gamma(P^*) + \frac{N-1}{2}\gamma(h^2P^* + P^*). \tag{8}$$

Since $R_{sum-inner(odd/even)} \leq C_{sum} \leq R_{sum-outer(odd/even)}$ and $R_{sum-inner} = R_{sum-outer}$, the formula (3) is the sum-capacity of N-user GZIC. □

Based on fairness, each user should share the sum-rate as equal as possible by rate-splitting [14]. So each user can achieve the same rate by the measure of splitting-rate, i.e. $R_Z = \frac{1}{2}\gamma(h^2P^* + P^*)$, when $1 \leq h^2 < P^* + 1$.

3.2 The Existing Capacity and Achievable Rate of Multi-user GZIC in Other Regimes

In N-user Gaussian Z very weak interference channel ($\frac{\sqrt{1+4P^*}-1}{2P^*} > h^2$), treating the interference as noise can achieve the capacity of N-user GZIC [15]. Therefore, the existing maximum achievable rate of each user is

$$R_Z = \gamma(SINR) \tag{9}$$

where $SINR = P^*/(1 + h^2P^*)$.

In N-user Gaussian Z weak and moderate interference channel ($\frac{\sqrt{1+4P^*}-1}{2P^*} \leq h^2 < 1$), By far, the best achievable rate R_{HK} [13] of multi-user GZIC with weak and moderate interference is

$$R_Z = R_{HK}. \tag{10}$$

In N-user Gaussian Z very strong interference channel ($P^* + 1 \leq h^2$), the interference links could be removed without affecting the capacity region [16]. Therefore, the existing maximum achievable rate of N-user GZIC is

$$R_Z = \gamma(P^*). \tag{11}$$

Based on formulas (3), (9), (11) and (10), the optimum multi-dimensional communication scheme is obtained as shown in the Table 1 below.

Table 1. Multi-dimensional communication scheme

Interference strength	Optimal transmission solutions	RSR
Very weak	Treat the interference as noise	$R_Z = \gamma(SINR)$
Weak and moderate	Rate-splitting and part interference cancelation	$R_Z = R_{HK}$
Strong	Rate-splitting and all interference cancelation	$R_Z = \frac{1}{N}C_{sum}$
Very strong	All interference cancelation without affecting the capacity	$R_Z = \gamma(P^*)$

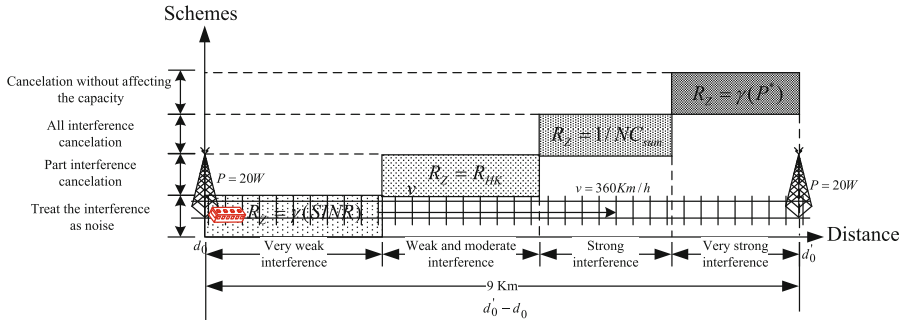


Fig. 3. Multi-dimensional communication scheme in simulation scenario.

4 Handover of HSR

In traditional system, reference signal receiving power (RSRP) is detected to make the decision of handover. Although Reference signal receiving quality (RSRQ) is applied as an auxiliary parameter to estimate the influence of interference, it cannot reflect the effect of interference cancelation. Therefore, we will use the optimal transmission solutions and the optimal rates above to construct a multi-dimensional communication scheme base on the rate-splitting and interference cancelation. In this scheme, the reference signal rate (RSR) is very important. The multi-dimensional communication scheme is shown in Table 1.

Similar with traditional system, handover margin (HOM) and Time-to-Trigger (TTT) are also needed to reduce the handover failure probability. We use the model of Cost231 [17] to simulate the high frequency and rural communication environment of HSR. The pass loss is $\mathrm{L_{HSR}} = 42.6 + 26 \lg f - 20 \lg d$, where f and d denote the frequency and the distance from the user equipment to original serving base station respectively. Let $f = 1800\,\mathrm{MHz}$, the transmit power of each base station is $P = 20\,\mathrm{W}$, the coverage of each base station is 5 km and the range of overlap between the two adjacent base stations is 1 km. We set $HOM = 0.5\,\mathrm{bits/s/Hz}$ and $TTT = 320\,\mathrm{ms}$. Considering the impact of Gaussian noise, we choose the $N_0 = 43\,\mathrm{dBm} + \mathrm{L_{HSR}}$ in the 5 km far from base station. It means the interference at the edge of cell has the same power of Gaussian noise.

The performance evaluation of multi-dimensional communication scheme is done by comparing it with the handover of traditional communication scheme. The scenario is that a train moving though the four areas (x-axis) interfered by a neighbor station has a fixed speed ($v = 360\,\mathrm{km/h}$) as shown in Fig. 3. We will choose the multi-dimensional communication scheme (y-axis) to improve the performance of communication system.

As the result of the simulation (Figs. 4 and 6), the multi-dimensional scheme can increase the rate and the coverage. Due to the rate-splitting of transmitter and the interference cancelation of receiver, the communication rate can be improved in weak, moderate, strong and very strong interference area. Meanwhile, the position of handover decision is extended.

4.1 Rate Increasing

In the fixed base station's arrangement, rate can be increased by multi-dimensional scheme. As Fig. 4 illustrates, d_t and d_m are the positions where the handover of traditional communication scheme triggered and the handover of multi-dimensional communication scheme triggered respectively. According to the criterion of event A3 triggering, $R(d_t)$ and $R_Z(d_m)$ are the lowest rates during the two communication process. We can easily find that the difference ΔR between $R(d_t)$ and $R_Z(d_m)$ is 1.2 bit. As shown in the Fig. 5, ΔR increases as frequency decreases. Therefore, the rate of HSR can be increased by multi-dimensional scheme especially in low frequency.

4.2 Coverage Increasing

Comparing with the traditional communication scheme, multi-dimensional scheme can extend the coverage of base station without decreasing the rate. The

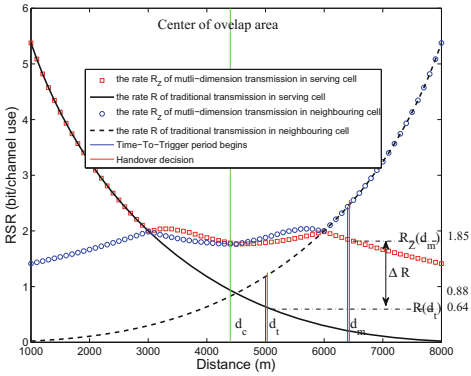


Fig. 4. Comparing the RSR of handover between traditional communication scheme and multi-dimensional communication scheme.

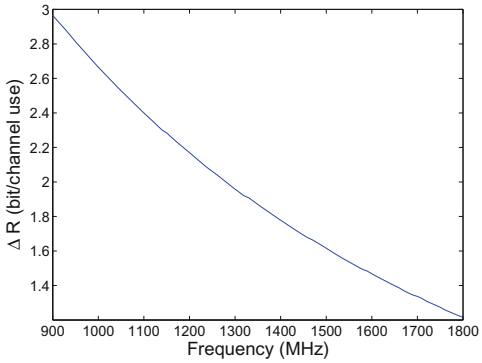


Fig. 5. ΔR versus frequency.

benefit of rate increasing by multi-dimensional scheme can be used to improve the coverage of base station. Let D_m denotes the position of handover decision, D_c and d_c denote the center position between the two adjacent base stations in multi-dimensional scheme and traditional communication scheme respectively. As shown in Fig. 6, the coverage of base station can be extended almost 5 km by multi-dimensional scheme. Meanwhile, $R_Z(D_m) = R(d_t)$. In Fig. 7, the difference ΔD between D_c and d_c increases as frequency decreases. Therefore, coverage of base station can be increased by multi-dimensional scheme in HSR and the frequency of handover can be decreased simultaneously. As shown in Fig. 8, during a 1800 km trip, users should endure 200 times of handover among the base stations deployed along the railway in traditional scheme. However, with the multi-dimensional scheme, as the coverage of each station rises, handover times are reduced obviously, especially in low frequency.

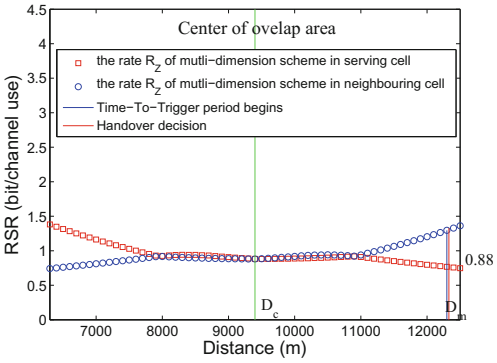


Fig. 6. The extended coverage of base station by multi-dimensional communication scheme.

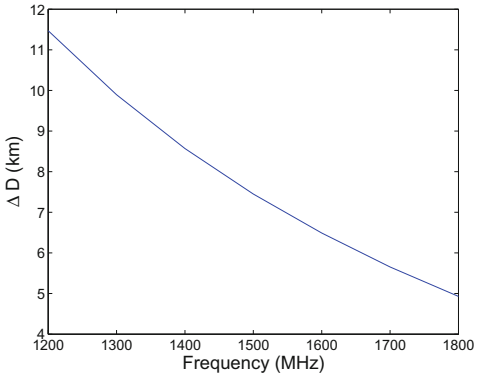


Fig. 7. ΔD versus frequency.

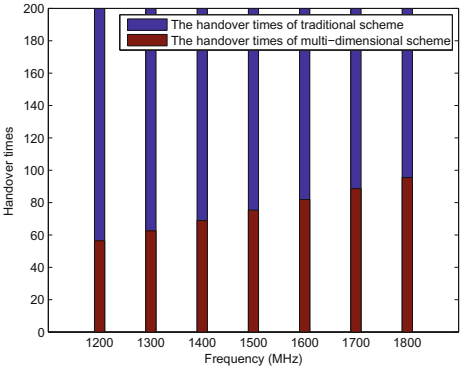


Fig. 8. Comparing the handover times between the traditional scheme and multi-dimensional scheme in a range of frequency.

5 Conclusion

We use the multi-user Gaussian Z interference channel model to represent the down link communication environment of HSR. The sum-capacity of multi-user Gaussian Z strong interference channel is obtained, which constructs a multi-dimensional communication scheme by the optimum rates and the optimum transmission solutions in all interference levels. The multi-dimensional communication scheme is applied into handover of HSR. By analyzing the result of simulation of multi-dimensional communication scheme, we find that the rate can be improved and the coverage of base station can be extended without decreasing the rate.

Acknowledgments. The authors would like to express their gratitude to the financial support of National Natural Science Foundation of China (Grant No. U1504601) and the education reform research project of Zhoukou Normal University (No. J2016015).

References

1. Ai, B., et al.: Challenges toward wireless communications for high-speed railway. IEEE Trans. ITs **15**, 2143–2158 (2014)
2. Wang, C.X., et al.: Channel measurements and models for high-speed train communication systems: a survey. IEEE Commmun. Surv. Tutor. **18**, 974–988 (2016)
3. Sun, N., Wu, J.: Maximizing spectral effciency for high mobility systems with imperfect channel state information. IEEE Trans. Wirel. Commun. **13**, 1462–1470 (2014)
4. Zhou, W., Wu, J., Fan, P.: High mobility wireless communications with Doppler diversity fundamental performance limits. IEEE Trans. Wirel. Commun. **14**, 6981–6992 (2015)
5. Chen, X., Ni, M.M.: Seamless handover for high mobility environments. In: International Wireless Communications and Mobile Computing Conference, pp. 281–286. IEEE Press, Cyprus (2016)

6. Ishii, H., Kishiyama, Y., Takahashi, H.: A novel architecture for LTE-B: C-plane/U-plane split and phantom cell concept. In: IEEE Globecom Workshops, pp. 624–630. IEEE Press, Anaheim (2012)
7. Cai, X., Giannakis, G.B.: Bounding performance and suppressing intercarrier interference in wireless mobile OFDM. IEEE Trans. Commun. **51**, 2047–2056 (2003)
8. Lu, Y., Xiong, K., Zhao, Z.Y., et al.: Remote antenna unit selection assisted seamless handover for high-speed railway communications with distributed antennas. In: 83rd IEEE Vehicular Technology Conference, pp. 1–6. IEEE Press, Nanjing (2016)
9. Xie, S.F., Yu, X.C., Luo, Y.: A seamless dual-link handover scheme with optimized threshold for C/U plane network in high-speed rail. In: Proceedings of the 83rd IEEE Vehicular Technology Conference, pp. 1–5. IEEE Press, Nanjing (2016)
10. Wu, J.X., Fan, P.Z.: A survey on high mobility wireless communications: challenges, opportunities and solutions. IEEE Access **4**, 450–476 (2016)
11. Liu, Z.Y., Zhou, E.Z., Fan, P.Z., Hao, L.: Tradeoff of capacity and handover performance in high speed railway wireless communications. In: 83rd IEEE the Vehicular Technology Conference, pp. 1–5. IEEE Press, Nanjing (2016)
12. Han, T.S., Kobayashi, K.: A new achievable rate region for the interference channel. IEEE Trans. Inf. Theory **27**, 49–60 (1981)
13. Liu, Y.P., Erkip, E.: On the sum capacity of K-user cascade Gaussian Z-interference channel. In: IEEE International Symposium on Information Theory, pp. 1382–1386. IEEE Press, Petersburg (2011)
14. Zhang, B., Li, Y., Bai, B.M., Zhou, S.H.: Decoding order algorithm based on rate-splitting in Gaussian interference channels. J. Electron. Inf. Technol. **35**, 1388–1394 (2013)
15. Sason, I.: On achievable rate regions for the Gaussian interference channel. IEEE Trans. Inf. Theory **50**, 1345–1356 (2004)
16. Carleial, A.B.: A case where interference does not reduce capacity. IEEE Trans. Inf. Theory **21**, 569–570 (1975)
17. European Cooperation in the Field of Scientific and Technical Research EURO-COST231: Urban Transmission LOSS Models for Mobile Radio in the 900 and 1800 MHz Bands, Revision 2, The Hague, September 1991

Satellite and Space Communications and Networking

A Novel Wavelet Denoising Algorithm Based on 2-Fold Cross Validation for BDS GEO Code Multipath Bias

Peng Wu(✉), Baowang Lian, Zesheng Dan, and Qiong Yang

Department of Electronics Engineering, Northwestern Polytechnical University, Xi'an, China
wupengrock@163.com

Abstract. The multipath bias is a non-ignorable factor that limits the BeiDou navigation satellite system (BDS) to improve positioning accuracy. Due to the relatively stationary geometry, the geostationary earth orbit (GEO) satellites not only have poor orbit accuracy, but also the multipath bias is difficult to eliminate by the existing approaches. In this paper, a new multipath error mitigation algorithm based on the wavelet and 2-fold cross validation methods are proposed, which can adaptive select the optimal decomposition level. The comparison of the proposed technique with the existing technique has been carried out with the simulated data series. In addition, a real date experiment shows that 17% to 57% improvement can be achieved in standard deviation compared with the raw multipath error.

Keywords: Wavelet filtering · 2-fold cross validation · Multipath bias
BeiDou navigation satellite system

1 Introduction

To date, China have accomplished the BeiDou regional navigation satellite system (BDS) that is especially consider as a valuable complement to the legacy systems for increasing coverage area and improving GNSS services in Asia-Pacific region [1]. The BDS constellation consists of 5 geostationary earth orbits (GEO) satellites, 5 inclined geosynchronous orbit (IGSO) satellites and 4 medium earth orbit (MEO) satellites, which is different from the other satellite navigation system, such as GPS, GLONASS, and Galileo. Because of GEO satellites being almost stationary relative to a static receiver on the Earth, the multipath interference may vary slowly along with time, even shows a likelihood 'DC offset' within a certain time zone. It is difficult to eliminate the multipath effect by merely increasing the observation period or averaging over multiple epochs. In addition, some scholars found that the multipath delay of GEO satellites have an extremely negative effect on code measurements [2], which eventually declines the positioning performance of BDS.

Due to the multipath effects are sensitive to the independent surroundings around the antenna, it is hard to achieve correlation for receivers at different locations and the multipath error cannot be eliminated by differential technique. Currently, there exists

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 297–306, 2018.
https://doi.org/10.1007/978-3-319-78130-3_30

three main techniques for mitigating the global navigation satellite system (GNSS) multipath effects as follows: antenna based techniques, such as choke-ring antenna, and array antennas; tracking loop techniques, such as the double delta correlator, multipath estimation delay lock loop (MEDLL) technique, and shaping correlator; data post-processing techniques, such as the sidereal filtering, adaptive method based on the signal-to-noise (SNR), empirical mode decomposition (EMD), and wavelet transform filtering [3].

In view of the good analysis performance both in time and frequency domains, the wavelet filtering has been widely used for signal extraction and denoising. [4] determines the optimal decomposition level by analyzing the relativity between the reconstructed signal and the original signal. But because there is no uniform rule for determine the correlation coefficient threshold, this method has a large amount of computation. According to the exponentially decaying properties of wavelet coefficients, the paper [5] designed an adaptive wavelet denoising threshold based on Bayesian rule. However, if the priori information is not accurate, it will lead to a poor denoising performance. Moreover, [6] studies the propagation characteristics of signal and additive white Gaussian noise in the wavelet decomposition process. However, this algorithm has a high requirement on sample size.

In this paper, we use the 2-fold cross validation method to assist the wavelet denoising algorithm to extract the low frequency components from the BDS GEO satellites multipath series. During this filtering procedure, the proposed method can automatically identify the denoising levels with less sample size and lower computation complexity.

2 Code Multipath Observations

As we know, the BDS receiver measures the distance to each satellite by tracking ranging code or carrier to calculate the localization of users. Meanwhile, the code observation ρ, and the carrier phase observation φ can be collected by the receiver. ρ and φ can be expressed as follow,

$$\rho = r + c(\delta t_u - \delta t^s) + I + T + M + \varepsilon_\rho \tag{1}$$

$$\phi = \lambda^{-1}(r + c(\delta t_u - \delta t^s) - I + T) + N + m + \varepsilon_\varphi \tag{2}$$

As shown in Eqs. (1) and (2), there are some common terms, such as receiver clock offset δt_u, satellite clock offset δt^s, troposphere delay T, and ionosphere delay I. Note that the ionosphere term is equal in magnitude but opposite in sign for the two measurements. In addition, λ is the wavelength of the carrier frequency, N is integer wavelength ambiguity, ε_ρ and ε_φ are the code and carrier tracking errors, respectively, M and m are the multipath delay in ρ and in φ, respectively. In the presence of multipath, tracking accuracy will be affected for both code- and carrier-tracking loops. Since the carrier center frequency is much higher than that of ranging code, carrier phase can be considered as a multipath-free observation, namely $M \gg m$. So the linear combination of single-frequency code and dual-frequency phase observations are

generally adopted to analyze the characteristics of code multipath error. And this combination is called multipath observation (*MP*). The *MP* combination on frequency i can be expressed as [2]

$$MP_i = \rho_i - (\eta_{(i,j)} + 1) \cdot \lambda_j \cdot \varphi_j + \eta_{(i,j)} \cdot \lambda_j \cdot \varphi_j + B_i \tag{3}$$

In which,

$$\eta_{(i,j)} = \frac{2f_j^2}{f_i^2 - f_j^2} \tag{4}$$

$$B_i = (\eta_{(i,j)} + 1) \cdot \lambda_j \cdot N_j - \eta_{(i,j)} \cdot \lambda_j \cdot N_j + \xi \tag{5}$$

where the linear factors $\eta_{(i,j)}$ is calculated by the frequency of the carrier f, and the subscripts i, j ($i, j = 1, 2, 3, i \neq j$) are used to denote different frequencies. The bias term B_i mainly contains the linear combination of the carrier phase ambiguities, the combination of phase multipath errors, and the constant term of hardware and software induced errors ξ. Assuming no cycle clips, B_i is a constant, and it can be deemed as the mean value of *MP*. However, in practice, the carrier cycle clips are difficult to be avoided, since the navigation signals are disturbed by various noise, which results in that the carrier PLL cannot track signal continuously. Therefore, it is significant to detect and restore the cycle clips for every continuous tracking sections before we use the multipath observation.

According to the above derivation for *MP*, the noises in code and carrier phase measurements cannot be eliminated by the linear combination. Therefore, the noise still exists in the *MP*. In order to study characteristics of the multipath error, it is essential to separate the *MP* signal from the noise. In the following, we will introduce the denoising algorithm.

3 Wavelet Threshold Denoising Algorithm

Threshold denoising in wavelet domain is an efficient algorithm to reduce the white noise in digital signal. Since the signal and noise have different relevance with the wavelet function, there will be different attenuations in wavelet decomposition process. In practical application, the Mallat algorithm based on multi-resolution analysis is often utilized to decompose the signal in $L^2(R^2)$ space [7]. As the Mallat algorithm can make the length of signal decomposition reduced by half each time, it is also a fast wavelet algorithm, which can be expressed as

$$\begin{cases} w_\varphi(j,k) = \sum_k s_{j-1}(k)\varphi(2n-k) \\ w_\psi(j,k) = \sum_k s_{j-1}(k)\psi(2n-k) \end{cases} \tag{6}$$

where $s(k)$ is the original signal. $\varphi(k)$ and $\psi(k)$ are corresponding with the low-pass filter of the scaling function and the high-pass filter of the wavelet function, respectively. $w_{\varphi}(j,k)$ is the scale coefficients and it is also the coefficients of signal's approximate part, $w_{\psi}(j,k)$ is the wavelet coefficients and it is also the coefficients of signal's details part. The diagram of three layers wavelet decomposition is shown in Fig. 1.

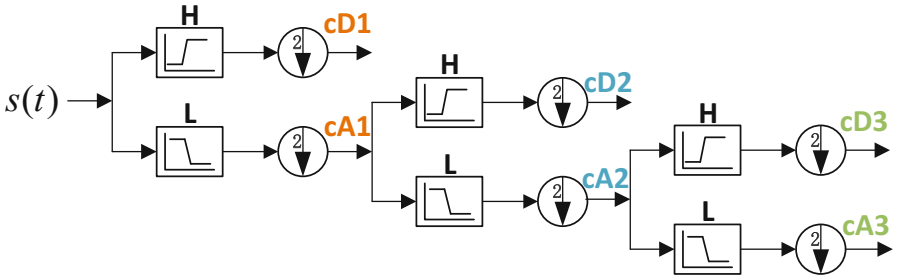


Fig. 1. The diagram of three layers wavelet

The noisy signal is transformed to the detail signal and the approximate signal by the Eq. (6). The Noise signal is mainly concentrated in the wavelet coefficients $w_{\psi}(j,k)$ with lower energy, such as cD1, cD2 and cD3 shown in the Fig. 1; the meaningful signal is mainly condensed in the scaling coefficients $w_{\varphi}(j,k)$ with higher energy, such as cA3 shown in the Fig. 1.

Generally, the denoising process of one-dimensional signal includes four steps: selecting wavelets bases, selecting the decomposition level, selecting the threshold function, and reconstructing signals. This paper will focus on determining an appropriate decomposition level.

4 Adaptive Wavelet Threshold Algorithm Based on 2-Fold Cross Validation

As mentioned above, if the layer is too shallow, the scaling coefficients contain much more noise; if the layer is too deep, the meaningful signal will be destroyed. So how to select a suitable wavelet decomposition layer directly affects the denoising performance. This section will introduce a new adaptive mitigation threshold technique based on 2-fold cross validation.

4.1 2-Fold Cross Validation

Cross validation is a model validation technique for assessing how the regression model fit to the original dataset. This method can be also used to compare the performance of two or more different regression algorithm and find out the best one for the available data. 2-fold cross validation is one of the basic form of cross validation.

In 2-fold cross validation the data is randomly divided into 2 roughly equal sized subsamples. One segment is tread as a training set to build the model, and the remaining segment is used for validation. Meanwhile, an error detection function is introduced to evaluate the performance of the different models, and then to find the best fit one. We take a quadratic regression model and a cubic spline interpolation model as an example to illustrate the core idea of the 2-fold cross validation.

The raw data shown in the Fig. 2(a) is sampled from the signal $s(n) = \sin(n) + \varepsilon(n)$, and the length of the data n is 40, SNR is 11 dB. Firstly, the raw data is divided into two part, illustrated in the Fig. 2(b). The odd series is training set, and the even is validation set. Subsequently, the quadratic regression model and the cubic spline interpolation model are used to fit the training set, and obtain two curves corresponding to these two models (red line in Fig. 2(c) and (d)). Then the validation set is employed to test the matching degree of the model. During the processing, the error detection function is mean squared error (MSE). According to the Fig. 2(c) and (d), the cubic spline interpolation model have a better fitting performance than the quadratic regression model.

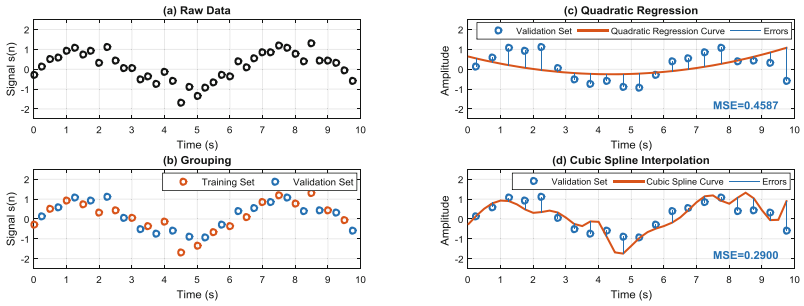


Fig. 2. The procedure of the 2-fold cross validation (Color figure online)

During the simulation, we also find that the matching degree of the cubic spline interpolation model is directly proportional to the SNR. Table 1 shows the relationship between the standard deviation of error and SNR.

Table 1. The relationship between the standard deviation of error and SNR of the cubic spline interpolation model

SNR(dB)	8	9	10	11	12	13	14	15
Std(Err)	1.2427	0.6701	0.4756	0.3672	0.3835	0.2949	0.2719	0.3398

4.2 Adaptive Selection of Decomposition Threshold

We proposed an adaptive threshold selection method with 2-fold cross validation. For the Mallat wavelet algorithm, if the level is too deep, it will make the useful signal loss some information. In case of this situation, FFT is also employed. The flow of the algorithm is as follows.

Step 1: To analyze the frequency spectrum of the raw multipath time series by using FFT, and recode these frequency components.

Step 2: To analyze the raw multipath time series by Mallat wavelet algorithm, and set the decomposition level k = 1.

Step 3: To abstract the approximate coefficients $s_k(n)$ and separate the data $s_k(n)$ into training set $s_{odd,k} = s_k(2n+1)$ and validation set $s_{even,k} = s_k(2n)$, n = 1, 2, 3, …, N/2. N is the length of the $s_k(n)$.

Step 4: In 2-fold cross validation method, the cubic spline interpolation model $f_{cb,k}(n)$ is applied, and then assess the variance of the $f_{cb,k}(n)$ relative to $s_{even,k}$ by using the following error detection function:

$$C(k) = \frac{1}{N}\sum_{n=1}^{N}[s_{even,k}(n) - f_{cb,k}(n)]^2, \quad n = 1, 2, 3 \ldots, N \tag{7}$$

Step 5: To calculate the FFT for the $s_k(n)$. If the main frequency components are NOT same with the that of raw multipath time series, the program will be stopped. Then find out the minimum value of $C(k)$, the corresponding level is the optimal level. Otherwise, set k = k + 1, and go to the Step 3.

Using the above procedures, the optimal decomposition level can be obtained for the wavelet transform with low computation complexity, and small size of the data.

5 Simulation and Analysis

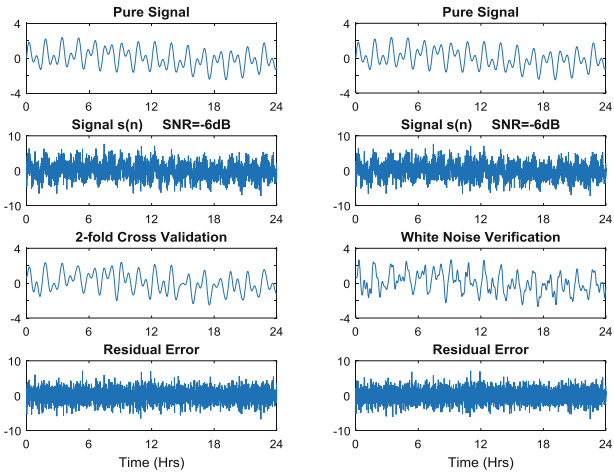
5.1 Simulation with Test Signal

The test signal $s(n)$ with three different frequencies sinusoidal wave can be expressed as the following:

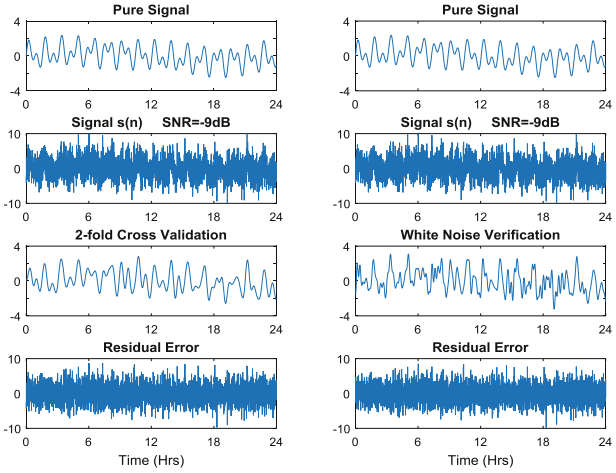
$$\begin{aligned} s(n) &= s_0(t) + \varepsilon(n) \\ &= \sin(2\pi n/2900) + \sin(2\pi n/5400) + 0.5\sin(2\pi n/86400) + \varepsilon(n) \end{aligned} \tag{8}$$

where $\varepsilon(n)$ is Gaussian white noise. 1/2900, 1/5400 and 1/86400 are the typical BDS GEO satellites multipath fading frequency. Figure 3 shows the comparison results of the proposed method and the white noise verification method at different SNR [8]. In this simulation, the soft threshold function is employed.

In each subgraph of Fig. 3, there are 8 waveforms arranged into 4 rows and 2 columns. The top row is the pure signal $s_0(t)$, the second row is the pure signal polluted by noise $s(t)$, the third row is filtered waveforms by the proposed method and white noise verification, respectively, and the bottom row shows residuals between the test signal $s(t)$ and the filtered values. After comparing the waves in the top and third row, it is clear that the 2-fold cross validation have a better denoising capability than white noise verification method.



(a) SNR is -6dB



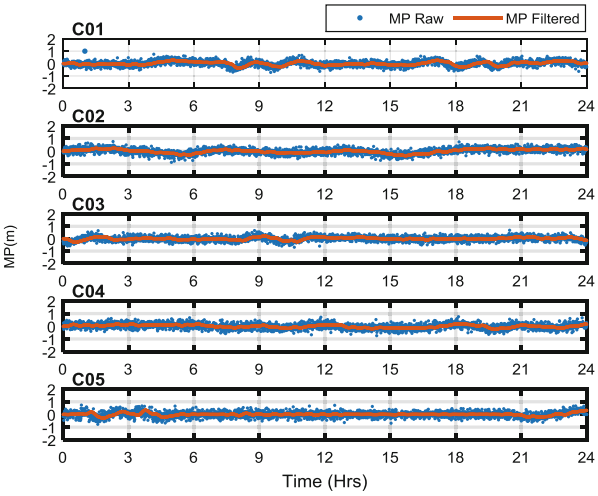
(b) SNR is -9dB

Fig. 3. Comparison of two kinds of wavelet denoising algorithms under different SNR

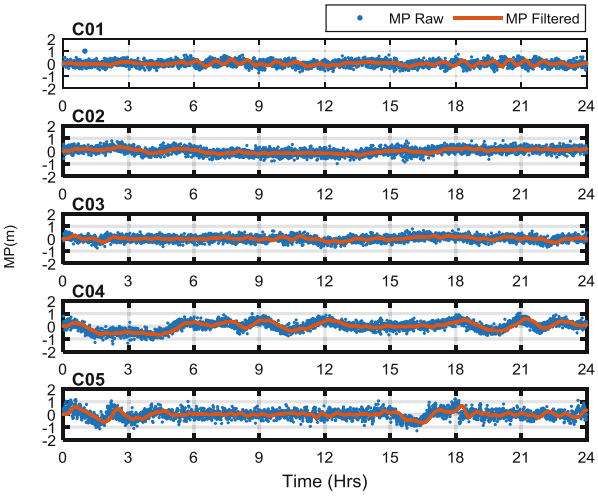
5.2 Simulation with Real BDS GEO Multipath Error Signal

The real BDS GEO measurements were collected by IGS MGEX (International GNSS Service Multi-GNSS Experiment). There are hundreds of observation sites under IGS around the world. In this simulation, we analyze the data from KARR site and JFNG site. These two stations are both equipped with Trimble NetR9 receiver connected with a Trm59800.00 None antenna.

According to Eq. (3), we obtain the MP2 series from the 5 GEO satellites. The 2-fold cross validation method is applied to assist the Mallat wavelet to mitigate the raw multipath series. And the results are shows in Fig. 4.



(a) JFNG



(b) KARR

Fig. 4. The mitigation results of multipath series

It is easily seen from the Fig. 4 that the noise levels of filtered multipath series are significantly decreased compared with the raw series. And the standard deviations of raw MP and filtered MP for five GEO satellites at these two stations are illustrated in Fig. 5. As shown in Fig. 5, the improvements range are from 17% to 57%.

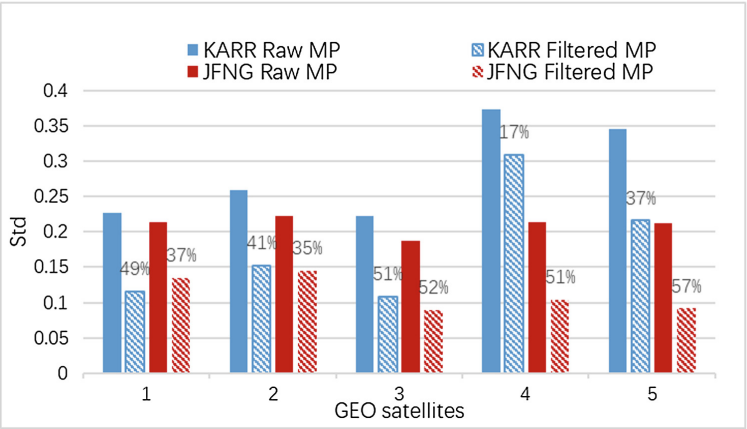


Fig. 5. The standard deviations of raw and filtered MP at two stations

6 Conclusion

In this paper, we proposed a new adaptive selection threshold level for wavelet by using 2-fold cross validation method. Simulation results show that the 2-fold cross validation have a better denoising capability than white noise verification method. In addition, we take the proposed algorithm to mitigate the real MP time series of JFNG and KARR stations. According to results, the standard deviation values of filtered MP achieve 17% to 57% improvements.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (contract no. 61301094) and BeiDou Special Fund (contract no. GFZX0301040115).

References

1. Montenbruck, O., Hauschild, A., Steigenberger, P., Hugentobler, U., Teunissen, P., Nakamura, S.: Initial assessment of the COMPASS/BeiDou-2 regional navigation satellite system. GPS Solut. **17**, 211–222 (2013)
2. Wang, G., de Jong, K., Zhao, Q., Hu, Z., Guo, J.: Multipath analysis of code measurements for BeiDou geostationary satellites. GPS Solut. **19**, 129–139 (2015)
3. Wang, X., Liu, W., Li, B., Sun, G.: The application of zero phase filter on BDS geostationary satellite code multipath mitigation. In: International Conference on Computer Science and Network Technology (2015)
4. Chunli, W., Chunlei, Z., Pengtu, Z.: Denoising algorithm based on wavelet adaptive threshold. Phys. Procedia **24**(Pt. A), 678–685 (2012)
5. Chen, Y., Ji, Z., Han, C.: Adaptive wavelet denoising threshold for based on bayesian rule. J. Optoelectron. Laser **19**, 120–124+134 (2008)

6. Du, W., Zhu, R., Li, Y.: Adaptive selection of optimal decomposition level in filtering algorithm based on wavelet transform. J. Optoelectron. Laser **21**, 1408–1411 (2010)
7. Huimin, C., Ruimei, Z., Yanli, H.: Improved threshold denoising method based on wavelet transform. Phys. Procedia **33**, 1354–1359 (2012)
8. Liu, X., Liu, X.: A new wavelet threshold denoising algorithm in speech recognition. In: Conference on Information Processing (2009)



Multi Block Overlapping Zero Padding Algorithm for Weak Signal Acquisition

Jinhua Sun(✉), Shanshan Lv, and Hao Wang

State Key Laboratory of Integrated Service Networks,
Xidian University, Xi'an, China
jhsun@xidian.edu.cn

Abstract. Global navigation satellite systems (GNSS) can implement high-precision navigation and positioning, requiring rapid and accurate signal acquisition under very weak signal conditions. Double block zero padding (DBZP) is an efficient algorithm for unaided weak signal acquisition. Considering the acquisition problem of weak direct sequence spread spectrum continuous phase modulation signal, the subblock combining and zero padding method is studied in detail. To solve the weakness of fixed number of subblock and fixed number of signal samples in each subblock given frequency resolution and Doppler searching range in DBZP, a multi block overlapping zero padding (MBOZP) algorithm is proposed, which combines with the ideas of data subblock overlapping. This algorithm optimizes the combination of data subblock by introducing the data subblock overlapping, which can increase the length of coherent accumulation and improve the performance of acquisition under the determined frequency resolution and Doppler searching range. Compared with the traditional DBZP, the simulation results show that MBOZP can obtain higher acquisition probability at small frequency offset.

Keywords: GNSS · Acquisition · Circular correlation
Continuous phase modulation

1 Introduction

High-precision navigation satellite systems play an important role in military fields and are widely used in various fields [1]. Phase shift keying commonly used in satellite communications was limited by power efficiency and bandwidth efficiency, however, continuous phase modulation [2] can solve this problem effectively. With the development of navigation technology, it has important significance to implement weak CPM signal acquisition under the harsh environment.

In traditional acquisition algorithms, serial search method [3] obtains signal acquisition by two-dimensional search in carrier frequency offset and local code phase; parallel frequency domain search method transforms two-dimension searching into one-dimension searching, which can raise calculation speed and decrease acquisition time; parallel code phase search method [4] changes time-domain convolution operation into frequency-domain multiplication, with one IFFT operation to calculate the correlation values at all code delays. In general, traditional acquisition algorithms exist

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 307–316, 2018.
https://doi.org/10.1007/978-3-319-78130-3_31

the problem of huge computational complexity. Therefore, Lin and Tsui [5] proposed the DBZP based on traditional acquisition algorithms, with the advantages of fast acquisition and high sensitivity. DBZP uses bi-frequency domain Fourier transformation for circular correlation, and the long correlation integral was divided into several blocks with the same length to perform correlation integral. In recent years, various improved algorithms based on DBZP were proposed, such as MDBZP [6], FMDBZP [7], IFMDBZP [8], DBZPTI [9] and restructured DBZP [10]. These improved algorithms mainly consider the decrease of the detection probability caused by date bit transition [11], and most of the algorithms focus on simplifying the detection decision and optimizing the frequency domain transform of correlation matrix.

Considering the influence of Doppler frequency offset on the length of data subblock, DBZP adopts fixed number of data subblock and fixed length of data subblock in order to guarantee the Doppler search range and frequency resolution. In terms of Doppler search range, it keeps the length of data subblock not too long to search the maximum Doppler frequency offset, whereas short data block will degrade the detection performance and increase the computation under the small or moderate frequency offsets. Thus, acquisition sensitivity can be increased by increasing the length of data subblock. The influence of the subblock combining and zero padding method on acquisition sensitivity in DBZP is further studied. When the new data subblocks are formed respectively by the received signal and the local signal, the overlapping multiple subblocks are introduced to increase the length of each new data subblock, which the number of data subblock remains unchanged. So the increase in the length of the coherent accumulation can improve the acquisition performance.

2 System Model

The system model of the direct sequence spread spectrum continuous phase modulation is shown in Fig. 1. The discrete time baseband signal arrived at the receiver can be expressed as

$$r_n = s(nT_s - \tau)\exp(j2\pi f_d nT_s) + w_n \tag{1}$$

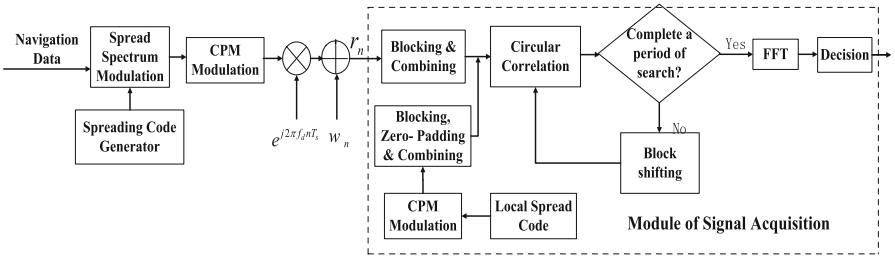


Fig. 1. Acquisition model of DSSS CPM system

where r_n is the nth sample of the received signal, $s(nT_s - \tau)$ is the complex baseband DSSS-CPM signal. τ is the code phase delay, T_s is the sampling interval ($f_s = 1/T_s$ is the sampling frequency), f_d is the Doppler frequency offset, w_n is the additive Gaussian white noise sample with power spectral density N_0.

The acquisition module is mainly to process the received complex baseband signal and the local spread spectrum signal, and get the cross-correlation function of them. Then the estimated values of code phase and Doppler frequency offset are obtained according to its strong correlation peak of the correlation values.

When the DBZP algorithm is applied, the parameters of algorithm need to be determined by the system requirement at first. We assume that the correlation integral time $T = 20$ ms, the sampling rate $f_s = 1.024$ MHz, and the Doppler frequency search bandwidth $B_{f_d} = 8$ KHz, which means the search range is $(-4\text{ KHz} \sim 4\text{ KHz})$. Hence the total data samples $N_{\text{total}} = f_s T = 20480$, the frequency resolution $f_{res} = f_s/N_{\text{total}} = 1/T = 50$ Hz, the total number of data block $N_s = B_{f_d}/f_{res} = 160$ and the length of each data block $L_s = N_{\text{total}}/N_s = 128$. Because of the periodicity of spread sequence, it just requires completing a period search for spread sequence when the data block is shifted in DBZP. The number of data block contained in 1 ms $N_{ms} = N_s/(T/1\text{ ms}) = 8$.

3 Multi Block Overlapping Zero Padding

The combination method of data subblock is discussed in this paper. First of all, the principle of DBZP is introduced so as to better understand the MBOZP algorithm.

3.1 Double Block Zero Padding

In DBZP, the search of all code phase is completed by circular correlation within the data subblock and the data subblock shifting. Then the correlation value matrix corresponding to different Doppler shift is obtained using FFT, which is equivalent to the parallel frequency search. The core of DBZP lies in double blocking and zero padding, which can ensure the search of data fully and without repetition, and avoid the loss of power caused by long data, and the FFT/IFFT operation of short data subblock in circular correlation can reduce the computation complexity. So DBZP not only reduces the influence of Doppler frequency offset to the correlation values but also reduces the computational complexity greatly. The algorithm integrated with the advantages of parallel frequency domain search and parallel code phase search. The main procedures of this algorithm are listed as follows:

Step1. Blocking, Double blocking and Zero padding

Blocking: The received signal and the local signal with the correlation integral time of T, are divided into N_s blocks respectively and each block contains L_s data samples.

Double blocking: The adjacent two data blocks of the received signal are combined to form a new data block, that is, the ith block and the $i+1$th block are combined into the new ith data subblock $(i = 1, 2, 3, \cdots\cdots, N_s)$. Each new data subblock contains $2L_s$ data samples.

Zero padding: The new N_s local spreading code subblocks are formed by filling the same length of zero after each data block of local signal, each new local spreading code subblock contains $2L_s$ data samples.

Step2. Circular correlation

Circular correlation within a subblock: The data subblock of the received signal is circular correlated with the corresponding subblock of local signal, the operation within the subblock is equivalent to the parallel code phase search. The length of subblock for circular correlation is $L = 2L_s$. The results of circular correlation only retain the first L_s points data in the $2L_s$ points.

Subblock shifting: The data of the received signal is block shifted to the left, and then the data subblock of the received signal after shifting is circular correlated with the subblock of local signal, the search is completed when the received signal is shifted $N_{ms} - 1$ times. Finally, a time domain correlation value matrix with the size $N_s \times (L_s \times N_{ms})$ is obtained, in which N_s stands for possible Doppler frequency offsets and $L_s \times N_{ms}$ stands for all possible code phases.

Step3. Detection and decision

After FFT, the time domain correlation value matrix is transformed into the frequency domain correlation value matrix, the energy of all points in the matrix is calculated. The signal is acquired precisely if the maximum in the matrix is greater than the given threshold and the estimated code phase is within a symbol, otherwise acquisition is failed.

3.2 Multi Block Overlapping Zero Padding

The length and the number of the blocks in traditional DBZP are determined by the frequency resolution and the range of Doppler frequency offset. It generally uses the smaller length of block in order to give consideration to the large Doppler frequency offset, however, for the smaller Doppler frequency offset, the shorter circular correlation data will make the acquisition performance lower than the longer block. The MBOZP is proposed in this paper keeping the number of blocks fixed, compared with original DBZP. It can properly increase the length of data subblock so as to obtain higher processing gain. This requires the introduction of data block overlapping, as shown in Fig. 2.

For MBOZP, the number of block is still N_s and the length of each block is still L_s, the length of data subblock for circular correlation is $L = 2L_s$ when the data overlapping is L_s. The new data subblock of the received signal is obtained by combining the adjacent three blocks and the new subblock of local spread signal is obtained by combining the adjacent two blocks and then added a block of zero with length L_s, so the length of new subblock is $3L_s$; it is required to perform the FFT/IFFT transform with $3L_s$ points in circular correlation within a subblock, only the first L_s points in the final $3L_s$ points are saved.

As shown in Fig. 3, this method is equivalent to the parallel frequency domain search when the section is also overlapping.

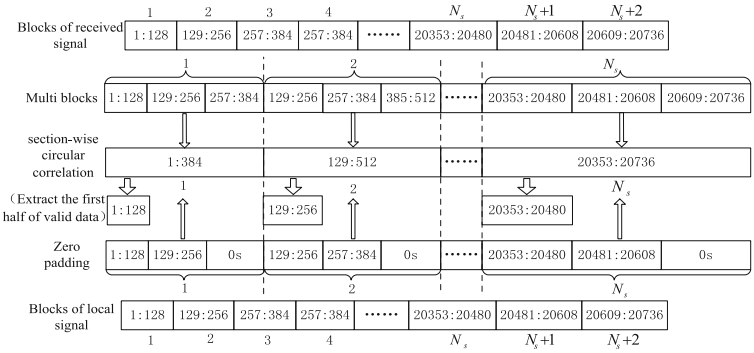


Fig. 2. Block and zero padding in MBOZP ($L_s = 128$)

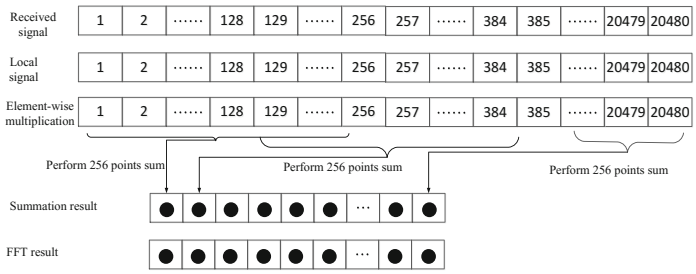


Fig. 3. Data subblock overlap correlation process in time domain

4 The Performance Analysis of MBOZP

The correlation result of MBOZP is equivalent to the parallel frequency search with data overlapping. The acquisition algorithm based on matched filter is on the basis of parallel frequency search, designing a kind of transversal filter by sequence correlation to accomplish the correlation operation and using the FFT to reduce the time of acquisition. The matched filter is divided into a set of short matched filter in PMF-FFT [12], which can reduce the performance of correlation power and decrease the hardware overhead. The blocking and correlation in MBOZP is equivalent to the segmentation and matching in PMF-FFT, thus, the performance analysis of MBOZP can be studied based on the performance of PMF-FFT.

The received signal is shown as (1). Assuming that the noise is ignored, when the received signal is aligned with the local signal, $s(nT_s - \tau)s^*(nT_s - \hat{\tau})$ is regarded as 1, so the normalized correlation output value of the received signal and the local signal can be expressed as

$$G_{DMF} = \frac{1}{N_{total}} \sum_{n=0}^{N_{total}-1} r_n s^*(nT_s - \hat{\tau}) = \frac{1}{N_{total}} \left| \frac{\sin(\pi f_d N_{total} T_s)}{\sin(\pi f_d T_s)} \right| \exp[j\pi f_d (N_{total} - 1) T_s] \quad (2)$$

The data with N_{total} samples is divided into N_s block, each data block has the length L_s so the correlation output of the kth block is shown as,

$$G_{PMF}(k) = \frac{1}{L_s}\sum_{n=(k-1)L_s}^{N_sL_s-1} r_n s^*(nT_s - \hat{\tau}) = \frac{1}{L_s}\left|\frac{\sin(\pi f_d L_s T_s)}{\sin(\pi f_d T_s)}\right| \exp[j\varphi]\exp[j\pi f_d(k-1)L_sT_s] \quad (3)$$

where $k = (1, 2, 3, \cdots, N_s)$, $\varphi = \pi f_d(L_s - 1)T_s$ represents the phase characteristic of PMF.

Determining N_s-point FFT for $R(k)$, the output of the normalized frequency response for lth points is

$$\begin{aligned} G_{PMF-FFT}(l) &= \frac{1}{N_{total}}\sum_{k=0}^{N_s-1} G_{PMF}(k)\exp[-j2\pi kl/N_s] \\ &= \frac{1}{N_{total}}\frac{\sin(\pi f_d L_s T_s)}{\sin(\pi f_d T_s)}\frac{\sin(\pi f_d N_s L_s T_s - \pi l)}{\sin\left(\pi f_d T_s L_s - \pi\frac{l}{N_s}\right)}\exp[j\Phi] \end{aligned} \quad (4)$$

where $l = (1, 2, 3, \cdots, N_s - 1)$, $\Phi = \pi(N_s - 1)(f_d L_s T_s - l/N_s)T_s - \pi f_d(L_s - 1)T_s$ represents the phase characteristic of PMF-FFT.

Accordingly, the final normalized amplitude frequency response is

$$G_{PMF-FFT}(f_d, l) = \left|\frac{\sin(\pi f_d L_s T_s)}{L_s \sin(\pi f_d T_s)}\right| \left|\frac{\sin(\pi f_d N_s L_s T_s - \pi l)}{N_s \sin\left(\pi f_d L_s T_s - \pi\frac{l}{N_s}\right)}\right| = G_1(f_d, l) * G_2(f_d, l) \quad (5)$$

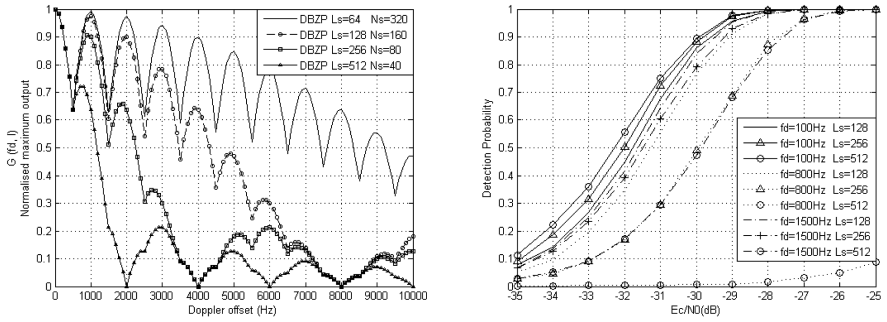
where $G_1(f_d, l) = 1/L_s|\sin(\pi f_d L_s T_s)/\sin(\pi f_d T_s)|$ is the attenuation which is caused by the process of correlation, the output correlation value shows significant attenuation when the larger Doppler frequency offsets exist in received signal. $G_2(f_d, l) = |\sin(\pi f_d N_s L_s T_s - \pi l)/N_s \sin(\pi f_d L_s T_s - \pi(l/N_s))|$ is caused by incomplete phase compensation in FFT operation, which gets the maximum value when $l = L_s f_d N_s T_s$. This indicates that the acquisition sensitivity is decided by the length and the number of data blocks, when the length of data block is fixed, increasing number of data block can improve the estimation accuracy of Doppler frequency offset and code phase.

5 Simulation Analysis

5.1 The Influence of the Length of Data Block and Doppler Frequency Offset

The search range of Doppler frequency offset changes when the number of data block changes in DBZP, so the acquisition sensitivity is influenced by f_d, L_s, and N_s. The influence of different numbers of data block acquisition sensitivity under different frequency offsets is given in Fig. 4. The search range $B_{f_d} = (-4\,\text{KHz},\ +4\,\text{KHz}), L_s = 128$. The search range of Doppler frequency will halve while the length of block doubles.

Figure 4 presents the influence of changing block length on acquisition performance under different frequency offsets. Figure 4(a) is the normalized gain of correlation output at different block length. It is seen that the larger Doppler frequency offsets will cause the attenuation of the output correlation value, when the block length is increasing, the search range of the corresponding frequency offsets will narrow. The detection probability under different block length is given in Fig. 4(b), where f_d is 100 Hz, 800 H and 1500 Hz. We can see that when the frequency offset is small, the higher detection probability can be obtained with longer block length, however, with the increase of the frequency offset, the long data correlation is affected by the Doppler offset. In summary, in the choice of the length and numbers of data block, the search range and Doppler frequency offset should be considered first. We can choose the long data block with small correlation loss when the frequency offset is small.



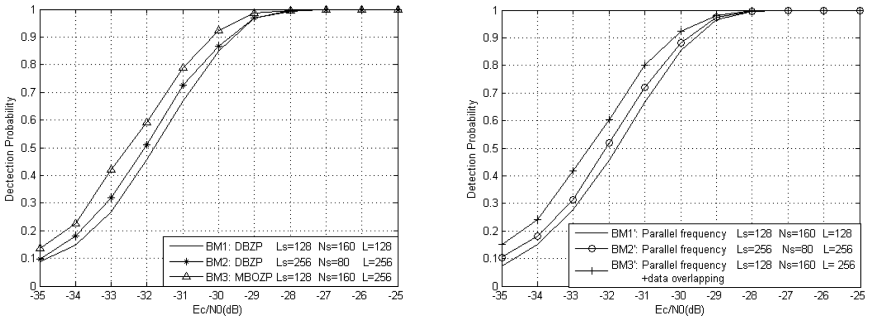
(a) (b)

Fig. 4. The influence of block length on acquisition performance under different frequency offset

5.2 The Blocking Method

The detection probabilities under the different blocking methods are compared in Fig. 5, where L_s is the block length, N_s is the number of data block and L is the length of data subblock for circular correlation which is equivalent to the length of piecewise summation in PMF-FFT. Figure 5(b) shows the detection probability corresponding to frequency parallel search algorithm. It can be seen that the correlation result of DBZP is absolutely equivalent to the frequency parallelize search from Fig. 4(a), we have seen that the correlation value is less vulnerable to small frequency offsets, the acquisition sensitivity is higher when the block length is longer, thus, the acquisition sensitivity can be improved by increasing the block length when the Doppler frequency offset is small, such as $f_d = 100$ Hz.

It can also be seen from Fig. 5 that the acquisition sensitivity of MBOZP is the highest. From the normalized amplitude frequency response we see that acquisition sensitivity is codetermined by N_s, L_s and f_d. We assume that f_d is 100 Hz, for BM3 (MBOZP with $L_s = 128$, $N_s = 160$ and $L = 256$) and BM1(DBZP with $L_s = 128$, $N_s = 160$ and $L = 128$), the numbers of data block are the same, but the length of data



(a) (b)

Fig. 5. The detection probability under the different block method

subblock for MBOZP increases, so its detection probability is higher. The length of data subblock for BM2 and BM3 remains the same, but the number of data block for BM3 increases because of the data overlapping, so that the MBOZP gets higher detection probability. From above, the acquisition sensitivity of MBOZP is improved by 1 dB at small frequency offsets.

Figure 6 is the acquisition performance of MBOZP and DBZP under different frequency offsets.

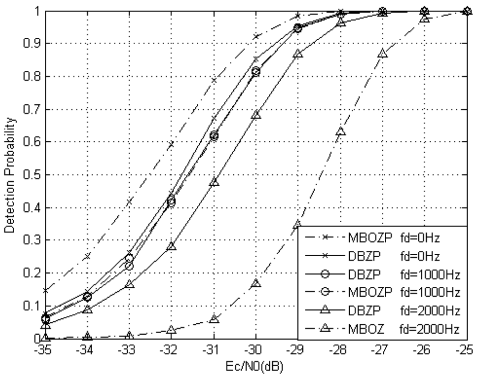


Fig. 6. The influence of frequency offset on acquisition performance

We can see that the performance of MBOZP is better than DBZP at small frequency offsets, however, the acquisition performance of MBOZP decreased faster than DBZP by increasing the frequency offsets. In other words, MBOZP is sensitive to frequency offset. This indicates that MBOZP is more suitable for fine estimation with small frequency offset. In practice, we can use DBZP to estimate the frequency offset firstly and obtain the coarse estimation of frequency offset, then, we can use MBOZP to improve the acquisition performance while the frequency offset is small.

5.3 Complexity Analysis of MBOZP

In this section we will analyze the computation complexity. Parallel frequency search algorithm includes multiplication, piecewise summation and FFT, and the algorithm shown in Fig. 3 adds data overlapping based on parallel frequency. DBZP includes double blocking, zero padding, FFT and IFFT, MBOZP changes the combination of data subblock, the overlapping data makes the length of data subblock for circular correlation increase.

Table 1 shows the computational complexity of several acquisition algorithms based on real multiplication and real addition.

Table 1. Computational complexity of several acquisition algorithms

Algorithm	Real addition	Real multiplication
Parallel frequency search	$2(2L_s - 1)N_s f_s T_s$	$4N_s L_s f_s T_s$
DBZP	$6(2 + N_{ms})N_s L_s \log_2^{2L_s} + 2N_{ms}N_s L_s$	$4(2 + N_{ms})N_s L_s \log_2^{2L_s} + 4N_{ms}N_s L_s$
Parallel frequency search with data overlapping (shown in Fig. 3)	$2(3L_s - 1)N_s f_s T_s$	$4N_s L_s f_s T_s$
MBOZP (shown in Fig. 2)	$9(2 + N_{ms})N_s L_s \log_2^{3L_s} + 4N_{ms}N_s L_s$	$6(2 + N_{ms})N_s L_s \log_2^{3L_s} + 8N_{ms}N_s L_s$

We assume that $f_s = 1.024\,\text{MHz}$, $N_s = 160$, $L_s = 128$, $N_{ms} = 8$, therefore, the real addition of parallel frequency search algorithm is 8.36e7 and the real multiplication is 8.39e7; if introduced the data overlapping in this algorithm, its real addition becomes 1.26e8 and its real multiplication remain the same. The real addition and real multiplication of DBZP are 1.02e7 and 7.21e6 respectively; while for MBOZP, its real addition becomes 1.65e7 and its real multiplication becomes 1.19e7.

Compared with parallel frequency search algorithm, when the data overlapping is introduced, its real addition rises by 33.7% and its real multiplication keeps unchanged. Its real addition and real multiplication of MBOZP rise by 38.2% and 39.4% compared with DBZP. This shows that the data overlapping will cause the increase of computation.

Compared to the capture algorithm in single frequency domain, the computation of MBOZP and DBZP decrease significantly. When $L_s = 128$, compared with the parallel frequency search algorithm, the real addition and real multiplication computation of DBZP are reduced by 87.8% and 91.4%, that of MBOZP are reduced by 86.9% and 85.8%, which means multi block combination acquisition algorithm in dual frequency domain has lower computation complexity.

From the foregoing, the acquisition sensitivity of algorithm based on DBZP is higher than the algorithm in single domain frequency, and its computation is lower; the smaller frequency offsets lead to the longer block length, the lower computation and the higher acquisition sensitivity. While MBOZP increases the acquisition performance by sacrificing computation, however, this sacrifice is acceptable.

6 Conclusion

Considering the problem of weak signal acquisition in spread spectrum communication systems, the acquisition performance is improved by MBOZP algorithm. Theoretical analysis and simulation results show that data block overlapping can achieve more accurate acquisition sensitivity without increasing too much complexity. However, the proposed algorithm is effective in engineering applications.

Acknowledgment. This work was supported by the National Nature Science Foundation of China (61271175) and the China 111 Project (B08038).

References

1. Foucras, M., Julien, O., Macabiau, C., Ekambi, B.: An efficient strategy for the acquisition of weak Galileo E1 OS signals. In: ENC 2013, European Navigation Conference (2013)
2. Chung, C.D., Cho, S.M.: Constant-envelope orthogonal frequency division multiplexing modulation. In: Communications, 1999, APCC/OECC 1999, Asia-Pacific Conference on Communications and Fourth Optoelectronics and Communications Conference, vol. 1, pp. 629–632. IEEE (1999)
3. Beifan, Z.: Research on the GPS satellite. Electron. Sci. **30**, 30–32 (2017)
4. Van Nee, D.J.R., Coenen, A.J.R.M.: New fast GPS code-acquisition technique using FFT. Electron. Lett. **27**(2), 158–160 (1991)
5. Lin, D.M., Tsui, J.B.Y.: Comparison of acquisition methods for software GPS receiver. In: Proceedings of the 13th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 2000), pp. 2385–2390 (2000)
6. Ziedan, N.I., Garrison, J.L.: Unaided acquisition of weak GPS signals using circular correlation or double-block zero padding. In: Position Location and Navigation Symposium, 2004, PLANS 2004, pp. 461–470. IEEE (2004)
7. Heckler, G.W., Garrison, J.L.: Implementation and testing of an unaided method for the acquisition of weak GPS C/A code signals. Navigation **56**(4), 241–259 (2009)
8. Zhang, W., Ghogho, M.: Improved fast modified double-block zero-padding (FMDBZP) algorithm for weak GPS signal acquisition. In: Signal Processing Conference, 2010 18th European, pp. 1617–1621. IEEE (2010)
9. Foucras, M.: Performance Analysis of the Modernized GNSS Signal Acquisition. Doctoral dissertation (2015)
10. Meng, Q., Liu, J.Y., Zeng, Q.H., Feng, S.J., Li, R.B.: BeiDou navigation receiver weak signal acquisition aided by block improved DBZP. Acta Aeronautica et Astronautica Sinica **38**(8), 144–154 (2017)
11. Li, Y.M., Li, J., Zhang, P., Zheng, Y.: Improved algorithm for weak GPS signal acquisition based on delay-accumulation method. Acta Geodaetica Cartogr. Sin. **45**(1), 44–49 (2016)
12. Ni, Y.H.: Research on fast acquisition algorithm of DSSS signal based on improved PMF-FFT. Electron. Measur. Technol. **36**(8), 33–36 (2013)



A Partially Repeated Scheme for Polar Codes

Daolong Wu(✉)

Key Laboratory of Technology on Datalink, 20th Institute, China Electronics Technology Group Corporation (CETC), Xi'an 710068, Shaanxi Province, China
rainl985_wu@163.com

Abstract. Since the code length of a polar code is an exponent of 2 if the Arikan's kernel is employed, it is hard to construct code that with flexible code length. In this paper, a partially repeated scheme for polar codes is proposed for constructing length flexible codes. And the Gaussian Approximation method is used to analyze the block error rate performance of the code. The simulation results verified the analysis.

Keywords: Polar codes · Gaussian approximation · Channel codes

1 Introduction

Polar codes, the first class of capacity-achieving codes, is proposed by Arikan in [1, 2] based on the phenomenon of channel polarization. The channel polarization theorem states that the bit-channel of polar codes tends to one of the two kinds as the code length goes to infinity. One the pure-noise channel, whose capacity is 0 while the other is the noise-less channel whose capacity is 1. By transmitting frozen bits which are known to the decoder over the pure-noise channels and information bits over the noise-less channels, polar codes can achieve the capacity of original binary input discrete memory-less symmetric channel (BI-DMSC) when the code length is infinity.

The construction of polar codes for binary erasure channel (BEC) is easy, however, the construction is quite complexity for additive white Gaussian noise (AWGN) channel. The conventional construction method of polar codes for AWGN channel is Monte Carlo approach [2], which need a large number of simulations before code construction. By observing the factor graph of polar codes, Mori and Tanaka viewed the process of channel polarization as an instance of density evolution (DE) [3, 4]. Unfortunately, the calculation complexity of the probability density function of each node is high, which makes the required memory increases exponentially with the code length. By extending the ideas of [3, 4], Ido and Vardy devised two approximation methods by which one can get the upper bound and lower bound on the error probability of each bit-channel efficiently [5]. Thus, polar codes can be constructed according to the upper bound and lower bound. Based on the ideas of the previous work, authors of [6] proposed a Gaussian Approximation (GA) method to construct polar codes for AWGN channel. The GA method can estimate the block error rate (BLER) performance of each code and construct polar codes with low complexity.

In practice, code length in practice is not always an exponent of 2. And the gap between 2^n and 2^{n+1} gets larger and larger with n increases, where n is a positive

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 317–322, 2018.
https://doi.org/10.1007/978-3-319-78130-3_32

integer. Thus, constructing length flexible codes is necessary. Usually, there are two methods to change the code length. One is puncture which is used to shorten the code length and increase the code rate. The other is concatenation which will enlarge the code length and decrease the code rate.

In this paper, we mainly consider the method of enlarge the code length. Since the error probability of each polarized bit-channel is different, we consider changing the code length by repeating part of the information bits which are transmitted over the bit-channels with poor error probability performance. By doing so, the code length can be enlarged, and the code rate is decreased.

The paper will be organized as follows: Sect. 2 will introduce the principle of polar codes and Gaussian approximation for polar codes. Section 3 will propose the partially repeated polar codes (PR-polar codes) and the corresponding Gaussian approximation method to estimate the BLER performance. Section 4 will give the numerical and the simulation results. Section 5 will conclude our work.

2 Preliminary

In this section, we will introduce the concept of polar codes and Gaussian approximation briefly.

2.1 Polar Codes

As presented in [1], the Arikan kernel $F = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ is usually chosen for the construction of binary polar codes. And $G_N = B_N(F^{\otimes n})$ is the generator matrix of a length N polar code, where $\otimes$ represents the Kronecker product, $N = 2^n$, and B_N represents the bit-reversal permutation matrix. A coded sequence X_1^N can be obtained by $X_1^N = U_1^N G_N$, where X_1^N denotes the vector $(X_1, X_2, \cdots, X_N)$ and $U_1^N = (U_1, U_2, \cdots, U_N)$. Based on the principle of polar codes, U_1^N consists of two parts. One subset contains the frozen bits which have been leaked to the decoder and the other contains the information bits. In this paper, we use the symbol $\mathcal{I}$ denote the set of the indices of the bit-channels used for transmitting information bits. Correspondingly, $\mathcal{I}^c$, the complement of $\mathcal{I}$ denotes the indices of the bit-channels used for transmitting frozen bits. For example, if the number of information bits is K and that of the frozen bits is $N - K$, the polar code is a rate K/N code because the mutual information of the frozen bits is zero.

A corrupted sequence Y_1^N can be obtained by transmitting the coded sequence X_1^N in N channel use. The N channel use can be viewed as a synthesized vector channel W_N. Suppose u_1^N is a realization of U_1^N and y_1^N is the corresponding channel output. The decoder should split vector channel W_N into N bit-channels $W_N^{(i)}$ $(i = 1, 2, \cdots, N)$. Each $W_N^{(i)}$ has an input u_i and output $\left(y_1^N, \hat{u}_1^{i-1}\right)$, where $\hat{u}_1^{i-1}$ is the estimation of the sequence u_1^{i-1}. Based on the splitting bit-channels, the probability $W_N^{(i)}\left(y_1^N, \hat{u}_1^{i-1} | u_i\right)$ of u_i can be obtained, where $u_i \in \{0, 1\}$. Let $h_i\left(y_1^N, \hat{u}_1^{i-1}\right)$ denote the hard decision of u_i, we have

$$\hat{u}_i = \begin{cases} u_i, & i \in \mathcal{I}^c, \\ h_i\left(y_1^N, \hat{u}_1^{i-1}\right), & i \in \mathcal{I}. \end{cases} \tag{1}$$

2.2 Gaussian Approximation

When the coded bits are transmitted via AWGN channels, the log-likely ratio (LLR) of the channel outputs are Gaussian consistent variables. It was shown in [7] that the LLR of each sub-channel is a Gaussian consistent variable if the hard decisions of the previous sub-channels are correct under a successive cancellation (SC) decoding.

To estimate the mean of the LLR of the i th sub-channel output, the initialized mean value $L_i = 2/\sigma^2$, where σ is the noise variance. Let $(b_1, b_2, \cdots, b_n)$ be the n-bit binary representation of i. Then for $j = 1, 2, \cdots, n$,

$$L_{\left(b_1, b_2, \cdots, b_{j-1}, 0\right)} = \phi^{-1}\left(1 - \left(1 - \phi\left(L_{\left(b_1, b_2, \cdots, b_{j-1}\right)}\right)\right)^2\right), \tag{2}$$

$$L_{\left(b_1, b_2, \cdots, b_{j-1}, 1\right)} = 2L_{\left(b_1, b_2, \cdots, b_{j-1}\right)}. \tag{3}$$

Where $\phi\ (x) \triangleq 1 - \frac{1}{\sqrt{4\pi x}} \int_{-\infty}^{\infty} \tanh\ (\tau/2)\ \ \exp\left[-(\tau - x)^2/(4x)\right] d\tau$.

To make the calculation simple, $\phi(x)$ can be approximated by

$$\phi(x) = \begin{cases} \exp\left(-0.4527x^{0.86} + 0.0218\right), & 0 < x < 10, \\ \sqrt{\frac{\pi}{x}} \exp\left(-\frac{x}{4}\right)\left(1 - \frac{10}{7x}\right), & x \geq 10. \end{cases} \tag{4}$$

By doing so, the mean of LLR of each sub-channel's output can be obtained. The first error probability of the ith sub-channel can be calculated by

$$P\left(W_N^{(i)}\right) = \frac{1}{2} erfc\ \left(0.5\sqrt{L_i}\right), \tag{5}$$

Where $erfc\ (x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-\eta^2} d\eta$.

3 Partially Repeated Scheme for Polar Codes

In this section, we propose the partially repeated scheme for polar codes firstly. And then, we analyze the block error performance of the scheme based on Gaussian approximation.

The structure of a partially repeated polar code is shown in Fig. 1. As the figure shows, the overall code length is N_1, and the length of the polar code is N. Thus, the code rate is K/N_1, where K must be no larger than N to make sure the polar code can be encoded. For the polar code, there are $N - K$ frozen bits which are shown by circles without filling. Those with different fillings are information bits. The N bits ($N - K$ frozen bits and K information bits) are encoded by a length N polar encoder. The coded

bits are denoted by boxes in Fig. 1. Since the error probability of each information bit is different, parts of the information bits are repeated several times in the PR-polar code scheme. In Fig. 1, the first information bit is repeated k_1 times and the second one is repeated k_2 times. Let k_i $(i = 1, 2, \cdots, K)$ denote the time of the i th information bit be repeated, we have $\sum_{i=1}^{K} k_i = N_1 - N$.

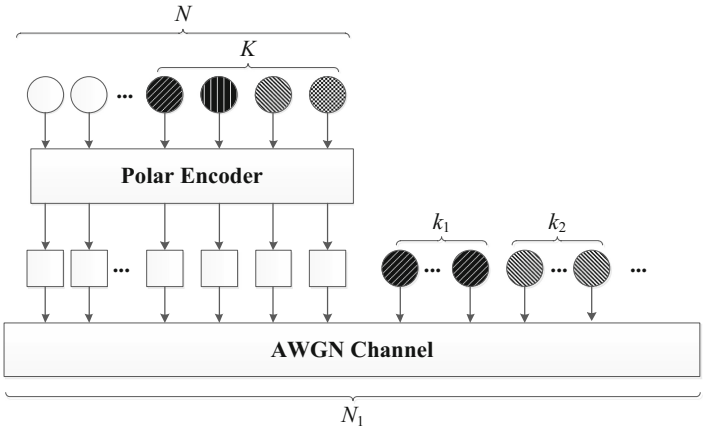


Fig. 1. Structure of a partially repeated polar code.

The decoder decodes the polar code by the first N channel outputs. When the LLR of the i th information bit is calculated, the decoder combines the calculated LLR with that of the k_i channel outputs of the repetitions of the i th information bits. By doing so, the reliability of the hard decision of the i th information bits is improved, and the decoding reliabilities of the following information bits can be improved by successive cancellation decoding.

The problem lies in that how to repeat the information bits properly to make the block error rate minimized.

Since the successive decoding is a sequential decoding, the bit-channel with the worst error performance dominates the error probability of the code word. Thus, the principle we determine the time of each information bit be repeated is repeating the information bit with the highest error probability at each time. The result of [6] shows that the higher the error probability of a bit-channel is, the lower the mean value of the bit-channel output is. Let $\mathcal{I}$ be the set of the indices of the bit-channel used to transmitting information bits, and $\mathcal{I}^{'} = \{1, 2, \cdots, K\}$ be the indices of the elements of $\mathcal{I}$, the algorithm to determine the repeating times can be described as follows:

(1) Initializing all $k_i = 0$, $i \in \mathcal{I}^{'}$.
(2) Calculating the LLR L_i $(i \in \mathcal{I})$ of the bit-channel whose index is an element of $\mathcal{I}$ by Gaussian approximation method.
(3) Sorting the element of $\mathcal{I}^{'}$ according to the value of L_i in ascending.

(4) Let j be the first element of $\mathcal{I}^{'}$, repeat the j th information bit once. And then, update $k_j = k_j + 1$, and $L_j = L_j + \frac{2}{\sigma^2}$.

(5) Repeat step 3 and 4 until $\sum_{i=1}^{K} k_i = N_1 - N$. The sequence k_i shows the times of each information bit need to be repeated.

By doing the steps above, times for each information bits will be repeated can be determined. And then, we can calculate the error probability of each information bit by (5). Noting that the L_i when using (5) is the updated mean value in step 4. The corresponding block error probability of PR-polar code can be obtained by

$$P_B = 1 - \prod_{i=1}^{K} \left(1 - P\left(W_N^{(i)}\right)\right). \tag{6}$$

4 Numerical and Simulation Results

The estimated block error rate performance and the corresponding simulation results are given in Fig. 2. To be specific, we give results of PR-polar codes with different length and code rate. Two PR-polar codes with rate 0.25 are shown in Fig. 2 which were denoted by code 1 and code 2. For code 1, the length of polar code is 1024, and there are totally 476 repeated bits. For code 2, the length of polar code is 512, correspondingly, the number of totally repeated bits is 988.

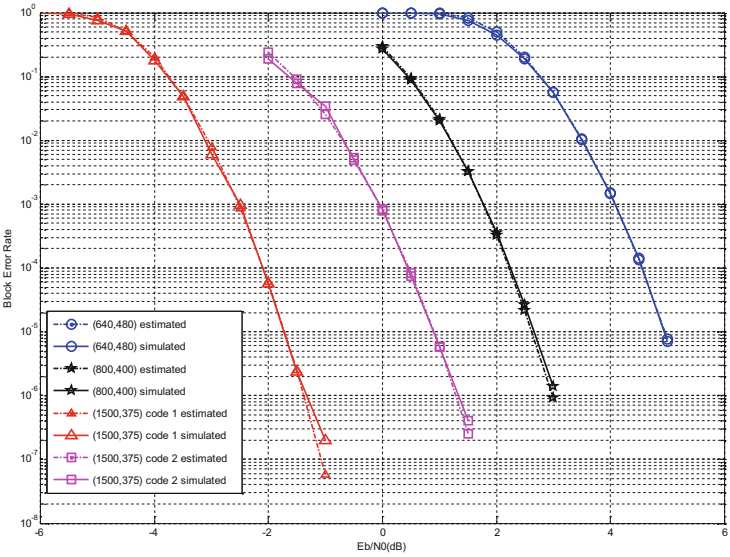


Fig. 2. BLER performance comparisons of different PR-polar codes with different lengths and rates.

We can observe that the estimated block error rates match the simulation results well no matter how the code lengths and the rates changes, which verified the accuracy of (6). The comparison between (1500, 375) code 1 and (1500, 375) code 2 suggests that the length of the polar code should be as near the total block length as possible when constructing a PR-polar code. The reason lies in that, larger code length of polar code can bring more channel polarization, and more coding gain can be expected.

5 Conclusion

In this paper, constructing of PR-polar codes is discussed. A method to construct PR-polar code and the corresponding analysis on block error rate based on Gaussian approximation of the constructed PR-polar code is also given. The numerical and simulation results verified the analysis method.

References

1. Arikan, E., Telatar, I.: On the rate of channel polarization. In: IEEE International Symposium on Information Theory, pp. 1493–1495. IEEE Press, Seoul (2009)
2. Arikan, E.: Channel polarization: a method for constructing capacity achieving codes for symmetric binary-input memoryless channels. IEEE Trans. Inf. Theor. **55**(7), 3051–3073 (2009)
3. Mori, R., Tanaka, T.: Performance of polar codes with the construction using density evolution. IEEE Commun. Lett. **13**(7), 519–521 (2009)
4. Mori, R., Tanaka, T.: Performance and construction of polar codes on symmetric binary-input memoryless channels. In: IEEE International Symposium on Information Theory, pp. 1496–1500. IEEE Press, Seoul (2009)
5. Tal, I., Vardy, A.: How to construct polar codes. IEEE Trans. Inf. Theor. **59**(10), 6562–6582 (2013)
6. Wu, D., Li, Y., Sun, Y.: Construction and block error rate analysis of polar codes over AWGN channel based on gaussian approximation. IEEE Commun. Lett. **18**(7), 1099–1102 (2014)



Performance Bound Analysis on Hamming-Weight-Analysis Algorithm for Blind Recognition of Linear Block Codes

Di Mao[1,2(✉)]

[1] Key Laboratory of Technology on Datalink, 20th Institute, China Electronics Technology Group Corporation (CETC), Xi'an 710068, China
mdcoml@l26.com
[2] State Key Laboratory of ISN, Xidian University, Xi'an 710071, China

Abstract. In this paper, the performance bound has been analyzed on Hamming-Weight-Analysis (HWA) algorithm. We first introduce the description of the linear block code and the theoretical analysis of the HWA algorithm. According to the theoretical analysis, we build simulation model and analyze the performance of HWA algorithm. HWA algorithm has a good recognition performance under the condition of low code rate and high BER. Finally, the performance bound of HWA algorithm is given based on the simulation results. The results in this paper have been validated through simulation on MATLAB.

Keywords: Blind recognition · Linear block codes
Hamming-Weight-Analysis · Performance bound

1 Introduction

Due to the deterioration property of the wireless channel, the general communication system is always designed according to the worst channel situation, which guarantees the reliable communication to achieve desirable performance in time-varying channel. However, this principle often leads to inadequate utilization of channel capacity. To solve the problem, adaptive modulation and coding (AMC) is a good solution [1]. AMC system will adjust the parameters of transmitter according to the variation of channel quality. Since the modulation pattern and coding rate are changing with channel conditions, the sender need to transmit the control information about them to the receiver. Thus the receiver can adjust demodulator and decoder to the most suitable state. However, in practical, it is always difficult to establish a synchronous communication system due to the delay, interference, interrupt and some other reasons during the transmission. This fact makes the relevant information cannot be sent to the receiver on time or correctly, so that communication will be failed. Therefore, it is an important problem to quickly identify the modulation pattern and FEC mode based on the received data, which is also the requisite to optimize the performance of the system.

Hamming-Weight-Analysis (HWA) algorithm enjoys the benefits of simple and fast, and it can be widely used in coding recognition. As long as the code rate and code length is in the reliable recognition area, any linear block code, no matter system code

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 323–331, 2018.
https://doi.org/10.1007/978-3-319-78130-3_33

or nonsystematic code, is readily applicable to recognize its code length. HWA algorithm also has a good performance under high BER condition, which makes it an effective algorithm in practical for blind recognizing and provides the basis for blind recognizing other parameters. With the development of cognitive radio and cognitive communication, blind recognition of channel coding will be one of the most important functions for intelligent communication system in the future.

The existing literature about coding recognition is rare. Most of them focus on the identification and reconstruction of convolution codes [2, 3]. In reference [4, 5], the authors use Euclidean Algorithm to identify convolution codes and punctured convolution codes. Few publications are there for recognition of parameters of block codes. Valembois has formalized the problem of finding the nearest block code from bit stream and proved it to be NP complete [6]. Galois FFT has been used to recover the parameters of BCH code in [7]. In reference [8, 9], the authors use cyclic property of the block codes which are not applicable in case of linear code that is not cyclic. Even though these methods are useful, there is little research on the performance analysis and performance bounds of these methods. Therefore, the researches on the performance of these methods are quite necessary.

This paper investigates the performance of Hamming-Weight-Analysis (HWA) algorithm through fundamental principle and simulation analysis. The simulation results show that the HWA algorithm has a good recognition performance under the condition of low code rate and short code length. Moreover, HWA algorithm also work well when the BER is high. Finally, base on the basic theory and simulation statistical characteristics, the performance bound of HWA algorithm is given.

The rest of the paper is organized as follows. Section 2 describes the mathematical model of linear block codes and the basic theory of HWA algorithm. In Sect. 3, we analyze the performance of HWA algorithm and give the performance bound of HWA algorithm. Simulation results on MATLAB are also provided in Sect. 3. Finally, the conclusion is drawn in Sect. 4.

2 Background

In this section, the basic concept and mathematical model of linear block codes is introduced. We will also discuss the fundamental principle of Hamming-Weight-Analysis algorithm.

2.1 The Description of Linear Block Codes

Linear Block Codes are the most important class of block codes [10]. A (n, k) Block Code is a kind of Algebraic code. Each codeword consists of k message symbols and $n - k$ parity symbols. There is an algebraic relation between the parity symbols and parity symbols, if the relation is linear, it is called linear block code. In GF(2), we consider symbols as bits (Fig. 1).

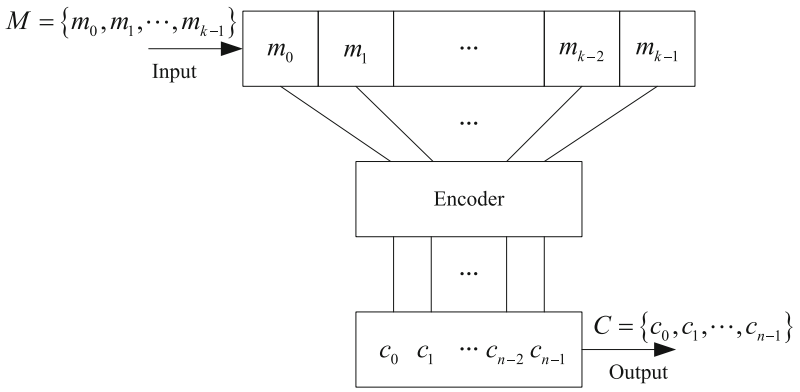


Fig. 1. The structure of encoder of linear block codes

The mathematical model of linear block codes can be described as: $C = M \times \mathbf{G}$, where $M = \{M_0, M_1, \cdots, M_i, \cdots\}$ denotes input message sequence which have k bits per group, $M_i = \{m_0, m_1, \cdots, m_{k-1}\}$ denotes k message bits which send into encoder at *i-th* moment. $C = \{C_0, C_1, \cdots, C_i, \cdots\}$ denotes output codewords which have n bits per group, $C_i = \{c_0, c_1, \cdots, c_{n-1}\}$ denotes the codeword generated by encoder at *ith* moment. $\mathbf{G}$ is called the generator matrix.

2.2 Hamming-Weight-Analysis Algorithm

Hamming-Weight-Analysis (HWA) is an effective algorithm to recognize length of codeword. The Hamming weight of a codeword is equal to the number of nonzero elements in the codeword. For a (n, k) linear block code, we define $W(C)$ as distribution of Hamming weight, $W(C) = \{N_i = d, 0 \le d \le n\}$, $W(C)$ is the number of codewords in C which Hamming weight is d. p_i is the weight distribution probability of d-weight codewords.

Theorem 1. If each transmitted bit has a probability q of being received incorrectly on BSC channel, q is quite small. Received sequence r has the minimum distance from codeword v_j, in other words, for $i = 1, 2, \cdots, q^k$, $\forall i \neq j$, satisfy

$$(r, v_j) < d(r, v_i) \tag{1}$$

So we have maximum conditional probability when v_j is sent message and r is received message, then

$$P(r|v_j) > P(r|v_i) \tag{2}$$

Theorem 2. For a digital communication system, the probability of t simultaneous errors in one code block is P(t), then

$$P(t) \gg P(t+1), \quad t = 0, 1, \cdots q^k \tag{3}$$

Assume BER is 10^{-3}, we have

$$P(1) = C_n^1 10^{-3}(1-10^{-3})^{n-1} \approx C_n^1 \times 10^{-3}$$
$$P(2) = C_n^2 10^{-6}(1-10^{-3})^{n-2} \approx C_n^2 \times 10^{-6}$$

Thus $P(1) \geq P(2)$, the change of code weight caused by error is only ± 1. Therefore, Theorems 1 and 2 show us that the change of code weight caused by error is not significant. Thus, errors have little effect on weight distribution, therefore, the length of block code n can be estimated by the weight distribution.

For a (n, k) linear block code, the size of codebook is $2^k \ll 2^n$ which is far less than 2^n, so the probability distribution of codeword in n-dimensional vector space is unequally.

Theorem 3. For a (n, k) linear block code, the codebook v, which is generated by a k-tuple vector, is a subset of n-dimensional space V, and the probability distribution of v in V is unequally.

If estimated codeword length $\hat{n}$ is not equal to n, there are no constraints between each $\hat{n}$-length "codeword". It can be considered that the occurrence probability of 0 and 1 is equally. Assuming that the weight distribution of the sequence is a equal- probability distribution model, the probability of code vectors with different weight is

$$p = 1/(\hat{n}+1) \tag{4}$$

Assuming that P_i is the actual probability of code vectors with weight i, the weight distribution distance is defined as

$$D(\hat{n}) = \frac{\hat{n}+1}{\hat{n}} \sum_{i=0}^{n} \left| P_i - \frac{1}{\hat{n}+1} \right|^2 \tag{5}$$

where $\hat{n}$ is the estimated codeword length, $D(\hat{n})$ is variance distance between actual distribution and uniform distribution. According to Theorem 1, $\hat{n}$ will be the estimated codeword length only if $D(\hat{n})$ is maximum.

3 Performance Analysis of HWA Algorithm

In this section, the performance of HWA algorithm is investigated through computer simulations. We choose BCH code as unknown FEC, and the parameter of BCH code is shown in Table 1.

Table 1. Parameters of BCH codes used for simulation in this paper.

n	k	Generator polynomial
31	6	313365047
31	11	5423325
31	16	107657
63	10	472622305527250155
63	16	6331141367235453
63	18	1363026512351725
63	24	17323260404441
63	30	157464165547
127	22	1233760704047225224354456266376470 43
127	29	403114461367670603667530141176455
127	36	3146074666522075044764574721735
127	43	177217722136512275212205743 43
127	50	5444651252331401242150 1421
127	57	335265252505705053517721

Simulation results of (31,6) BCH Codes is shown in Figs. 2 and 3.

From Fig. 2, it is visible that weight distribution distance is maximum when estimated codeword length $\hat{n} = 31$, which is also the code length of the BCH Code we used. Figure 3 shows the weight distribution when $\hat{n} = 31$, and weight distribution is the most unbalanced under this circumstance. Therefore, HWA algorithm has a good performance for blind recognition.

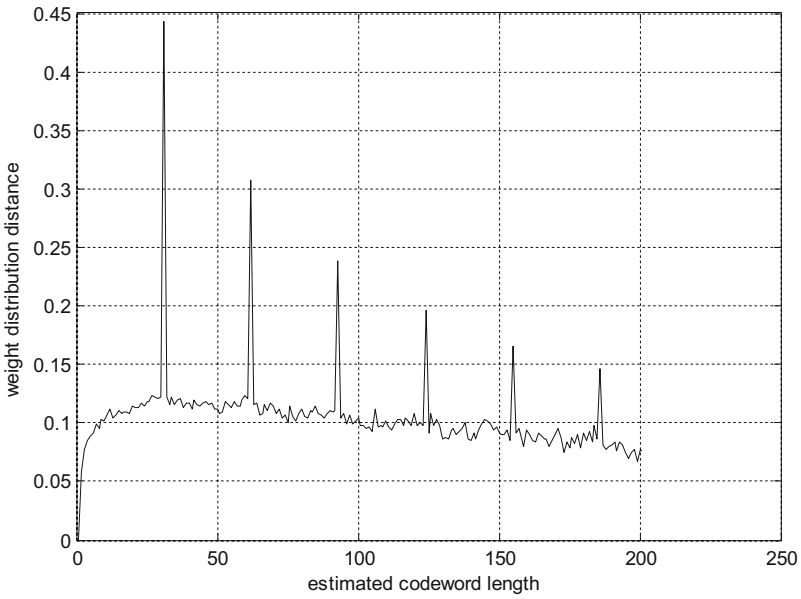


Fig. 2. Plot of weight distribution distance calculation using HWA algorithm

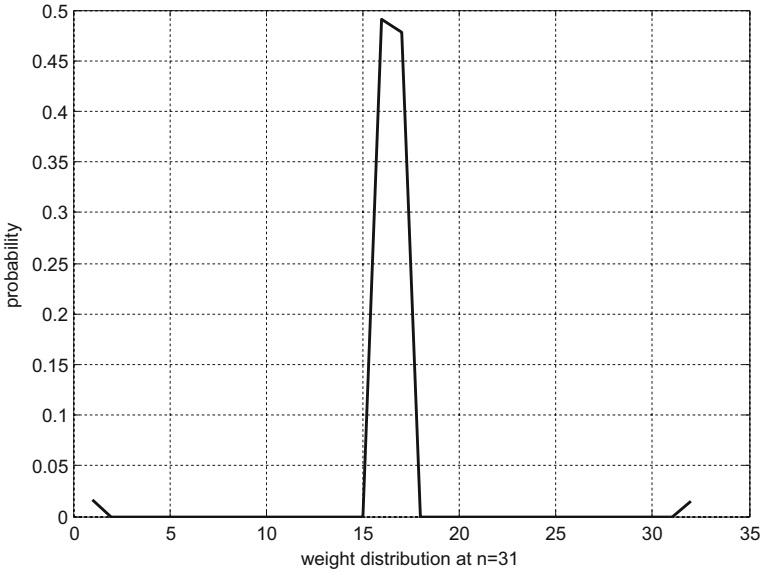


Fig. 3. Plot of weight distribution when estimated codeword length $\hat{n} = 31$

According to Theorem 3, the code weight distribution will gradually approach to uniform distribution with the increase of code rate. Therefore, the performance of HWA algorithm will decreases with the increase of code rate. On the other hand, the increase of code length will also cause performance decreases which is shown in Fig. 4. From Fig. 4, it is visible that the effective range of HWA algorithm decreases rapidly when code rates are close to 0.5, in other words, the HWA algorithm will achieves performance upper bound when code rate is 0.5. In order to obtain the complete performance bound, we need to do the simulations for different code length and code rate.

The reliable recognition points are shown in Fig. 5. According to the distribution of these points and theoretical analysis, we can get the performance bound curve of HWA algorithm. The black line in Fig. 5 is the performance bound, the area under the curve is the reliable recognition area of HWA algorithm and the points in this area are theoretical reachable.

HWA algorithm also has a good performance against bit errors, which is derived from Theorem 2. From Fig. 6, it is visible that the success rate of recognition can be kept at high level under high BER condition. The range of HWA algorithm decreases rapidly with the increase of code rate and code length.

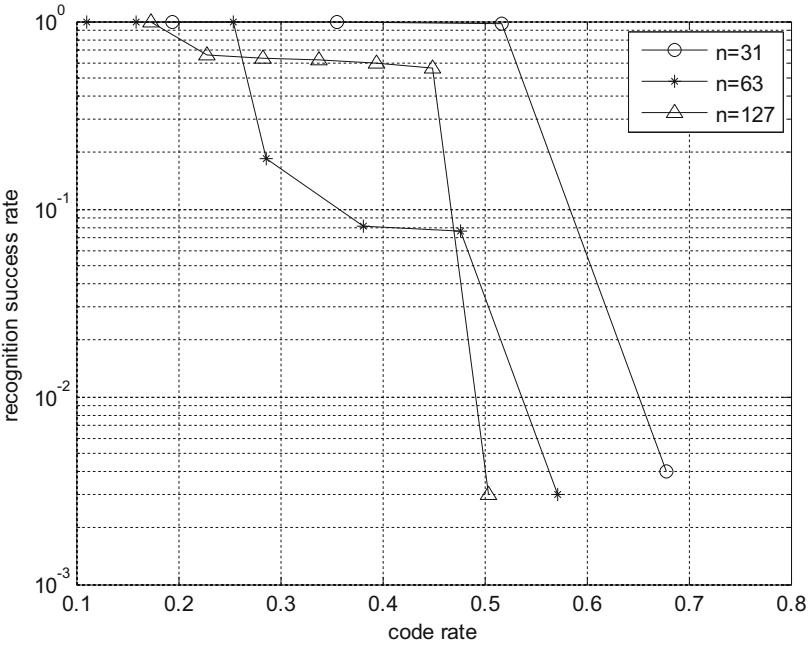


Fig. 4. Plot of recognition success rate vs code rate

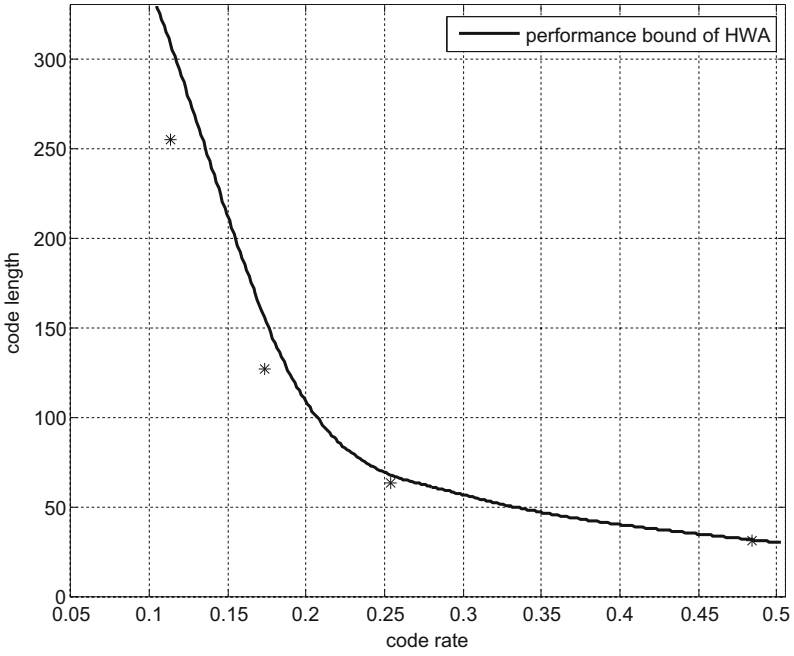


Fig. 5. Plot of performance bound of HWA algorithm

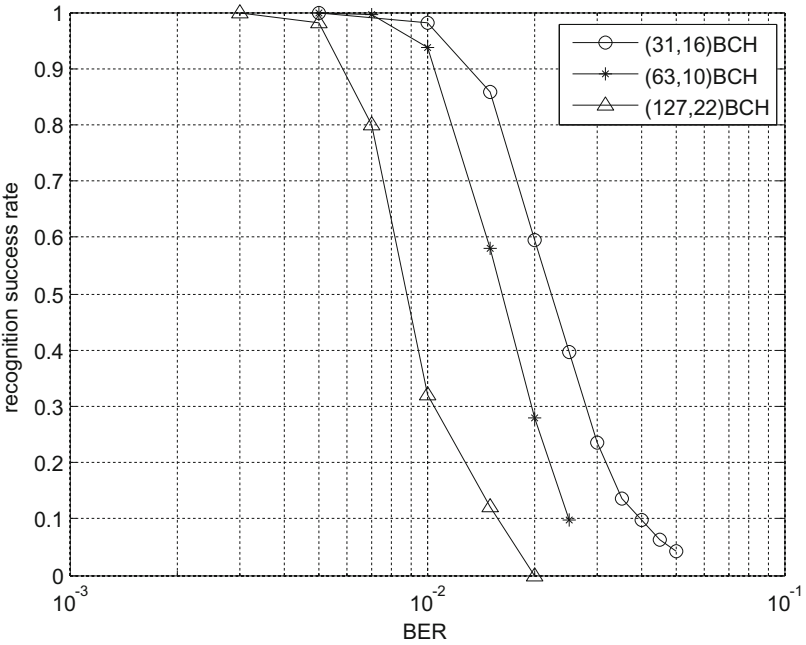


Fig. 6. Plot of recognition success rate vs BER

4 Conclusion

HWA algorithm is one of the simplest and fastest algorithms for coding recognition. In this paper, we investigate the performance bound of HWA. With the basic introduction on the linear block code and the HWA algorithm, the simulation has been conducted. By analyzing the simulation results, the HWA bound has be given which is a significant parameter for blind recognition of channel coding.

References

1. Goldsmith, A.J., Chua, S.-G.: Adaptive coded modulation for fading channels. IEEE Trans. Commun. **46**(5), 595–602 (1998)
2. Filiol, E.: Reconstruction of convolutional encoders over GF(q). In: Darnell, M. (ed.) Cryptography and Coding 1997. LNCS, vol. 1355, pp. 101–109. Springer, Heidelberg (1997). https://doi.org/10.1007/BFb0024454
3. Rice, B.: Determining the parameters of a rate 1/n convolutional encoder over GF(q). In: Proceedings of Third International Conference on Finite Fields and Application, Glasgow (1995)
4. Wang, F., Huang, Z., Zhou, Y.: A method for blind recognition of convolution code based on Euclidean algorithm. In: IEEE International Conference on Wireless Communications, Shanghai, pp. 1414–1417 (2007)

5. Peizhong, L., Li, S., Yan, Z., Xiangyang, L.: Blind recognition of punctured convolutional coding. Sci. China **35**(2), 173–185 (2005)
6. Valembois, A.: Detection and recognition of a binary linear code. Discrete Appl. Math. **111** (1–2), 199–218 (2001)
7. Wu, G., Zhang, B., Wen, X., Guo, D.: Blind recognition of BCH code based on Galois Field Fourier Transform. In: International Conference Wireless Communications & Signal Processing (WCSP 2015) (2015)
8. Wang, J., Yue, Y., Yao, J.: Statistical recognition method of binary BCH code. Commun. Netw. **3**, 17–22 (2011)
9. Sharma, A., Pillai, N.R.: Blind recognition of parameters of linear block codes from intercepted bit stream. In: International Conference on Computing, Communication and Automation (ICCCA 2016), pp. 1262–1266 (2016)
10. Lin, S., Costello Jr., D.J.: Error Control Coding, 2nd edn. Pearson Prentice Hall, Upper Saddle River (2004)



An Efficient Algorithm for Refining Position and Velocity Outputs of Space Borne GNSS Receivers

Shuhao Chang[1(✉)], Xi Chen[1], and Menglu Wang[2]

[1] Research Institute of Tsinghua University in Shenzhen, Shenzhen 518057, China
chang-sh15@mails.tsinghua.edu.cn, chenxiee@tsinghua.edu.cn
[2] School of Aerospace Engineering, Tsinghua University, Beijing 100084, China
wangml14@mails.tsinghua.edu.cn

Abstract. Space borne GNSS receivers may not have enough tracked GNSS space vehicles due to the geometry caused by high altitude and limited receiver sensitivity. In many cases, such as real-time orbit determination for communication satellites, it is necessary to refine the position and velocity outputs of a space borne GNSS receiver for improved accuracy and robustness before actually using them. Toward this problem, an efficient algorithm jointly using weighted Runge-Kutta integration and cubic Hermite polynomial interpolation is proposed in this work. Simulations are conducted based on the GNSS data of GRACE-B satellite and LING QIAO satellite, the result of which show the proposed algorithm can effectively eliminate outliers and significantly reduce the root mean square error of GNSS position and velocity outputs.

Keywords: GNSS · Space borne · Runge-Kutta · Refinement

1 Introduction

As one of the most important technologies of human history, global navigation satellite systems (GNSS) have been used for various purposes in both civil and military areas. It can provide precise, continuous, world-wide, three dimensional position and velocity information to users with appropriate positioning algorithms. Space borne GNSS receivers are those GNSS receivers used by space crafts and satellites. The use of a GNSS receiver as the orbit position, velocity and even attitude sensor, has become a standard component of satellites nowadays, especially LEO satellites. Precision and cost-effectiveness, among others, are main contributors to the popularity of GNSS in space missions.

Due to higher altitude, space borne GNSS receivers generally have less average trackable GNSS space vehicles than their ground counterparts. In order to obtain a position and velocity reference of LEO satellites, at least four tracked navigational space vehicles should be in sight. However, space borne GNSS receivers may have not enough tracked GNSS satellites in view. For LEO satellites, a lack of trackable GNSS space vehicles can only occur occasionally. Swarm

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 332–340, 2018.
https://doi.org/10.1007/978-3-319-78130-3_34

C satellite of European Space Agency (ESA) flies in an orbit 530 km above the earth approximately and the possibility that it tracks less than four GPS satellites is 2.6% [1]. The number of trackable GNSS space vehicle will decrease as orbit altitude increases. Actually, Space-borne GNSS receivers can be used not only for orbit determination of LEO satellites, but also for orbit determination of satellites with higher altitude. NASA Goddard Space Flight Center (GSFC) had developed a new space-borne GPS receiver called Navigator that can operate effectively in the full range of Earth orbiting missions from LEO to GEO and beyond. For such GNSS receivers, the opportunity of tracking four or more satellites drops more significantly than LEO GNSS receivers. Besides orbit height, there are other reasons causing a space borne GNSS receiver to have inferior observations of tracked GNSS space vehicle signals. For example, erroneous bit synchronization due to limited acquisition sensitivity, scintillations on GNSS systems are much more severe and frequent for LEO satellites, particularly during posts unset hours [2], because of which observations may contain big error. Space radiation effects such as single element upset, may also cause large observation error.

A space borne GNSS receiver typically uses iterative least square or a Kalman filter to figure out a positioning solution. These L2 algorithms are not robust enough when there are large observation deviation. Theoretically, large observation deviation can be checked out and eliminated in pretreatment. However, preprocessing only works well when there are enough observations.

In existing work, we proposed a position and velocity refining algorithm based on short-arc filtering [3] at the expense of computation complexity. As a continuation of existing work, a more robust and efficient algorithm for refining position and velocity outputs of space borne GNSS receivers is proposed in this work. It is based on weighted Ronge-Kutta method and cubic Hermite polynomial interpolation. To overcome speed deficiency of Runge-Kutta method, enough but less data with outliers being screened out are used. Simulations with orbit data from GRACE-B satellite and GNSS data of LING QIAO (NORAD ID: 40136, International code: 2014-051A) were undertaken to show the efficiency and robustness of the proposed algorithm.

The paper is organized as follows: Sect. 2 gives the details and discussion of the proposed algorithm. Section 3 presents simulation performance and numerical results. The paper concludes in Sect. 4.

2 The Proposed Algorithm

2.1 General Idea

The proposed algorithm for refining position and velocity outputs of space borne GNSS receivers works as follows:

1. Save the output of a GNSS receiver for the last N epochs.
2. Approximately obtain the curve of position and velocity to get knowledge of a gentle and non-anomalous trajectory through cubic Hermite polynomial interpolation or polynomial fitting.

3. Calculate the Euclidean distance from each point to the three-order curve and omit anomalous points once they are too far away from the curve. A confirmed threshold q is required to determine whether the point we are aiming at is largely affected by non-ideal factors.
4. Gill's formula of weighted forth-order Runge-Kutta method is involved to attain better reference of position and velocity of the current epoch. We only use the "fine points" we have already screened out in the former step and reciprocals are used as weights.
5. The robustified position and velocity reference of the current epoch we have already got can be used in the following epochs to revised the orbit (Fig. 1).

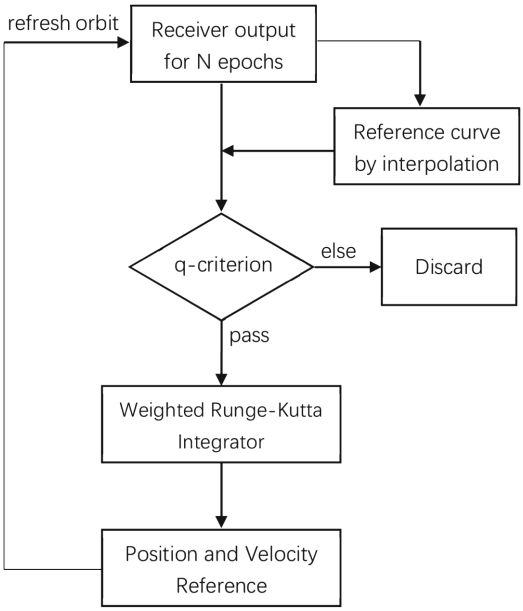


Fig. 1. Flowchart of the proposed algorithm

2.2 Algorithm Details

For the sake of convenience, we first declare a mathematical abstraction as a unified language in this paper. We map each output of GNSS receiver into the seven-dimension space as a point, epoch by epoch. Consider the six-tuple containing positions and velocities of x,y and z axis as a description of a point's state and, naturally, we obtain a series of six-tuples as

$$\mathbf{X}(t) = \begin{bmatrix} P_x(t) \\ P_y(t) \\ P_z(t) \\ V_x(t) \\ V_y(t) \\ V_z(t) \end{bmatrix} = \begin{bmatrix} P(t) \\ V(t) \end{bmatrix} \tag{1}$$

Now we can easily write down a relational expression:

$$V(t) = \begin{bmatrix} \mathbf{0}_{3\times3} & \mathbf{I}_{3\times3} \end{bmatrix} \begin{bmatrix} P(t) \\ V(t) \end{bmatrix} = \mathbf{A}\mathbf{X}(t) \tag{2}$$

Low earth orbit communication satellites' movements are mainly affected by J2 perturbation in earth's non-spherical perturbation among all kinds of perturbation forces [4]. Taking J2 perturbation into consideration, the acceleration of x, y and z axis can be denoted as follows:

$$\begin{cases} a_x = \left[-\frac{GM}{r^3}(1 + \frac{3}{2}J_2(\frac{r_e}{r})^2(1 - 5(\frac{r_e}{r})^2)) + \omega_e^2\right] P_x + 2\omega_e V_y \\ \quad = \xi_1 P_x(t) + 2\omega_e V_y \\ a_y = \left[-\frac{GM}{r^3}(1 + \frac{3}{2}J_2(\frac{r_e}{r})^2(1 - 5(\frac{r_e}{r})^2)) + \omega_e^2\right] P_y - 2\omega_e V_x \\ \quad = \xi_1 P_y(t) - 2\omega_e V_x \\ a_z = \left[-\frac{GM}{r^3}(1 + \frac{3}{2}J_2(\frac{r_e}{r})^2(3 - 5(\frac{r_e}{r})^2))\right] P_z(t) \\ \quad = \xi_2 P_z \end{cases} \tag{3}$$

where GM is the geocentric gravitational constant, $J_2 = 1.082628 \times 10^{-3}$ is the Earth's second dynamic form factor in the Nodal precession, ω_e is the angular velocity of the earth, r_e is the equatorial radius of the earth and $r = \sqrt{P_x^2 + P_y^2 + P_z^2}$ is the orbit radius. In a unified format, we arrange the formulas above to get the derivative of state variable X:

$$a(t) = \begin{bmatrix} \xi_1 & 0 & 0 & 0 & 2\omega_e & 0 \\ 0 & \xi_1 & 0 & -2\omega_e & 0 & 0 \\ 0 & 0 & \xi_2 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} P(t) \\ V(t) \end{bmatrix} = \mathbf{B}\mathbf{X}(t) \tag{4}$$

Thus

$$\frac{d}{dt}\mathbf{X}(t) = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \mathbf{X}(t) = \mathbf{\Phi}_{6\times6}\mathbf{X}(t) \tag{5}$$

Now we involve the forth-order Runge-Kutta method. We adopt Gill's formula of the Runge-Kutta method, for rounding errors are largely reduced in it [6]. In case we are given the initial value $\mathbf{X}_0 = \mathbf{X}(t_0)$ as well as the derivative $\frac{d}{dt}\mathbf{X}(t) = f(t, \mathbf{X}_0)$, then the value $\mathbf{X}(t_0 + h)$ could be estimated with confidence as follows:

$$\mathbf{X}(t_0 + h) = \mathbf{X}_0 + \frac{1}{6}(k_1 + (2 - \sqrt{2})k_2 + (2 + \sqrt{2})k_3 + k_4) \tag{6}$$

where

$$\begin{cases} k_1 = hf(t_0, \mathbf{X}_0) \\ k_2 = hf(t_0 + \frac{h}{2}, \mathbf{X}_0 + \frac{k_1}{2}) \\ k_3 = hf(t_0 + \frac{h}{2}, \mathbf{X}_0 + \frac{\sqrt{2}-1}{2}k_1 + \frac{2-\sqrt{2}}{2}k_2) \\ k_4 = hf(t_0 + h, \mathbf{X}_0 - \frac{\sqrt{2}}{2}k_2 + \frac{2+\sqrt{2}}{2}k_3) \end{cases} \tag{7}$$

Take each one of the N points as a initial point and calculate a reference result for the state of current epoch and give it a weight inversely proportional to time interval. For instance, the result we calculate from the point n epochs before should be conferred with weight $1/n$. Note that the initial points we use ought to be screened out by the judgement whether they are too *far* from the three-order fitting curve. Each point with a six-dimension Euclidean distance to the fitting curve larger than a delicately designed threshold q should be omitted. Add all reference results to a temporary register X_{temp} and calculate the sum of wights p. Thus, a better state reference for current epoch is derived as follows:

$$\mathbf{X}_{ref} = \mathbf{X}_{temp}/p \tag{8}$$

where

$$p = \sum_{n} \frac{1}{n} \ , \ n \in \{n \leq N \mid q - criterion\ pass\} \tag{9}$$

2.3 Rationale Analysis

In this part, several topics worthy for discussion are expounded. Feasibility and designation matters are included here.

During the motion of satellites, measurement errors including offset in bit synchronization, tracking loop instability as well as hatching effect can cause inaccuracy of orbit parameters. Analysis based on the real in-orbit data show that an abrupt decrease of actual tracked satellite number may be a major reason for outliers. Thus, in order to get accurate orbit parameters, we need an approximate and available reference to pick out those outliers which should be excluded from the following integration. It has been testified that piecewise polynomial interpolation, such as splines or piecewise Hermite interpolation, adopted on short intervals is such a desirable choice. Furthermore, piecewise cubic interpolation with sampling intervals of about 100–200 s is considered to have a perfect behaviour with slight error [5]. Negligible computational complexity is required during this step and continuity at the junction point is ensured. Meanwhile, cubic polynomial fitting is also feasible to generate the reference curve. More time cost, however, is inevitable to solve the linear equation group with Vandermonde matrix of time.

Here gives a third-order fitting curve to illustrate this procedure. Figures 2 and 3 shows a practical local fitting error and we can determine the threshold q from these graphs as priori knowledge. For deep fading situations, it both works to exclude outliers by thresholds for separate parameters or six-dimension Euclidean distance. Outliers can be ignored by q-criterion in this way and will not have impact on following integration procedures. It's worth noting that according to different circumstances, the parameters are supposed to be given different values. As for the epochs involved in the algorithm, computing time will increase and the interpolated or fitting curve will not work well with current fixed order if N is too large. While if N is chosen improperly small, the long-time error of the

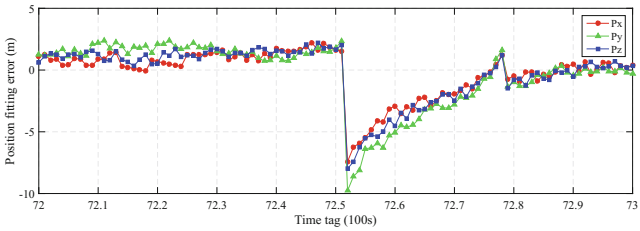


Fig. 2. Third-order fitting error of position

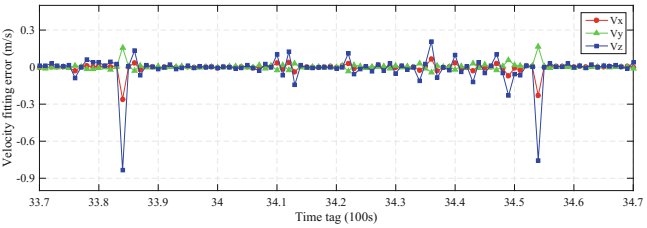


Fig. 3. Third-order fitting error of velocity

satellite will be neglected and acts out as wave property in the final estimate of error. It demonstrates the relationship between the performance of the algorithm and N in Fig. 4, taking LING QIAO satellite for example. And we can conclude that with the value of N between 10 and 15 a relatively reasonable performance of the algorithm on LING QIAO satellite can be achieved.

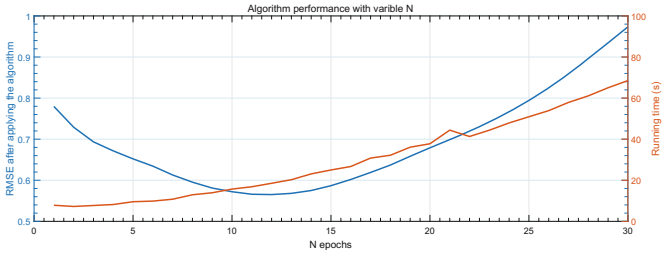


Fig. 4. Relationship between the performance of the algorithm and N

3 Simulation Results

3.1 Correctness Check

Before applying the algorithm to LING QIAO satellite, we first take correctness check on PVT data collected from GRACE B receiver to prove that the algorithm is Gaussian noise-resisting. Since GRACE-B flies in a polar orbit 500 km over

earth, the precision of its position and velocity data given by the receiver is relative high. To imitate a bad situation, additive white Gaussian noise(AWGN) is applied on original position and velocity data from the receiver. The position and velocity errors before going through the process system is depicted in Figs. 5 and 6 with red circles, while blue cubes shows the effect after going through the system. Here we involve last $N = 15$ epochs in integration and threshold is set as $q = 10$ m. Results shows the RMSE of position reference decreased from 3.911 m to 0.863 m and for velocity reference it decreased from 0.434 m/s to 0.131 m/s, which is the normal range of GRACE's position error. This check shows the robustness of the proposed algorithm.

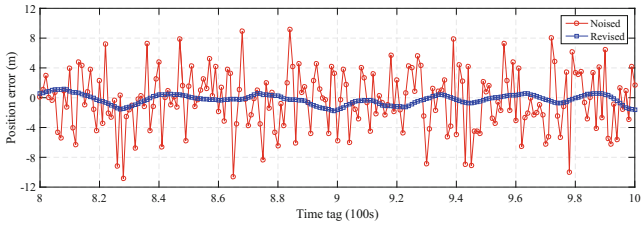


Fig. 5. Comparison of position errors of x-axis

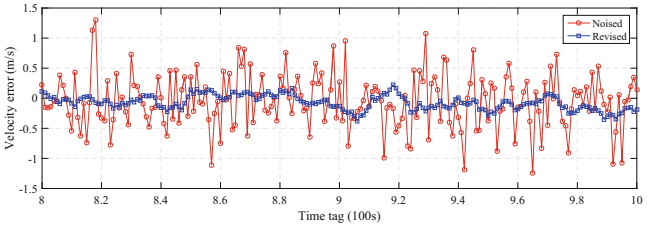


Fig. 6. Comparison of velocity errors of x-axis

3.2 Simulation on LING QIAO

In this part, simulations are conducted based on the original data provided by LING QIAO GPS receiver on Mar. 11th, 2015. LING QIAO satellite flies in an orbit 780 km above the earth approximately and the GPS receiver output suffers stronger interference than GRACE satellites. In the simulations, the algorithm factors are adopted as follows: $N = 10$, $q = 10$ m. In Fig. 7, position errors of x, y and z axis before and after robustifying are depicted and the results show a perfect property against signal deep-fading of the algorithm. The thing is, outliers of the original orbit data, such as the point at $t = 36.25, 31.30$, are revised by the algorithm. These outliers are very likely caused by abrupt

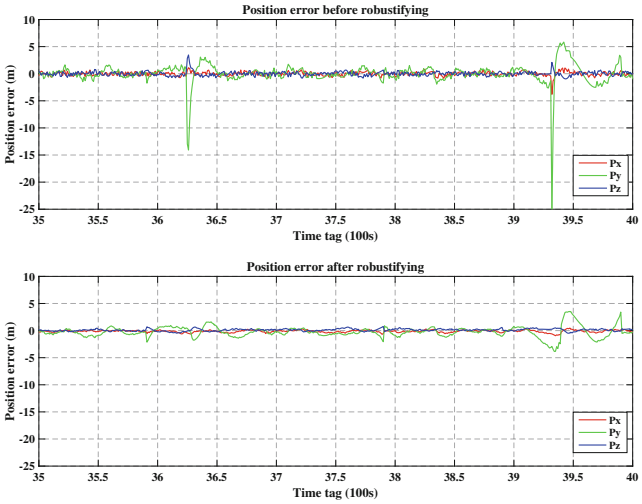


Fig. 7. Comparison of position errors

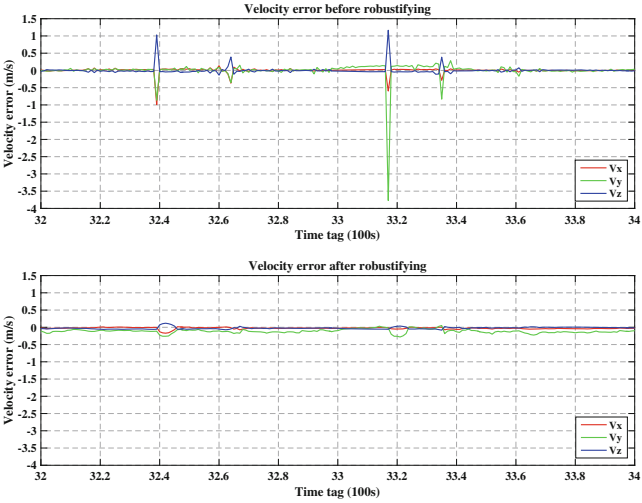


Fig. 8. Comparison of velocity errors

short-time decrease of actual tracked satellite number. Through the process of screening outliers and weighted integration, better reference for actual position is obtained. Analysis of velocity simulation is the same in Fig. 8. The RMSE of position reference decreased from 1.672 m to 0.577 m and for velocity reference it decreased from 0.086 m/s to 0.029 m/s. Thus the algorithm can approximately reduce the root mean square error of receiver output by 60% to 70%.

4 Conclusions

In this paper, we proposed an algorithm to acquire refined position and velocity as orbit state reference from space borne GNSS receiver outputs. Piecewise Hermite interpolation or similar fitting method are used to screen out nonanomalous points and reciprocals of time epochs are used to weight the Runge-Kutta solutions. It is also shown by simulation results based on real in-orbit data of GRACE-B and LING QIAO that the proposed algorithm can effectively eliminates outliers and significantly reduce the root mean square error of position and velocity data.

Acknowledgement. The authors would like to thank the Science, Technology and Innovation Commission of Shenzhen Municipality (No. JSGG20160429165838848).

References

1. Xiong, C., Stolle, C., Lxefhr, H.: The Swarm satellite loss of GPS signal and its relation to ionospheric plasma irregularities. Space Weather **14**, 563–577 (2016)
2. Basu, S., Groves, K.M., Basu, S., Sultan, P.J.: Specification and forecasting of scintillations in communication/navigation links: current status and future plans. J. Atmos. Solar-Terr. Phys. **64**, 1745–1754 (2002)
3. Xi, C., Menglu, W., Wenyun, G.: Robustifying position and velocity references for communication satellites by short-arc filtering. In: International Conference on Wireless Communications & Signal Processing (WCSP), Nanjing (2015)
4. Montenbruck, O., Gill, E., Lutze, F.: Satellite orbits: models methods and applications. Appl. Mech. Rev. **55**, 2504–2510 (2002)
5. Korvenoja, P., Piche, R.: Efficient satellite orbit approximation. In: Proceedings of International Technical Meeting of the Satellite Division of the Institute of Navigation, pp. 1930–1937 (2000)
6. Fuyang, K., Qing, W., Shuguo, P.: GLONASS orbit Runge-Kutta integral algorithm by automatic integral step length. J. SE Univ. **40**, 755–759 (2010)



Fountain-Coding Aided Symmetrical Encryption for Secure Wireless Transmission

Chao Lyu[1,2(✉)], Pinyi Ren[1,2], Qinghe Du[1,2], and Li Sun[1,2]

[1] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China
lvchao@stu.xjtu.edu.cn, {pyren,duqinghe,lisun}@mail.xjtu.edu.cn
[2] National Simulation Education Center for Communications and Information Systems, Xi'an, China

Abstract. It is difficult to ensure secure transmissions in wireless networks because of the openness of radio signal. In this paper, a fountain-coding aided symmetrical encryption scheme is proposed to deal with this problem. In the proposed scheme, all source packets are first encrypted with a secret key. Then, all encrypted source packets are encoded with fountain code and next transmitted on wireless channels. Finally, the secret key is encrypted with all the fountain-coded packets the legitimate receiver has received for sharing the secret key between the legitimate users. Due to the independent fading characteristic between the legitimate channel and the wiretap channel, if the eavesdropper fails to correctly receive some fountain-coded packets the legitimate receiver receives, the wireless transmission can be secured under appropriate conditions. Compared with the counterparts, the proposed scheme achieves lower intercept probability of the eavesdropper with little receiver feedback burden under appropriate conditions.

Keywords: Secure wireless transmission · Fountain code · Secret key
Eavesdropper · Intercept probability · Receiver feedback burden

1 Introduction

Due to the openness of wireless medium, data transmissions in wireless networks can be easily overheard by the eavesdropper. Traditional approaches including symmetrical and asymmetrical encryption are employed to protect the confidential data from being intercepted. However, the management and distribution of the secret keys are hard to handle in dynamic wireless networks.

The research work reported in this paper is supported by the National Natural Science Foundation of China under the Grant No. 61431011, the Key Research and Development Program of Shannxi Province under Grant No. 2017ZDXM-GY-012 and the Fundamental Research Funds for the Central Universities.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 341–350, 2018.
https://doi.org/10.1007/978-3-319-78130-3_35

Apart from the traditional cryptography-based paradigm, physical layer security, which exploits the inherent characteristics of noise and wireless channels, is another way to ensure secure wireless transmission. Most existing physical layer security schemes try to maximize the secrecy capacity which gives the maximum transmission rate of legitimate users while the eavesdropper can not achieve any information from received signal. To ensure the *perfect secrecy*, any transmission rate which is higher than the secrecy capacity is prohibited. From a practical perspective, this requirement does not always need to be satisfied when the transmitted file is composed of correlated source packets and the eavesdropper can not achieve useful information even though a small number of the source packets are intercepted [2].

Implementing reliable communications without retransmission [4], Fountain code, such as LT code, can be introduced to solve the aforementioned issues. In fountain-coded data transmission, the source file is first divided into K source packets. Then, fountain-coded packets are continually generated by the modulo-2 addition of randomly chosen distinct source packets and transmitted to the legitimate receiver. Once the legitimate receiver has correctly received N fountain-coded packets, where N is slightly larger than K, and recovered the source file, the fountain-coded data transmission is terminated. If the legitimate receiver successfully receives enough fountain-coded packets and recovers the source file before eavesdropper does, the wireless transmission between legitimate users can be secured. Compared with the existing physical layer security strategies aiming at maximizing the secrecy capacity, the data rate of the fountain-coding aided scheme is only bounded by the Shannon capacity of the legitimate channel.

In [3], a dynamic fountain code design scheme is proposed to enhance the security of wireless transmissions. However, in the dynamic fountain code design scheme, the legitimate receiver has to send the sequence set of all received source packets back to the legitimate transmitter every time when the legitimate receiver has correctly received a source packet. Sending a large number of feedbacks adds heavy burden to the legitimate receiver.

In this paper, a novel fountain-coding aided symmetrical encryption scheme is proposed for secure wireless transmission. Compared with the dynamic fountain code design scheme, although the proposed scheme also exploits the independent fading characteristic between the legitimate channel and the wiretap channel, the proposed scheme ensures secure wireless transmission with better security performance and less receiver feedback burden under appropriate conditions. It should be noted that the proposed scheme does not address the *perfect secrecy* for secure wireless transmission either. The supreme aim is still to lower the signal-to-noise-ratio (SNR) at the eavesdropper to decrease its intercept probability. The proposed scheme is appropriate and convenient to complement the existing encryption techniques in upper layers. Especially, when the secret key of the existing encryption techniques is cracked, the proposed scheme can effectively strengthen the security of the vulnerable wireless network.

The rest of this paper is organized as follows. Section 2 presents the system model. Section 3 illustrates the proposed scheme in detail. Section 4 presents the

performance analyses of the proposed scheme. Simulation results are shown in Sect. 5. Finally, conclusions are given in Sect. 6.

2 System Model

2.1 System Descriptions

As is shown in Fig. 1, this paper considers a wireless network consisting of three nodes including one legitimate transmitter (Alice), one legitimate receiver (Bob), and one eavesdropper (Eve). In this system, Alice wants to send a confidential file to Bob while Eve seeks to intercept it. To deliver a confidential file, Alice first divides it into K source packets denoted by $(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K)$. Then, a potentially infinite fountain-coded packets denoted by $(\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \ldots)$ are generated by the modulo-2 addition of randomly chosen distinct source packets. For each fountain-coded packet, the encoding information is inserted in the packet header to generate a complete data packet. Next, every data packet is further encoded by CRC coding and channel coding. Finally, all encoded data packets are modulated and transmitted to wireless channels. Due to the independent fading characteristic between the legitimate channel and the wiretap channel, Bob and Eve may correctly receive different fountain-coded packets. Once Bob has correctly received N fountain-coded packets, where N is slightly larger than K, and successfully recovered the confidential file, Bob will send a feedback to Alice to terminate the wireless transmission. While if Eve has not received enough fountain-coded packets to recover the confidential file, the wireless transmission of the confidential file is secured.

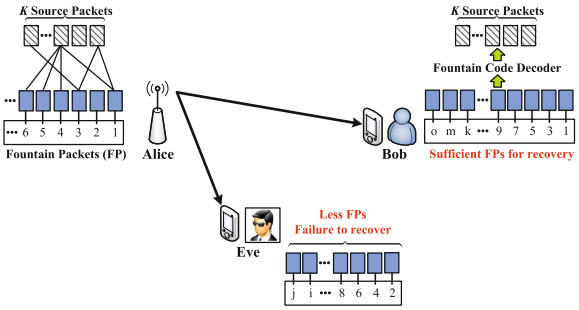


Fig. 1. System model.

The legitimate channel from Alice to Bob and the wiretap channel from Alice to Eve are modeled as block flat Rayleigh fading channels which means the channel coefficient is a constant in a transmission slot but changes independently across different transmission slots. The channel coefficient between a node pair (i, j) is denoted by h_{ij}, which is modeled as a complex circularly symmetric

Gaussian variable with mean zero and variance $\delta_{ij}^2 = d_{ij}^{-\alpha}$ where d_{ij} denotes the distance between node i and node j and α is the path loss exponent. The average transmitting power of Alice is $\bar{P}$, and the additive noise at Bob and Eve are assumed as complex Gaussian variables with mean zero and variance N_0. Therefore, the average SNR of the system is represented by $\rho = \bar{P}/N_0$. Accordingly, the SNR at node j is represented by $\gamma_{Aj} = \rho|h_{Aj}|^2$.

Algorithm 1. Fountain Code Encoding Operation

1: Randomly generate the degree d_n of a fountain-coded packet $\mathbf{f}_n$ with a degree distribution $\rho(d)$. Different fountain codes generally have different degree distributions.
2: Uniformly randomly choose d_n different source packets and the fountain-coded packet $\mathbf{f}_n$ is generated by the modulo-2 addition of those d_n source packets.

Algorithm 2. Iterative Decoder

1: Search a fountain-coded packet $\mathbf{f}_n$ with the degree d_n satisfying $d_n = 1$, which signifies that the fountain-coded packet is connected to only one source packet $\mathbf{s}_k$. If such a fountain-coded packet does not exist, the decoding operation stops at this moment and fails to recover all source packets.
(a) Set $\mathbf{s}_k = \mathbf{f}_n$ and add $\mathbf{s}_k$ to connected fountain-coded packets: $\mathbf{f}_{n'} = \mathbf{f}_{n'} \oplus \mathbf{s}_k$.
(b) Remove the edges between the source packet $\mathbf{s}_k$ and the fountain-coded packets.
2: Repeat 1 until all source packets are recovered.

2.2 The Encoding and Decoding Principles of LT Code

LT code retains the good performance of the random linear fountain code with low encoding and decoding complexities [4]. The encoding operation is described by Algorithm 1. Two kinds of decoder including the iterative decoder which is described by Algorithm 2 and the Gaussian elimination decoder which is interpreted by Algorithm 3 can be used for fountain code decoding. The time complexity of the iterative decoder is $O(n \log(n))$ while the time complexity of the Gaussian elimination is $O(n^3)$. However, the decoding capability of the Gaussian elimination decoder is stronger than the iterative decoder which indicates the Gaussian elimination decoder can decode some fountain-coded packet sets that the iterative decoder can not decode. Considering the time consumption, Bob intends to apply the iterative decoder for decoding. Considering the decoding capability, Eve intends to employ the Gaussian elimination decoder for decoding.

3 Scheme Descriptions

As is shown in Fig. 2, the basic principle of the proposed scheme is as follows: First, Alice divides the source file into K source packets denoted by

Algorithm 3. Gaussian Elimination Decoder

Input: An augmented matrix $\mathbf{A}_{N\times(K+1)} = (\mathbf{G}, \mathbf{F})$, where $\mathbf{G}_{N\times K}$ is the generator matrix of fountain-coded packets and $\mathbf{F}_{N\times 1}$ is the corresponding fountain-coded packet vector.

Output: Decoding state or the source packet vector $\mathbf{S}$.

```
for (i = 0, j = 0; i < N AND j < K; i++, j++)
    maxRow = argmax(A[row][j], row = i, ..., N − 1);
    if (A[maxRow][j] == 0)
        i = i − 1;
        continue;
    if (maxRow != i)
        swapRows(i, maxRow);
    for (k = i + 1; k < N; k++)
        if (A[k][j] != 0)
            for (l = j; l <= K; l++)
                A[k][l] = A[k][l] ⊕ A[i][l];
if (i < K)
    return "Fail to Decode!";
for (k = i; k < N; k++)
    if (A[k][j] != 0)
        return "Fail to Decode!";
for (k = K − 1; k >= 0; k − −)
    S[k] = A[k][K];
    for (l = k + 1; l < K; l++)
        if(A[k][l] == 1)
            S[k] = S[k] ⊕ S[l];
return S;
```

$(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K)$. Then, all the source packets are encrypted with a secret key $\mathbf{k}$ which is as long as the source packet. Next, all the encrypted source packets denoted by $(\mathbf{s}'_1, \mathbf{s}'_2, \ldots, \mathbf{s}'_K)$ are encoded using fountain code, such as LT code, and transmitted to Bob. Once Bob has decoded all the encrypted source packets $(\mathbf{s}'_1, \mathbf{s}'_2, \ldots, \mathbf{s}'_K)$, Bob sends the sequence set $\mathbf{V}$ of all the correctly received fountain-coded packets $\mathbf{F}$ back to Alice. After receiving the sequence set, Alice encrypts the secret key $\mathbf{k}$ with the corresponding fountain-coded packets $\mathbf{F}$ in the sequence set $\mathbf{V}$ and sends the encrypted secret key $\mathbf{k}'$ to Bob. After receiving the encrypted secret key $\mathbf{k}'$, Bob decrypts it with the received fountain-coded packets $\mathbf{F}$. Finally, Bob decrypts all the encrypted source packets $(\mathbf{s}'_1, \mathbf{s}'_2, \ldots, \mathbf{s}'_K)$ and recovers the source file. Therefore, if Eve intends to intercept the source file, Eve has to correctly receive not only all the fountain-coded packets $\mathbf{F}$ that Bob received but also the sequence set $\mathbf{V}$ and the encrypted secret key $\mathbf{k}'$.

With stricter limitations for Eve to intercept the source file and fewer feedbacks for Bob to send, the proposed scheme is safer than the dynamic fountain code design scheme with less receiver feedback burden. It is worth mentioning when the source packet length as well as the secret key length is long enough, and the number of source packets is large enough, Eve is unable to decrypt the

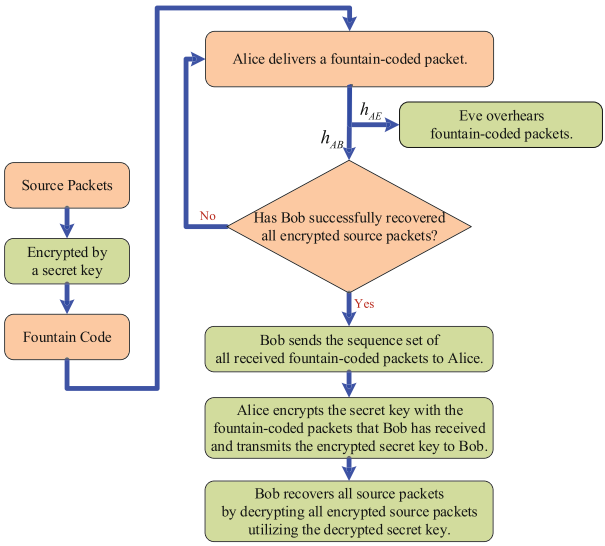


Fig. 2. Transmission mechanism.

encrypted data by brute force attacks in an effectively short time. To better evaluate the performance of the proposed scheme, Eve's brute force attacks are ignored in analyzing Eve's intercept probability.

4 Performance Analyses

4.1 Eve's Intercept Probability

According to [6], in the fading channel with additive Gaussian white noise, the relationship between the packet error rate (PER) and the SNR at Bob satisfies Eq. 1, where n is the transmission mode number, γ represents the SNR at Bob and a_n, g_n, γ_{pn} are fitting parameters. In [6], two tables give values of those fitting parameters according to the specific transmission mode.

$$\mathrm{PER}_n(\gamma) \approx \begin{cases} 1, & if \quad 0 < \gamma < \gamma_{pn} \\ a_n \exp(-g_n \gamma), & if \quad \gamma \geq \gamma_{pn} \end{cases} \tag{1}$$

From Sect. 2.1, it is learnt that the statistical average packet error rate of the legitimate channel and the wiretap channel can be calculated as Eqs. 2 and 3.

$$\lambda = 1 - \exp\left(-\frac{\gamma_{pn}}{\rho {d_{AB}}^{-\alpha}}\right) + \frac{a_n}{g_n \rho {d_{AB}}^{-\alpha} + 1} \exp\left(-\frac{g_n \rho {d_{AB}}^{-\alpha} + 1}{\rho {d_{AB}}^{-\alpha}} \gamma_{pn}\right) \tag{2}$$

$$\beta = 1 - \exp\left(-\frac{\gamma_{pn}}{\rho {d_{AE}}^{-\alpha}}\right) + \frac{a_n}{g_n \rho {d_{AE}}^{-\alpha} + 1} \exp\left(-\frac{g_n \rho {d_{AE}}^{-\alpha} + 1}{\rho {d_{AE}}^{-\alpha}} \gamma_{pn}\right) \tag{3}$$

It is assumed that Alice divides the source file into K source packets. From Sect. 3, it is learnt that if Eve intends to intercept the source file, Eve has to correctly receive at least all the fountain-coded packets that Bob received. Therefore, the upper bound of the intercept probability of Eve can be expressed as Eq. 4, where k represents the number of fountain-coded packets that Alice transmits, l represents the number of fountain-coded packets that Bob receives, and δ_l represents the probability that Bob receives l fountain-coded packets but fails to decode.

$$P_{Intercept} = \sum_{k=K}^{\infty} \left[\sum_{l=K}^{k} \left[\binom{k-1}{l-1} \lambda^{k-l} (1-\lambda)^{l} (1-\delta_l) (1-\beta)^{l} \right] \right] \tag{4}$$

4.2 Receiver Feedback Burden

First, receiver feedback burden is defined as M/K, where M represents the number of feedbacks that Bob sends and K represents the number of source packets. The receiver feedback burden indicates the estimated number of feedbacks that Bob has to send when Bob has successfully acquired a source packet. In real situations, Bob is usually a mobile device which is powered by lithium battery, so the receiver feedback burden significantly influences the usage time of Bob. Therefore, good security schemes should ensure secure wireless transmission with less receiver feedback burden.

From Sect. 3, it is learnt that Bob has to send the sequence set $\mathbf{V}$ back to Alice and receive the encrypted secret key $\mathbf{k}'$ that Alice transmits after recovering all encrypted source packets in the proposed scheme. Therefore, when system employs the proposed scheme in this paper, considering the PER of the fading channel, the receiver feedback burden can be calculated as $2/[(1-\mathrm{PER})K]$.

5 Simulation Results and Discussions

In the simulations, all nodes are located in the rectangular coordinate system. Alice and Bob are located at (0, 0), (1, 0). It is assumed that Eve is located at (x, 0). The path loss exponent α is equal to 3 and the system SNR ρ satisfies $\rho = 10\,\mathrm{dB}$. 16-QAM modulation is employed and the coding rate equals 9/16. The fitting parameters a_n, g_n, γ_{pn}, which can be determined with TABLE II of [6], satisfy Eq. 5. Then, the PER of the legitimate channel and the wiretap channel can be calculated as Eqs. 2 and 3. Alice divides the source file into K source packets.

$$a_n = 50.1222, g_n = 0.6644, \gamma_{pn} = 7.7021. \tag{5}$$

Alice employs LT code with the robust soliton distribution. Bob applies the iterative decoder for fountain code decoding while Eve employs the Gaussian elimination decoder for fountain code decoding. The total simulation times is set to be 10^5. Eve's intercept probability equals to Eve's successfully intercepting

times divided by the total simulation times. The receiver feedback burden equals to the number of feedbacks that Bob sends divided by the number of source packets in one simulation.

To better evaluate the performance of the proposed scheme, three kinds of baseline schemes are adopted in this paper.

(1) Basic LT code with the robust soliton distribution [4].
(2) The optimal degree distribution scheme [7] of which the degree distribution satisfies Eq. 6, where K represents the number of total source packets, θ represents the number of undecoded source packets and $\lceil\delta\rceil$ indicates the minimum integer greater than δ.

$$d^*(K,\theta) = \left\lceil \frac{K+1-\theta}{\theta} \right\rceil \tag{6}$$

(3) The dynamic fountain code design scheme in [3].

For the baseline scheme 1, the receiver feedback burden can be calculated as $1/[(1-\mathrm{PER})K]$, and the simulation results of Eve's intercept probability approximates the theoretical results. For the baseline scheme 2, as it is configured that Bob sends feedbacks to Alice without loss, the lower bound of the theoretical receiver feedback burden and the lower bound of Eve's theoretical intercept probability can be achieved by the simulation. For the baseline scheme 3, the receiver feedback burden equals to 1, and the lower bound of Eve's theoretical intercept probability can be obtained through the simulation because of the same reason as the baseline scheme 2. For the proposed scheme, the receiver feedback burden can be calculated as $2/[(1-\mathrm{PER})K]$, and the upper bound of Eve's theoretical intercept probability can be acquired from the simulation results, as it is configured that Bob sends the feedback to Alice and Eve without loss and Alice sends the encrypted secret key to Bob and Eve without loss while Eve may probably fail to correctly receive the feedback that Bob sends or the encrypted secret key that Alice sends in real situations.

The impact of the distance from Alice to Eve on Eve's intercept probability is investigated in Fig. 3(a), where the number of source packets satisfies $K = 128$. As is observed from Fig. 3(a), Eve's intercept probability in all schemes fast decreases as the distance from Alice to Eve increases. This is reasonable because as Eve is farther away from Alice, the SNR at Eve gets worse and the PER of the wiretap channel increases. What's more, Eve's intercept probability in the proposed scheme is much smaller than that in the baseline scheme 1, 2, and 3 for the different locations of Eve. Therefore, the proposed scheme is safer than other three schemes when $K = 128$ for the different locations of Eve.

Figure 3(b) plots Eve's intercept probability as a function of the value of K. It can be seen from Fig. 3(b) that Eve's intercept probability in the baseline scheme 1 and the lower bound of Eve's intercept probability in the baseline scheme 2 nearly stays the same for different values of K, and the upper bound of Eve's intercept probability in the proposed scheme and the lower bound of Eve's intercept probability in the baseline scheme 3 fast decreases as the value

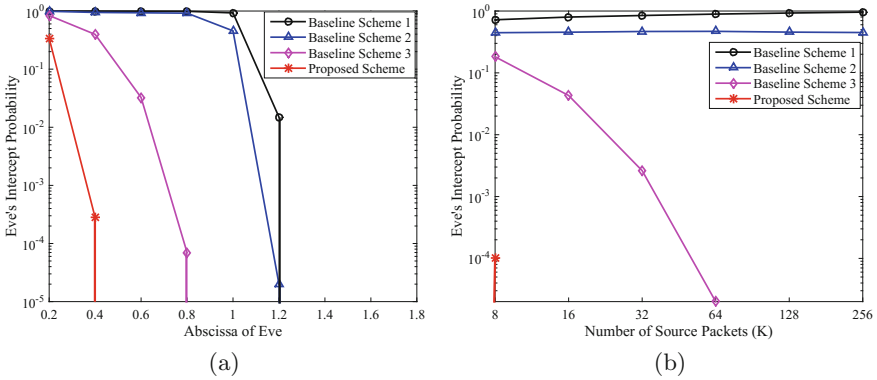


Fig. 3. (a) The intercept probability of Eve versus the distance from Alice to Eve, where $K = 128$. (b) The intercept probability of Eve versus the value of K, where Eve is located at (1, 0).

of K increases. In addition, Eve's intercept probability in the proposed scheme is much smaller than that in the baseline scheme 1, 2, and 3 for different values of K. Therefore, the proposed scheme is safer than the other three schemes when Eve is located at (1, 0) for different values of K. To protect the confidential file from brute force attacks, the proposed scheme is appropriate to be applied to application scenarios with a large value of K.

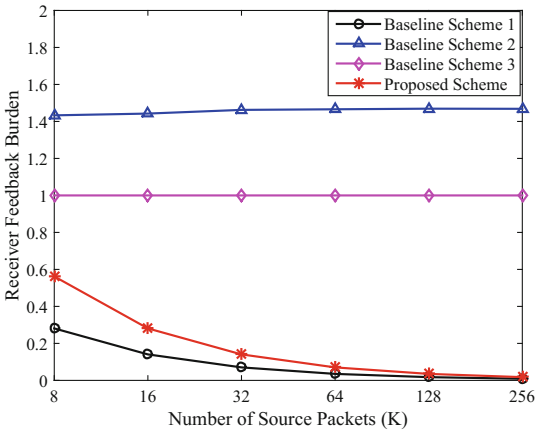


Fig. 4. The receiver feedback burden versus the value of K, where Alice is located at (1, 0).

Figure 4 exhibits the relationship between the receiver feedback burden and the value of K. As is observed from Fig. 4, the receiver feedback burden in the

proposed scheme and the baseline scheme 1, which are less than that in the baseline scheme 2 and 3, are nearly equal to zero for large values of K. Therefore, the proposed scheme almost has the same receiver feedback burden as the baseline scheme 1 for large values of K, and it is friendlier to Bob than the baseline scheme 2 and 3. Accordingly, the proposed scheme is appropriate to be applied to application scenarios with a large value of K.

6 Conclusions

In this paper, a fountain-coding aided symmetrical encryption scheme is proposed for secure wireless transmission. According to the basic characteristic of fountain code, a sufficient number of fountain-coded packets have to be correctly received to recover all encrypted source packets. Through exploiting the independent fading characteristic between the legitimate channel and the wiretap channel, if the eavesdropper fails to correctly receive some fountain-coded packets the legitimate receiver receives, the eavesdropper can not intercept the source file under appropriate conditions. The theoretical analyses and simulation results of performance both prove that the proposed scheme achieves lower intercept probability of the eavesdropper with little receiver feedback burden under appropriate conditions.

References

1. Niu, H., Iwai, M., Sezaki, K., Sun, L., Du, Q.: Exploiting fountain codes for secure wireless delivery. IEEE Commun. Lett. **18**(5), 777–780 (2014)
2. Sun, L., Ren, P., Du, Q., Wang, Y.: Fountain-coding aided strategy for secure cooperative transmission in industrial wireless sensor networks. IEEE Trans. Ind. Inform. **12**(1), 291–300 (2016)
3. Li, W., Du, Q., Sun, L., Ren, P., Wang, Y.: Security enhanced via dynamic fountain code design for wireless delivery. In: Proceedings of IEEE WCNC 2016, Doha, Qatar, 3–6 April 2016
4. MacKay, D.J.C.: Fountain codes. IEE Proc. Commun. **152**(6), 1062–1068 (2005)
5. Luby, M.: LT codes. In: Proceedings of IEEE 43rd Annual Symposium on Foundations of Computer Science, pp. 271–280, November 2002
6. Liu, Q., Zhou, S., Giannakis, G.B.: Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links. IEEE Trans. Wirel. Commun. **2**(5), 1746–1775 (2004)
7. Zhang, X., Du, Q.: Adaptive low-complexity erasure-correcting code based protocols for QoS-driven mobile multicast services over wireless networks. IEEE Trans. Veh. Tech. **55**(5), 1633–1647 (2006)
8. Barros, J., Rodrigues, M.R.D.: Secrecy capacity of wireless channels. In: Proceedings of IEEE International Symposium on Information Theory, pp. 356–360 (2006)
9. Bloch, M., Barros, J.: Physical-Layer Security: From Information Theory to Security Engineering. Cambridge University Press, Cambridge (2011)
10. Goldsmith, A.: Wireless Communications. Cambridge University Press, Cambridge (2005)

A New Anti-blackout Communication Method Based on Carrier Aggregation of OFDMA and Frequency Diversity

Jieqi Li[1], Xiaoya Zuo[2(✉)], Yong Zhang[1], Zhengxue Li[1], Shaohui Mei[2], and Zhongqi Tian[2]

[1] Research and Development Center, China Academy of Launch Vehicle Technology, Beijing 100076, China
[2] School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China
zuoxy@nwpu.edu.cn

Abstract. Aiming at the problem of communication blackout in the near-space aviation, a new anti-blackout communication method is proposed based on carrier aggregation of OFDMA and frequency diversity. The plasma sheath covered by the aerocraft is treated as a wireless fading channel. The same message is modulated onto every subcarrier of OFDMA in different frequency band. Then frequency diversity reception is used to improve the robustness of the data link. The analysis and simulation results show that the method could significantly improve the bit error performance of the communication link through the plasma sheath. This research provides an effective and simple way for anti-blackout communications.

Keywords: Blackout · Plasma · OFDMA · Frequency diversity
Bit error

1 Introduction

In recent years, with the development of manned space flight, outer space exploration and hypersonic vehicle, more and more attention has been paid to near space communications. When a spacecraft flies in the atmosphere with ultra high speed, a plasma sheath is formed at the front end. The plasma is responsible for the reflection, refraction, absorption and dispersion of electromagnetic waves. When the aircraft's signal can not pass through the plasma sheath, it will interrupt communication with the outside world, which is the "blackout" phenomenon [1]. The presence of the black obstacle increases the acquisition difficulty and the possibility of target loss. The system communication error rate will increase at the same time [2].

The main reason for the plasma sheath to produce the "Blackout" phenomenon is the free electron interaction, so the key point to solve signal interruption problem is how to change the free electron distribution and decrease the free electron density. The corresponding technical measures can be classified into three categories: one is to improve the design of the aircraft itself, such as changing the aerodynamic configuration, improving the heat resistant material, arranging the antenna window reasonably, improving the

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 351–360, 2018.
https://doi.org/10.1007/978-3-319-78130-3_36

transmission power or changing the communication frequency; the second is to reduce the free electron density near the antenna, such as the use of chemical reduce technology; the third is to increase the magnetic field from outside to limit the movement of the free electron and change the signal propagation model [3–6].

At present, the elimination methods of blackout lack of simple and effective communication methods. The changing of the aerodynamic shape, improving the hear resistance material and arranging the antenna window reasonably are based on the aircraft structure; Increasing the transmission power or changing the communication frequency have a large cost and the effect is not obvious; Reducing the electron density near the free antenna and the external magnetic field need to install extra special equipment on the aircraft and will cause great negative impact on the aircraft appearance design.

According to the research of the wave propagation in the plasma, the wave fading characteristic change with various parameters is often non-linear. Even for this non-linear relationship with a thorough understanding, limiting the influence of the blackout by using this nonlinear influence is a tough problem. How to compensate for the nonlinear influence caused by each influence parameter is still a problem at present. In view of this, this paper proposes a simple and effective method based on orthogonal frequency division multiplexing access (OFDMA) carrier aggregation and frequency diversity to anti the blackout communication. Considering the plasma sheath as a wireless channel, the same measurement and control information are transmitted in a plurality of OFDMA subcarriers, so as to improve the communication reliability through frequency diversity. The analysis and simulation results demonstrate the effectiveness of the proposed method, thus providing a simple and effective method for anti-blackout communication. To our knowledge, this is the first time that this method has been proposed for anti-blackout communications.

2 Anti-blackout Communication Method Based on OFDMA Carrier Aggregation and Frequency Diversity

2.1 Carrier Aggregation

In the evolution of the LTE system to the LTE-Advanced system, there are many factors that affect its evolution, in which the wide spectrum requirements are the most important factor. At present, LTE-Advanced system has 6 candidate frequency bands, but these applications are not enough to carry the entire bandwidth of the 100 MHz bandwidth required by the LTE-A system. Therefore, in order to solve this problem, 3 GPP uses the carrier aggregation technology to extend the system bandwidth, in order to meet the capacity of LTE-A system, which makes the carrier aggregation technology become one of the key technologies of LTE-Advanced. The so-called carrier aggregation technology is to aggregate the continuous or discrete LTE system carrier to a large bandwidth LTE-Advanced system carrier, where the system bandwidth can reach 100 MHz. This carrier aggregation technology has many advantages. First, it can use the origin LTE system physical channel and modulation coding scheme, and without the redesign of the system decreases the influence for the LTE-Advanced system. It greatly reduces the cost

of the system. Secondly, the carrier aggregation technology reuses the resources of the original LTE system, which not only reduces the difficulty of designing the LTE-Advanced system but also implements a smooth transition from LTE system to LTE-Advanced system. Finally, the application of carrier aggregation makes the LTE-Advanced system backward compatible, thus greatly shortening the commercial time of the system.

Carrier aggregation technology provides bandwidth scalability, facilitates more flexible use of different bandwidths, and achieves higher spectrum utilization. At the same time, the introduction of carrier aggregation technology provides greater flexibility and some gain for mobility management (Fig. 1).

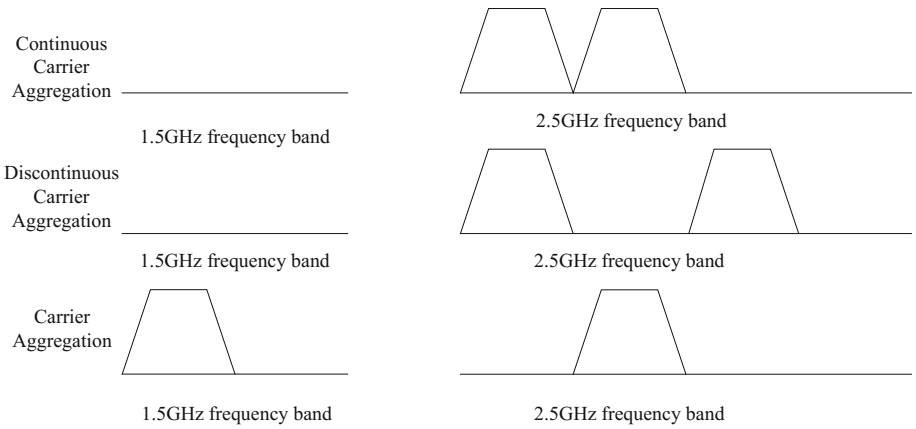


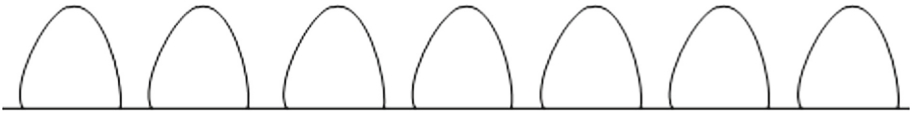
Fig. 1. Three categories of carrier aggregation.

There are three different combinations of carrier aggregation models: continuous carrier aggregation in the frequency band, discontinuous carrier aggregation in the frequency band, and carrier aggregation between the frequency bands. By changing the size of IFFT (the number of subcarriers) and calculating the center frequency of the carrier, the continuous carrier aggregation can be realized in the frequency band easily. There are two steps to aggregate the discontinuous carries based on OFDM frequency band. First, the same as aggregation of the continuous carrier in frequency band, a larger IFFT and carrier transformation include the two discontinuous channels. Second, using the subcarrier nulling limits the carriers in other channels. At present, the carrier aggregation between the frequency bands is implemented by using the individual IFFT module and radio frequency links. And the current backend hardware conditions are also difficult to share the same radio frequency link. The carrier aggregation between the bands needs to add base-band IFFT module and extra radio frequency link.

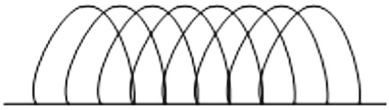
2.2 OFDMA Frequency Diversity

The traditional frequency division multiplexing is to divide the whole frequency band into several disjoint sub frequency, which brings parallel transmission data stream, and

each sub channel needs to keep enough wide protection frequency band. While the OFDM system has orthogonality between subcarriers, allows the sub channel spectrum overlapping and can maximize the use of spectrum resources. The principle of the OFDM is given in Fig. 2. Through coding the data at the transmitter gains the useful data and then though the serial to parallel transformation transforms the high speed data stream to low speed data streams. The data streams spread on several channels and the data on each channel is independent from the other channels. Then after the FFT transformation and parallel to serial transformation, the data stream together with the cyclic prefix will gain the analog signal after the digital to analog transformation. Last the analog signal will be modulated and transmited. At the receiver, first the signal will be demodulated and after the analog to digital transformation the gained digital signal will have the FFT to gain the data. OFDM is well fit to the high speed transmission over wireless broadband channels because of the orthogonal frequency division multiplexing can well combat frequency selective fading in wireless transmission environment.



a）Traditional frequency division multiplexing.



b）Orthogonal frequency division multiplexing.

Fig. 2. OFDM frequency division multiplexing.

By assigning different subcarriers to the user, the OFDMA provides a natural multiple access and diversity scheme. Because of the independence of user channel fading, the quality of service (QOS) can be satisfied by using the joint subcarrier allocation which can have the benefit of improving performance of multiuser diversity gain. Compared with OFDM technology, each user can transmit data with better channel condition sub channel, while the data will transmit in the whole frequency band for the OFDM technology. So the OFDMA ensures that each sub-carrier is corresponding to the better channel condition of users and get the frequency diversity gain.

2.3 The Communication Method Based on OFDMA Carrier Aggregation and Frequency Division

Figure 3 presents a simple, effective and easy approach to solve the problem of blackout based on carrier aggregation communication theory and frequency diversity

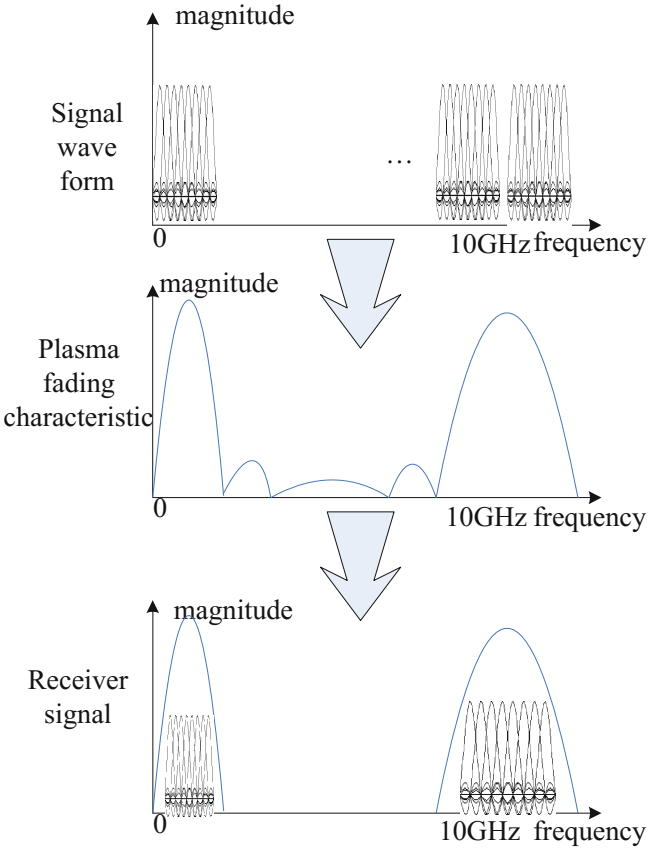


Fig. 3. Broadband communication system based on carrier aggregation.

theory. Based on OFDMA and carrier aggregation technology, the signal waveform is generated, which is used as the basic waveform for ultra high speed aircraft wireless communication. As a kind of wireless channel, the frequency response characteristic of plasma sheath is similar to that of fading channels, and the frequency response of the channel is very frequency selective fading. For ultra high speed vehicle communication, because of low communication rate, we can repeat the same information on each OFMDA carrier to achieve frequency diversity. Despite the fading characteristics of the plasma sheath is complex and nonlinear, but for a number of OFDMA subcarriers, because the OFMDA carriers in different frequency band are aggregated, there will always be sub-carriers in non fading or small fading sub-carrier positions. Therefore, some part of OFDMA sub-carrier information can be received correctly. The use of OFDMA frequency diversity technology and combined with low signal to noise ratio detection technology can solve the blackout phenomenon of high-speed aircraft communication.

3 Transmission Model Based on OFDMA Carrier Aggregation and Frequency Diversity

In order to verify the effectiveness of anti-blackout communication based on OFDMA carrier aggregation and frequency diversity, the following gives the analysis of the performance of TT&C communication link.

The system structure diagram of TT&C communication link is shown in Fig. 4. The source signals are modulated in different OFDMA carriers and pass through plasma fading channel and at the receiver end combining reception is realized for diversity. At last BER performance is calculated after demodulation at the baseband. In this system, baseband modulation is the BPSK modulation which is commonly used in measure and control communication systems. OFDMA carrier aggregation frequency band is the common 2 GHz of S frequency band. In order to anti-blackout communication, we choose another two frequency bands, 30 GHz in millimeter wave band and 1 MHz in low frequency band. The frequency diversity uses the best performance method, which is the maximum ratio combining.

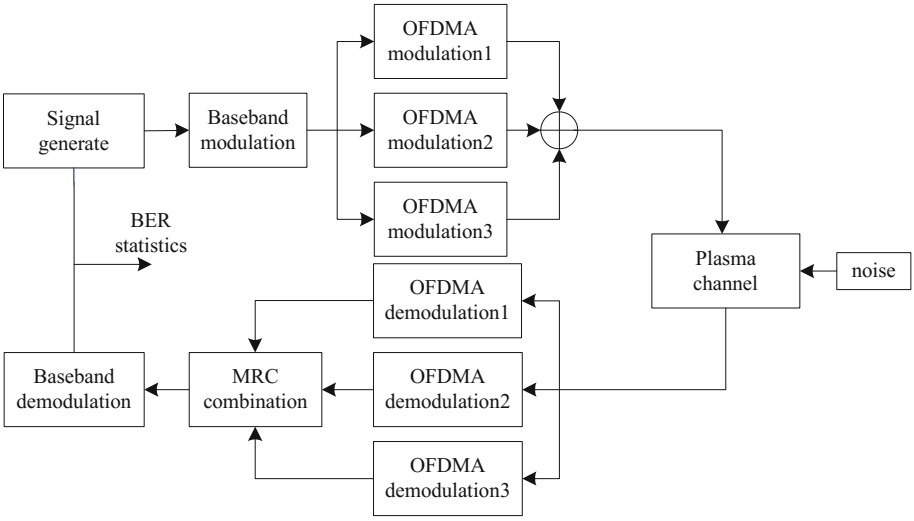


Fig. 4. Structure of TT&C communication link.

In the usual TT&C link, the BER performance of the link is as follows because the system uses the BPSK modulation,

$$P_e = \frac{1}{2} erfc(\sqrt{\gamma}) \tag{1}$$

Where $\gamma = a^2/2\sigma_n$, σ_n is the noise power, a is the magnitude of receiver signal after channel attenuation. Assume the magnitude of transmitting signal is A, then $a = A|h|$, where h is the corresponding coefficient of the channel.

The plasma has a great attenuation because of the blackout, and near the characteristic frequency of the blackout $|h|$ will cause a deep fading for around 30 dB, which can cut down the communication. $|h|$ can have the different attenuation at different frequency bands when using the multi-carrier aggregation based on OFDMA. Therefore we can improve the system performance by diversity reception.

Assume the channel fading caused by plasma channel at frequency band k is h_k, then the magnitude of the receiver signal $a_k = A|h_k|$. Assume using M frequency bands for diversity acceptance, then the receiver signal based on maximum ratio combining will be

$$R = \sum_{k=1}^{M} c_k a_k \tag{2}$$

In the equation, c_k is the weighting coefficient for the *kth* frequency band maximum ratio combining, proportional to the amplitude of the signal and inversely proportional to the noise power, thus $c_k = a_k/\sigma_n^k$.

Assume the noise power in each frequency band is independent of each other, and the average noise of the combined output is the sum of the various noise powers, which is $\sum_{k=1}^{M} c_k^2 \sigma_n^k$. Thus, the output signal to noise ratio (SNR) of maximum ratio combing will be

$$\gamma_R = \frac{\left(\sum_{k=1}^{M} c_k a_k \Big/ \sqrt{2}\right)^2}{\sum_{k=1}^{M} c_k^2 \sigma_n^k} \tag{3}$$

Because of the SNR for each branch is

$$\gamma_k = \frac{a_k^2}{2\sigma_n^k}$$

Where

$$a_k = \sqrt{2\sigma_n^k \gamma_k}$$

Apply it to Eq. (3), we get

$$\gamma_R = \frac{\left(\sum_{k=1}^{M} c_k \sqrt{\gamma_k \sigma_n^k}\right)^2}{\sum_{k=1}^{M} c_k^2 \sigma_n^k} \tag{4}$$

According to the Hu Waltz inequality

$$\left(\sum_{k=1}^{M} pq\right)^2 \leq \left(\sum_{k=1}^{M} p^2\right) \bullet \left(\sum_{k=1}^{M} q^2\right)$$

Such that

$$\left(\sum_{k=1}^{M} c_k \sqrt{\gamma_k \sigma_n^k}\right)^2 \leq \left(\sum_{k=1}^{M} c_k^2 \sigma_n^k\right) \bullet \sum_{k=1}^{M} \gamma_k \tag{5}$$

Combine with Eq. 4,

$$\gamma_R \leq \frac{\left(\sum\limits_{k=1}^{M} c_k^2 \sigma_n^k\right) \bullet \sum\limits_{k=1}^{M} \gamma_k}{\sum\limits_{k=1}^{M} c_k^2 \sigma_n^k} = \sum_{k=1}^{M} \gamma_k \tag{6}$$

From the above formula, we can see that the maximum output SNR is the sum of the output SNR of each frequency bands, therefore the system BER performance can be improved greatly.

4 Simulation Results and Analysis

The error performance of TT&C link based on OFDMA carrier aggregation and frequency diversity is simulated and analyzed. A plasma sheath decay model [7] is employed in this paper.

Figure 5 shows the fading characteristics of the plasma at 10 cm thickness in the electron density $n_e = 10^{11}\text{cm}^{-3}$. The characteristic frequency of the plasma is about 2 GHz, and attenuation is about 37 dB, while the plasma has almost no attenuation at the frequency band for 30 GHz and 1 MHz.

In simulation, the system bandwidth is 1.25 MHz, the size of FFT is 128, modulation is BPSK, and the channel uses the plasma attenuated channel with Gaussian white noise channel as shown in Fig. 5.

Figure 6 shows the error performance curve of the TT&C communication link under OFDMA carrier aggregation. In the traditional single carrier frequency, namely 2 GHz band communication, the characteristic of the plasma frequency at 2 GHz has serious signal fading, so the communication error is large. When carrier aggregation is used to fuse conventional frequency band 2 GHz and a high frequency 30 GHz or low frequency 1 MHz, because of the small attenuation in the high and low frequency bands, the BER performance of the system is greatly improved after diversity reception. After combining three frequency bands, the BER performance of the system can be further improved. This shows that the effectiveness of OFDMA carrier aggregation and diversity reception for anti-blackout communication.

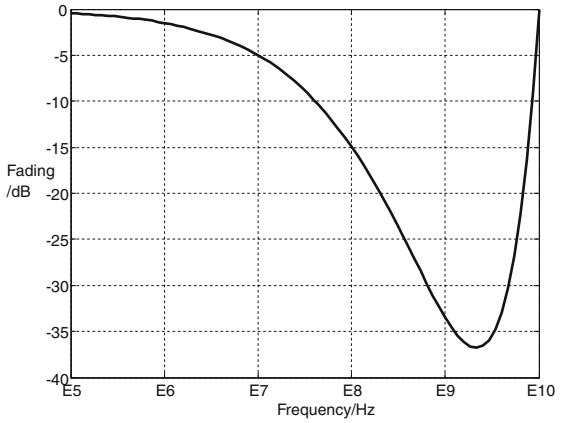


Fig. 5. The fading characteristic of plasma to air waves with frequency.

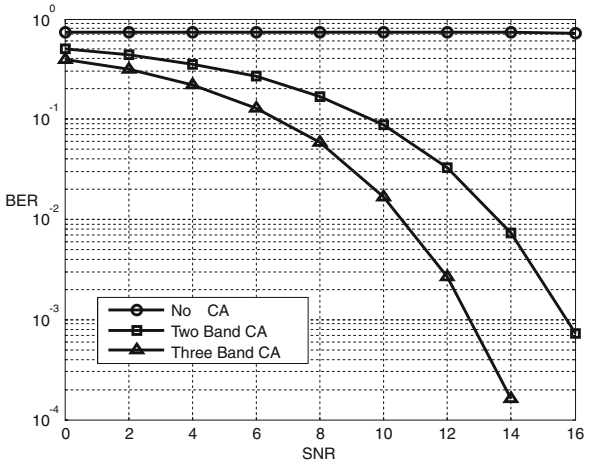


Fig. 6. Performance of communication system based on OFDMA carrier aggregation.

5 Conclusions

This paper proposed a simple and effective method to solve the blackout problem based on carrier aggregation communication theory and OFDMA frequency diversity theory which deals with the blackout communication problem generated by the hypersonic aircraft in the atmosphere. The basic waveform of hypersonic aircraft communication is generated by the technology based on OFDMA and carrier aggregation. The plasma is considered as one of the wireless channels. The frequency diversity technique of OFDMA is utilized to achieve a reliable communication. Analysis and simulation results show that the proposed method can effectively improve the BER performance of the TT&C link communication, and demonstrate the effectiveness of the proposed method.

Acknowledgments. This work is supported by the fund of China Academy of Launch Vehicle Technology (CALT201710, CALT201601). This work was also supported in part by the National Natural Science Foundations of CHINA (Grant No. 61701407 and 61501376).

References

1. Mitchell, F.H.: Communication-system blackout during reentry of large vehicles. Proc. IEEE **55**(5), 619–626 (1967)
2. Shi, L., Guo, B., Liu, Y., et al.: Characteristic of plasma sheath channel and its effect on communication. Progr. Electromagnet. Res. **123**, 321–336 (2012)
3. Kundrapu, M., Loverich, J., Beckwith, K., et al.: Modeling radio communication blackout and blackout mitigation in hypersonic vehicles. J. Spacecr. Rockets **52**(3), 853–862 (2015)
4. Stenzel, R.L., Urrutia, J.M.: A new method for removing the blackout problem on reentry vehicles. J. Appl. Phys. **113**(10), 103303 (2013)
5. Mehra, N., Singh, R.K., Bera, S.C.: Mitigation of communication blackout during re-entry using static magnetic field. Progr. Electromagnet. Res. B **63**, 161–172 (2015)
6. Garg, P., Dodiyal, A.K.: Reducing RF blackout during re-entry of the reusable launch vehicle. In: 2009 IEEE Aerospace Conference, pp. 1–15. IEEE (2009)
7. Xie, K.: Study on plasma sheath reproducing technology and related issues in EM-wave propagation experiment, Xidian University (2014)



Real-Time Indoor Positioning System Based on Background Training Model Using IR-UWB Radar

Hui-Seon Gang, June-Buem Park, and Jae-Young Pyun(✉)

Department of Information and Communication Engineering, Chosun University, 309, Pilmun-daero, Dong-gu, Gwangju, Republic of Korea
greenysun@chosun.kr, pjb@iitp.kr, jypyun@chosun.ac.kr

Abstract. In order to detect and track a moving target using an IR-UWB radar in indoor environment, the signal processing system using filters is normally used. Basically, filters cut the clutters off from the received signal based on the pre-defined characteristics of the target or clutters. Therefore, if the target and clutters are clearly distinguished, filters can be useful. However, the positioning range of the radar can be severely limited when the target signal is not distinguished because of the effect of signal attenuation caused by the path loss. This paper presents a radar signal processing method using codebook-based background training model. While the filter based signal processing system predicts the errors, proposed method measures and stores the informations of background components in the radar signal. In this way, the proposed system can distinguish the target and background more accurately. Experimental results in this paper shows that the proposed system is more robust to signal attenuation.

Keywords: IR-UWB · Radar · Signal processing
Indoor positioning · Codebook · Background subtraction

1 Introduction

In order to detect the location of a moving target from received IR-UWB radar signal, filter based signal processing algorithms are generally used. The filter based target detection system works effectively when each component in the signal can be distinguished from others in terms of amplitude, frequency, phase. However, the filter should be designed for the low complexity and dealt with the received signal within limited time. Therefore, filter based signal processing methods are generally used for implementation of a real time positioning system.

The IR-UWB radar transmits the impulse radio and receives a signal reflected back by the target. In case of short-range target detection, the position of the target can be detected by using difference of amplitude because the intensity of the signal reflected by the target is higher than the background noise. However, as the distance of the target increases, its signal strength is attenuated due to

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 361–371, 2018.
https://doi.org/10.1007/978-3-319-78130-3_37

the path loss at the wireless channel. It makes harder to detect a target from the received radar signal when the signal strength of a target decreased into the background noise level. To handle this problem, positioning system needs to employ the signal processing technique which can separate the target signal from others. As we can see, the complexity of the system is inevitably increased.

Background subtraction is widely used in the general signal processing systems to separate desired target from the received signal. In radar signal processing, background subtraction is also used for the detection of target position. But, simple model of background subtraction comparing the current signal with a previous signal may not detect a target correctly because the measured raw signal is attenuated and mixed with background noise and clutters. Background training and machine learning techniques are used in applications where the characteristics of the background or target in a signal can not be modeled. While background training and machine learning accumulate the information extracted from the input signal to separate the background and the target more precisely, it is not suitable for real-time system implementation because the amount of computation and memory usage increases over time. The codebook model proposed in [1] for image processing compresses and stores the background data into codewords instead of storing all the pixel values from input images. Therefore, the required memory size to store the trained data does not increase significantly over time, and the amount of data to be compared is small in the background subtraction step, which is advantageous in terms of the complexity.

In this paper, we present a method detecting objects more accurately from IR-UWB radar signals using codebook-based background training method. In Sect. 2, we explain a conventional background subtraction method using kalman filter. In Sect. 3, we introduce the proposed background subtraction based indoor positioning system. Moreover, we show the performance evaluation through the experimental results and conclude, respectively in Sects. 4 and 5.

2 Related Work

2.1 IR-UWB Radar Signal Processing

This section explains briefly the real-time radar signal processing system using filters. Radar signal processing is performed in the following order: DC offset

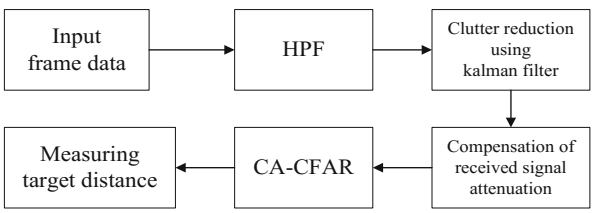


Fig. 1. IR-UWB radar signal processing with kalman filter.

removal, clutter reduction, compensation of the signal attenuation, desired target detection, and distance measurement [2]. Figure 1 shows the real-time positioning system consisting of the various signal processing steps.

Generally, radar based on the direct conversion amplifies signal through LNA and mixer. Here, a DC offset occurs due to a leakage component of the local oscillator signal [3]. Moreover, the level of the received signal fluctuates, i.e., the DC offset of the signal is not constant. In this study, HPF is used to eliminate this DC offset.

The receiving radar signal includes background components like clutters and noise. Since such clutter interferes detecting the desired target, a clutter reduction process is required to remove the background components for accurate detection. In this study, the kalman filter in [4] was used in progress of the clutter reduction.

IR-UWB radar signal experiences attenuation in proportion to the distance between the target and radar. Therefore, the attenuated signal should be compensated as follows:

$$s'[n] = s[n] \times \alpha[n], \tag{1}$$

where n is the index of sample in the captured radar frame, $s[n]$ is the received signal, and $\alpha[n]$ is the weighted vector, respectively. Here, n can be converted to a distance by using a time resolution of sample and impulse transmission velocity of radar. In addition, the weighted vector $\alpha[n]$ is proportional according to the index of sample (n) [4].

$$\alpha[n] = WF \times n, \tag{2}$$

Even with signal attenuation compensation and filtering, the processed radar signal still contains the target and clutter signals. The remaining clutter signals can cause the target to be detected at the location departed far away from the actual position or to generate a false alarm even when there is no target. To detect a target more accurately by reducing the influence of the clutter, this study employed cell averaging constant false alarm rate (CA-CFAR) [5]. CA-CFAR can adaptively select a threshold level to maintain a constant probability of false alarm. Then, it can detect target returns against background noises and clutters.

In the end, the distance from the radar to the target can be calculated by multiplying impulse velocity and time of arrival (ToA) of the signal returns from the detected position.

2.2 Codebook

The codebook model proposed in [1] has been designed to process the foreground-background segmentation more efficiently in a real-time image processing. For each pixel, it builds a codebook consisting of codewords for quantization and clustering of its intensity. Unlike the conventional method storing and accumulating changes of pixel intensity, the codebook model is compressing the trained data set by generating and updating codewords. This codebook based background

subtraction method is useful for real-time signal processing system because it has low complexity and usage of the required memory. In addition, the codebook model has the advantage that it is not affected by the movement of the target. For example, when the target has stopped moving for a while, conventional system using kalman filter has a problem that this target disappears after a few seconds because it updates the prediction data using previous target signal. In contrast, the codebook model can continuously detect a target regardless of the movement of the target unless the codebooks are changed or newly updated. However, in the image processing using vision camera, the codebook model has a disadvantage that it is vulnerable to the variation of illumination. In addition, fake foregrounds can be appeared due to the partial illumination changes such as shadows or highlights [6]. However, unlike vision camera, IR-UWB radar signals are not affected by illumination changes. Moreover, while vision camera uses multiple color channels and two-dimensional signal, IR-UWB radar receives only one-dimensional and one-channel signal, thereby realizing a real-time system more easier. Therefore, the codebook model can be more useful in IR-UWB radar signal processing than an image processing.

3 Proposed IR-UWB Radar Signal Processing Method

In this section, we propose the real-time IR-UWB radar signal processing system consisting of both codebook based background training model and background subtraction function. The background training model manage the codebooks for each sample and the background subtraction function separates the target and backgrounds in the received radar signal using these codebooks.

In the first step of the radar signal processing, the DC offset of the received frame signal is removed by using HPF and the compensation of the attenuated signal is performed. Background training function based on the codebook model is executed as a second step. Figure 2 shows how the background training function generates a codebook conceptually.

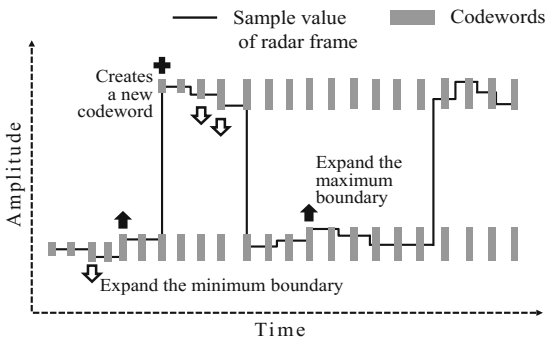


Fig. 2. Principle of the codebook-based background training function.

As the same with the image processing, codebook model generates codebook for each sample. Indeed, each codebook generates or updates a codeword according to the variation of the input value. Then, the current sample is compared with the codeword. If the change of sample is smaller than the threshold, the maximum and the minimum values of the adjacent codeword are updated. Conversely, if the change of sample is larger than the threshold, the codebook generates a new codeword. Through this process, each codebook can represent a background model with number of n codewords. Here, each codeword does not require a huge capacity of memory because it does not record all the sample values, but the maximum and minimum boundaries. By adopting codebook to radar signal, codeword is generated according to the strength of the samples each radar frame when there is no target, which is considered as background signal.

After the training of background radar signal, background subtraction function is executed to remove the background elements in the frame data. In the background subtraction step using the codebook, the foreground can be separated from the background by comparing the pre-processed radar sample with the trained codebook. Through this process, the radar sample is classified as background if it belongs into the range of one codeword in the codebook. On the other hand, it is classified as foreground if there is no matched codeword is found. Since the codebook model only compares the boundaries of codewords without performing a comparison on all values, it is possible to remove the background very quickly.

4 Performance Evaluation

4.1 Experiment Environment

The ranging test environment using IR-UWB radar is described in Fig. 3. The radar used in this experiment consists of NVA-R640 module with NVA 6100 chip of novelda [7], vivaldi type antenna, and low noise amplifier (LNA) for Rx



Fig. 3. Experiment environment for target distance measurement using IR-UWB radar.

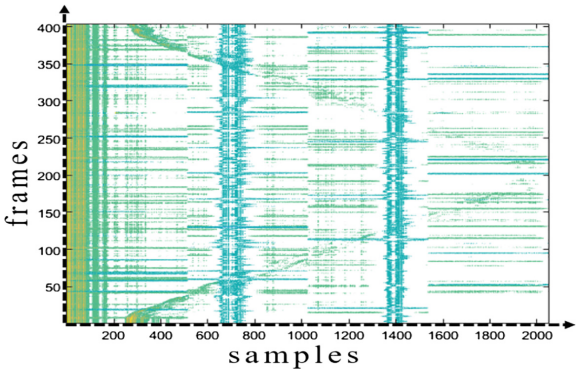


Fig. 4. Measured raw signal at IR-UWB radar.

terminal [2]. The center frequency of radar system is 4.5 GHz, the maximum pulse repetition frequency (PRF) is 48 MHz, and the 2048 samples are obtained per frame. The range of the target moves from 0 to 8 m. Also, the proposed system processes 16 frames per second in real time.

In this experiment, we placed the IR-UWB radar in the middle of the hallway and trained the background during 16 s. The target moves back and forth within the positioning range of the radar after the training time. Figure 4 shows a raw signal received from IR-UWB radar when the target moves from 1 to 8 m.

As shown in Fig. 4 the strength of signal describing the target located far from the radar is weak. Therefore, we need to amplify the target signal for the compensation of the attenuated signal. However, the software-based compensation using weighted vector (α) of (1) amplifies all the samples in the frame. Hence, target signal, clutters, and the background noise are amplified together, which makes it difficult to accurately detect the target. In this experiment, we evaluate the performance by comparing the signal processing results according to the changes of α under the same test scenario. In order to change the α, we adjusted the weight factor (WF) of (2) value of the system.

4.2 Experimental Result

In this paper, two experiments were conducted to evaluate the performance of the proposed system. In the first experiment, we compared the positioning performance between the proposed codebook (CB) based background training and conventional kalman filter (KF) based signal processing models.

Figures 5 and 6 show the measured distances using two models according to the adjustment of WF (0.002 and 0.032). In this experiment, WF of 0.002 corresponds to the case of weak compensation to the attenuated signal, whereas WF of 0.032 corresponds to the case of strong compensation to the signal. In the first experiment, we can see that the results of the two models are different depending on the size of the WF. When the signal compensation is small (WF

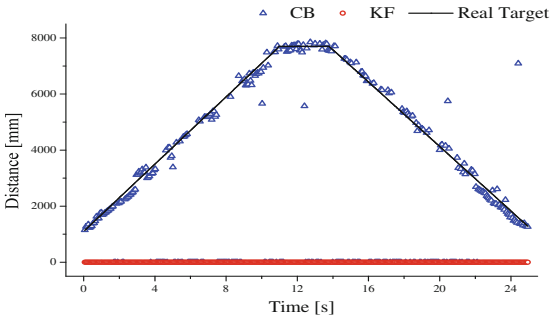


Fig. 5. Measured distance using the CB and KF models with WF of 0.002.

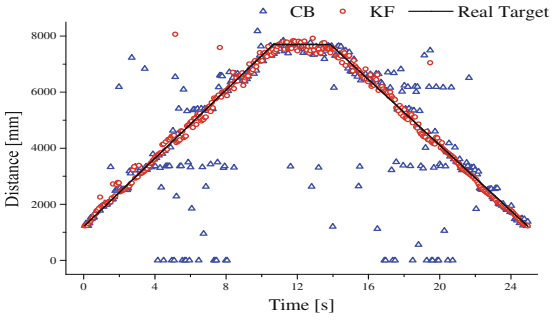


Fig. 6. Measured distance using the CB and KF models with WF of 0.032.

is 0.002), the proposed CB model shows that it is powerful at the background noise reduction since it is based on the background learning process, but the KF model can not detect the target signal since the measured attenuated signal is coexisting with the background noise. On the contrary, when the compensation of the attenuated signal is large (WF is 0.032), the KF model separates the target, background noise, and clutter signals well and shows better measurement results than CB model performs. The reason is that CB model works worse when the noise and clutter signals are amplified unnecessarily together with the target signal. That is, the number of codeword can be increased in the training step and the range (minimum $\sim$ maximum) of each codeword is much large to eliminate the noises. This unwanted codewords can increase the probability that the target is classified as background. As a result, WF should be adjusted well to compensate the attenuated target signal for CB model, but it is difficult to find the suitable WF value under the various target movement environment. In contrast, kalman filter can mitigate the influence of the amplified background noise effectively because it estimates the current background by using previous estimates [4].

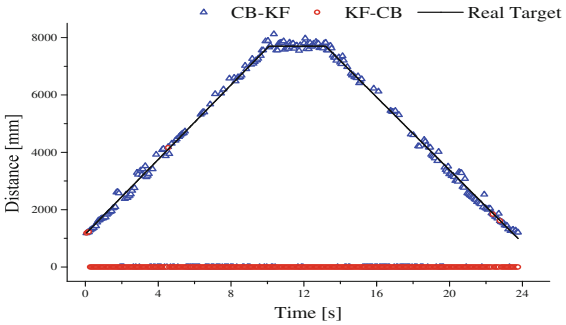


Fig. 7. Measured distance using the combined models with WF of 0.002.

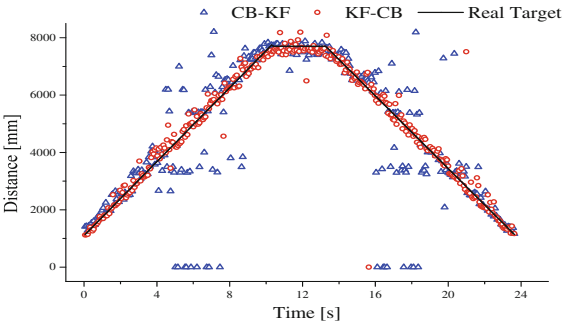


Fig. 8. Measured distance using the combined models with WF of 0.032.

Table 1. Confusion matrix of each ranging system model.

	WF	0.002	0.004	0.008	0.016	0.032
	α	0.25–4.0	0.5–8.0	1.0–16.0	2.0–32.0	4.0–64.0
CB	TP/FP	0.62/0.38	0.85/0.15	0.97/0.03	0.97/0.03	0.93/0.07
	FN/TN	0.04/0.96	0.10/0.90	0.13/0.87	0.15/0.85	0.76/0.24
KF	TP/FP	0.00/1.00	0.10/0.90	0.64/0.36	0.99/0.01	1.00/0.00
	FN/TN	0.00/1.00	0.00/1.00	0.01/0.99	0.05/0.95	0.89/0.11
CB-KF	TP/FP	0.66/0.34	0.85/0.15	0.93/0.07	0.99/0.01	0.95/0.05
	FN/TN	0.01/0.99	0.05/0.95	0.10/0.90	0.21/0.79	0.15/0.85
KF-CB	TP/FP	0.02/0.98	0.15/0.85	0.63/0.37	0.98/0.02	1.00/0.00
	FN/TN	0.00/1.00	0.00/1.00	0.00/1.00	0.01/0.99	0.05/0.95

Based on the results of the first experiment, we performed the second experiment for hybrid method combining CB and KF models. A model named CB-KF removes background by using CB first and then filters the signal by using KF. This model was designed to improve the accuracy of proposed background sub-

Table 2. Accuracy comparison for each ranging system model (higher is better).

WF	0.002	0.004	0.008	0.016	0.032
α	0.25–4.0	0.5–8.0	1.0–16.0	2.0–32.0	4.0–64.0
CB	0.79	0.87	0.92	0.91	0.59
KF	0.50	0.55	0.82	0.97	0.56
CB-KF	0.83	0.90	0.91	0.89	0.90
KF-CB	0.51	0.57	0.82	0.98	0.97

Table 3. RMSE comparison for each ranging system models (lower is better).

WF	0.002	0.004	0.008	0.016	0.032
α	0.25–4.0	0.5–8.0	1.0–16.0	2.0–32.0	4.0–64.0
CB	424.70	485.31	677.41	609.47	690.89
KF	5227.63	942.80	167.46	246.38	357.75
CB-KF	183.97	403.05	422.91	917.68	804.85
KF-CB	2085.53	743.62	162.67	256.73	338.10

traction when both the target and background signals are weakly measured. The other model named KF-CB filters the signal using KF first and then trains the filtered signal by using CB. This model was designed to improve the accuracy of proposed background subtraction by mitigating the influence of amplified background noise.

Figures 7 and 8 show the signal processing results of the two combined models according to changing of WF. From this experimental result, we can see that the CB-KF model shows better performance when the WF is low. On the contrary, KF-CB model shows better when the WF is high. For the performance comparison, accuracy and root mean square error (RMSE) are shown in Tables 1, 2 and 3.

In order to show the accuracy of each system model, we summarized measurement results as True-Positive (TP), False-Positive (FP), True-Negative (TN), and False-Negative (FN) in Table 1. By using these values, accuracy of each system can be calculated as follows:

$$Accuracy = \frac{\Sigma TP + \Sigma TN}{\Sigma total\ population} \tag{3}$$

We observe that the CB model shows the more stable accuracy in the low and medium WF conditions and the KF model shows better RMSE at the higher signal compensation. In the case of CB-KF, the performance for RMSE and accuracy is best under the condition of low WF. On the other hand, KF-CB performs very well under the condition of high WF, since the noise errors amplified at the compensation step are removed by KF and noise-reduced signals are learned for the future background subtraction.

Table 4. Average consumed time for computation overhead comparison [μs].

HPF+KF	156
HPF+Background subtraction using CB	126
Background training using CB model	4589

Table 4 shows the computation time of each step in the background subtraction process. As we can see from the results, background subtraction using CB has little effect on the computation overhead of the system. Although background training takes longer computation time, it is executed before starting real-time target tracking. Therefore, it could be suitable to use a CB model in a real-time target tracking system.

5 Conclusion

Background subtraction method is widely used in the first step of signal processing, but it is not simple to generate a background model when the background components including clutter and noise are not static. Furthermore, when the target signal is weakened to the background noise level, it is difficult to detect the target by using the background subtraction even if the background model is generated well. In this paper, we proposed a signal processing method using the CB based background model. From the experimental results, we found that the CB model is robust against signal attenuation but vulnerable to the amplified noise, whereas the KF is robust against noise but vulnerable to signal attenuation. Based on these features, we also figure out that the accuracy and RMSE can be improved more if the CB and KF are combined. However, when we design a background subtraction process that combines a CB and KF, the processing order should be determined depending on how much the target signal and the background noise are amplified together by the compensation of the attenuated signal. In the next study, we do research for the effect of combined method on the various background environment and its application.

Acknowledgment. Following are results of a study on the "Leaders in INdustry-university Cooperation +" Project, supported by the Ministry of Education and National Research Foundation (NRF) of Korea. Also, this research was supported by Basic Science Research Program through the NRF (No. 2014R1A1A2059952).

References

1. Kim, K., Chalidabhongse, T.H., Harwood, D., Davis, L.: Real-time foreground-background segmentation using codebook model. Real-Time Imaging **11**(3), 172–185 (2005)
2. Kim, H.-J., Kim, B.-H., Gang, H.-S., Hwang, S.-S., Kwon, G.-R., Pyun, J.-Y.: Human detection and tracking system using IR-UWB radar. In: International Conference on Electronics, Information and Communication, pp. 526–527, January 2015

3. Abidi, A.A.: Direct-conversion radio transceivers for digital communications. IEEE J. Solid-State Circuits **30**(12), 1399–1410 (1995)
4. Nguyen, V.H., Pyun, J.-Y.: Location detection and tracking of moving targets by a 2D IR-UWB radar system. Sens. J. **15**, 6740–6762 (2015)
5. Maali, A., Mesloub, A., Djeddou, M., Mimoun, H., Baudoin, G., Ouldali, A.: Adaptive CA-CFAR threshold for non-coherent IR-UWB energy detector receivers. IEEE Commun. Lett. **13**(12), 959–961 (2009)
6. Guo, J.M., Hsia, C.H., Liu, Y.F., Shih, M.H., Chang, C.H., Wu, J.Y.: Fast background subtraction based on a multilayer codebook model for moving object detection. IEEE Trans. Circuits Syst. Video Technol. **23**, 1809–1821 (2013)
7. Taylor, J.D., Wisland, D.T.: Novelda nanoscale impulse radar. In: Taylor, J.D. (ed.) Ultrawideband Radar: Applications and Design, 1st edn, pp. 373–388. CRC Press, Boca Raton (2012)



Bandwidth Adaptive Image Communication via Similarity Based Auto-Selection

Jieqi Li[1], Mingyang Ma[2], Yong Zhang[1], Zhengxue Li[1], Shaohui Mei[2(✉)], and Shuai Wan[2]

[1] China Academy of Launch Vehicle Technology, Beijing 10076, China
[2] School of Electronics and Information, Northwestern Polytechnical University, Xian 710072, Shaanxi, China
meish@nwpu.edu.cn, sxtyljq@163.com

Abstract. In image acquisition and communication systems on small or micro platforms, such as small satellite or unmanned aerial vehicle platforms, the imaging system can generate huge amount of image data while communication system can only deliver a very small part of them due to the limited communication bandwidth. In this paper, a novel bandwidth adaptive image communication strategy via similarity based auto-selection is designed to select a certain number of most informative images acquired by the imaging system for transmission, where the number of selected images is adaptive to the communication bandwidth. Specifically, the image that is more distinguishing to previously transmitted images measured by the similarity, instead of the instantly acquired image, is selected. Experimental results on simulated image sequence has demonstrated the effectiveness of the proposed bandwidth adaptive image communication algorithm.

Keywords: Bandwidth adaptive · Image communication · Image acquisition
Bandwidth adaptive

1 Introduction

The rapid development of small-satellite techniques makes it possible to test innovative science and engineering technologies on a smaller scale in the space environment. For example, NASA's PhoneSat project [1], which builds spacecraft by using off-the-shelf consumer smartphones, has demonstrated the ability to launch the lowest-cost and easiest to build satellites ever flown in space. The QB50 project aims to launch 50 CubeSats ($10 \times 10 \times 10$ cm, weighing 1 kg) to demonstrate the possibility of launching a network of 50 CubeSats built by Universities Teams all over the world as a primary payload on a low-cost launch vehicle to perform first-class science in the largely unexplored lower thermosphere [2]. These small-satellites provide convenient platforms to equip different kinds of sensors for space scientific research.

This work is partially supported by National Natural Science Foundation of China (61671383) and Funds by China Academy of Launch Vehicle Technology (CALT201601, CALT201710).

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 372–380, 2018.
https://doi.org/10.1007/978-3-319-78130-3_38

The remote sensing missions, e.g. earth observing, have also been widely verified on the small-satellite platform. However, the limitation of communication bandwidth restricts transmission of valuable images significantly. Nowadays, enormous amounts of images can be acquired through the imaging system over a small satellite platform, however, only a very small portion can be transmitted back to earth due to the limitation of bandwidth, power-supply, and etc. For example, if a one megapixel micro-camera is equipped for earth observing, it will cost several seconds or more to transmit just one image back to the earth or repeater satellites. However, during the transmission process, the micro camera is capable to acquire tens of hundreds of images which cannot be transmitted back to earth. For high-resolution or multi-camera sensing systems, much more data will be missed.

Under such circumstance where the data that can be transmitted is only a fraction of the data acquired, how to transmit the information efficiently is of significant importance. Existing methods can be divided into two categories using different perspectives, one is the image compression which aims to reduce the amount of data of every image to transmit more images, and the other is image extraction that selects the most representative images to transmit.

A. image compression. Image compression technology can effectively reduce the amount of data transmitted in remote sensing system. Li *et al.* [3] proposed a rate pre-allocated compression algorithm (RPCA) based on bit plane coding to compress the large aerial image data from the large-array aerial CCD camera, the quality of proposed image compression was comparable to the JPEG2000 standard with a lower complexity. Gong *et al.* [4] also proposed a low-complexity image-sequence compressing algorithm based on homography transformation, in which the image sequences were divided into frame-groups, and the intermediate frames in the same frame-group was predicted by the first-frame and the end-frame with homography transformation, finally the first frame and the residual images of the intermediate frames of the same frame-group were merged into a big image and coded to generate fixed-size code streams. Zhang *et al.* [5] presented an improved fast SPIHT compression algorithm, where the remote sensing images were decomposed by Le Gall 5/3 wavelet, and wavelet coefficients were indexed, scanned and allocated by the means of family blocks, and the N bit-planes could be encoded at the same time to improve the compression speed. In the aerospace platform, the chip for compression is also often used to achieve image compression. For example, Li *et al.* [6] proposed an image compression scheme based on ADV212 in multi-spectral space camera. The custom-specific mode was presented based on the working principle of adv212 and CCD image properties, and a (10, 5) parameters setting the quantization step method was proposed.

B. image extraction. Image extraction methods extract some most representative images to transmit from the massive acquired images. The clustering technique has been utilized to summarizing remotely sensed imagery. Hayden *et al.* [7] proposed an onboard clustering method to summarize remotely sensed imagery, this approach autonomously clustered the data into classes of similar imagery which enabled selective downlink of representatives of each class to reducing downlinked data volume. Sarmah *et al.* [8] proposed a grid density based clustering technique (SATCLUS) for multi-spectral satellite images in which a combination of density based and grid based clustering was used. The clustering process was divided into two steps, in the first step

a rough clustering of the image space was obtained and the second step dealt with cluster smoothening for quality cluster identification.

However, image compression technology can only reduce the amount of data per image, the total data of all images is still relatively large. Moreover, if the compression rate is too low, the image quality will be greatly reduced, which seriously influence the result of subsequent image processing. Therefore, we focus on the image extraction technology. As for the clustering technique, which is performed in an off-line manner, it cannot be utilized for real-time analysis. Generally, in a real-time image communication system, when one image is being transmitted, many new images will be reacquired by the imaging system during the transmission. Furthermore, the number of clusters is not easy to determine.

Therefore, in this paper, a bandwidth adaptive real-time image communication strategy is proposed to identify the most representative image to be transmitted during the transmission of current image. As a result, images with more information will be acquired through satellite platform, and thus the information-acquisition ability of the imaging system is fully explored in satellite based remote sensing missions.

2 Algorithm Description

Due to the limitation of communication bandwidth, hundreds or thousands of new images could be acquired from the beginning of one image transmission to the end of transmission, and this process is shown in Fig. 1. In order to exploit maximally the capability of the imaging system, selecting one of the most valuable images from the newly acquired images as the input of the communication system is of great significance.

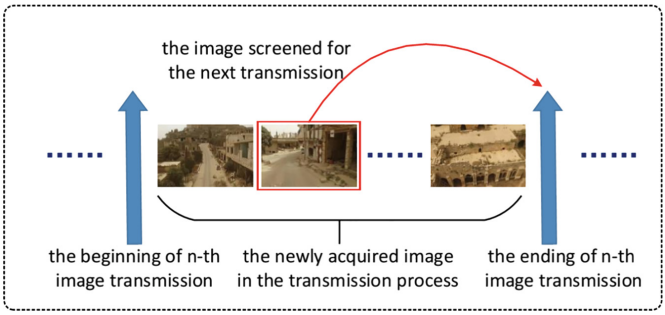


Fig. 1. The process of image acquisition and transmission. The image marked with red rectangle is screened for the next transmission by an image extraction method. (Color figure online)

As shown in Fig. 2, the large number of new images acquired by the imaging system in the transmission process are used as the input images, then the images are represented by feature vectors. According to the designed screening criteria, the most valuable image is screened for transmission, under the constraint of the transmitted images. Finally, the screened image is transmitted by a communication system. In addition, the screened image should be added into the library of transmitted images.

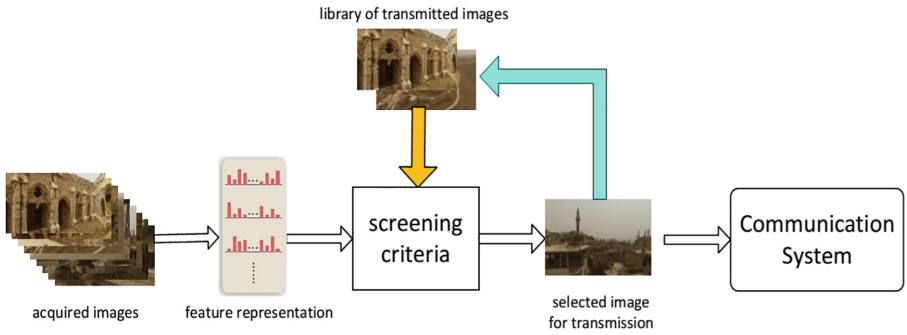


Fig. 2. The scheme of proposed bandwidth adaptive image communication strategy.

2.1 Feature Representation

In order to perform the proposed image communication strategy, each acquired image should be represented by its feature vector to alleviate the influence of illumination, rotation, and etc. In this paper, a 360-dimensional feature vector is utilized to represent each image, though many other features can also be employed. This feature vector consists of two parts: a 252-dimensional feature vector extracted by CENTRIST [9, 10], and a 108-dimensional feature vector accounting for color moment.

The CENTRIST feature captures local structures of an image without color information by utilizing a spatial pyramid structure. Only the last two spatial levels, each of which contains 5 and 1 image patches, are adopted. Thus, each patch is represented by a 42-dimensional feature, where 40 dimensions are for eigenvectors and the other two are for the mean and variance of each patch, respectively. Hence, the dimension of each CENTRIST feature is $6 \times 42 = 252$. More details can be found in [9, 10].

To calculate the color moment feature of a video frame, each frame is represented in HSV color space and partitioned in to 3×4 patches. A three-order color moment (i.e., Mean, Standard Deviation, and Skewness) is adopted.

$$E_i = \sum_{j=1}^{N} \frac{1}{N} p_{ij}. \tag{1}$$

$$\sigma_i = \sqrt{\frac{1}{N} \sum_{j=1}^{N} \left(p_{ij} - E_i\right)^2}. \tag{2}$$

$$s_i = \sqrt[3]{\frac{1}{N} \sum_{j=1}^{N} \left(p_{ij} - E_i\right)^3}. \tag{3}$$

where E_i, σ_i, s_i are mean, standard deviation, and skewness respectively, p_{ij} represents the j-th image pixel of the i-th color channel, and N is the total number of image pixels. As a result, each frame is represented with a $3 \times 4 \times 3 \times 3 = 108$-dimensional color moment.

2.2 The Proposed Algorithm

Assume $\mathbf{F}^{(j)} = [F_1, F_2, \cdots, F_n] \in \Re^{d \times n}$ represents the set of newly acquired images when the j-th image is transmitted back to earth by the communication system on a satellite, where each column vector $F_i \in \Re^d (i = 1, 2, \cdots, n)$ denotes the feature vector of the i-th image in $\mathbf{F}^{(j)}$. The key element for our image selection task is to select one most distinguishing image (denoted by I_{j+1}) to previous transmitted image set (denoted by $\mathbf{I}^{(j)} = [I_1, I_2, \cdots, I_n] \in \Re^{d \times j}$) from $\mathbf{F}^{(j)}$.

(1) *adaptivity*: How many images the screening is from is a key problem, which is affected by communication bandwidth and other factors. Assuming that the communication bandwidth is b kbps and the number of one image data is c kbits, then the time required to transmit one image is $t = c/b$ seconds. Assuming that the acquisition rate of the imaging system is f frame per second, therefore the number of newly acquired images in the transmission process is $n = f*t = fc/b$.

The screening process should be performed from $n = f*t = fc/b$ images to find the most distinguishing image. If the screening is from less than n images, the previous image has not been transmitted yet. On the contrary, the previous image has been transmitted early, but the next image has not been selected. The parameters n which is adaptive to the communication bandwidth will be adopted in our algorithm.

(2) *screening criteria*: The proposed screening criteria is similarity-based. Generally, the more different the images are from previous transmitted images, the more information they present. Therefore, the image that is different to previous transmitted images is selected to be transmitted back to earth. In other words, the image that is least similar to previous transmitted images in the n new images is selected:

$$\mathbf{I}_{j+1} = \arg \min_{F \in \mathbf{F}^{(j)}} : \sum_{k=1}^{j} Cor(F, I_k), \tag{4}$$

where $Cor(\cdot)$ measures the similarity of two frame images. Many similarity measurement can be utilized to calculate $Cor(\cdot)$, such as correlation coefficient, Euclidean distance, and etc. In this paper, the correlation coefficient is adopted to calculate $Cor(\cdot)$.

(3) *updating the transmitted image library*. This is another important issue should be solved before implementation. Due to the limited onboard memory resources, the transmitted image library $\mathbf{I}^{(j)}$ cannot increase continuously without a boundary. Therefore, the maximum size of the image library should be predefined according to the onboard memory resources. When the size of the transmitted image library exceeds its maximum size, the most non-representative image in this library should be excluded to keep a fixed size of the transmitted image library. In this paper, the criteria to exclude an image is also similarity-based, i.e., the image that is similar to all the other transmitted images is excluded:

$$E = \arg\max_{I \in \mathbf{I}^{(j)}} \sum_{k=1}^{j} Cor(I_k, I). \tag{5}$$

The updated transmitted image library will be

$$\mathbf{I}^{(j)} = \mathbf{I}^{(j)} \backslash E, \tag{6}$$

where '\' means E is excluded from $\mathbf{I}^{(j)}$. By limiting the previous transmitted image library in the proposed algorithm, the computation time can also be saved for real-time applications.

The flowchart of our proposed bandwidth adaptive image communication via similarity based auto-selection method for satellites is shown in Fig. 3. Every time when a new image is acquired, its feature is extracted for representation. Then its feature vector is compared with the current transmitted image library. If the similarity is smaller than that of current candidate image, the current acquired image will be selected as a new candidate image. Otherwise, it will not be selected. When n images have been compared with the current transmitted image library, the candidate image will be transmitted. Meantime, it is added to the current transmitted image library and the new transmitted image library is refined according to (6).

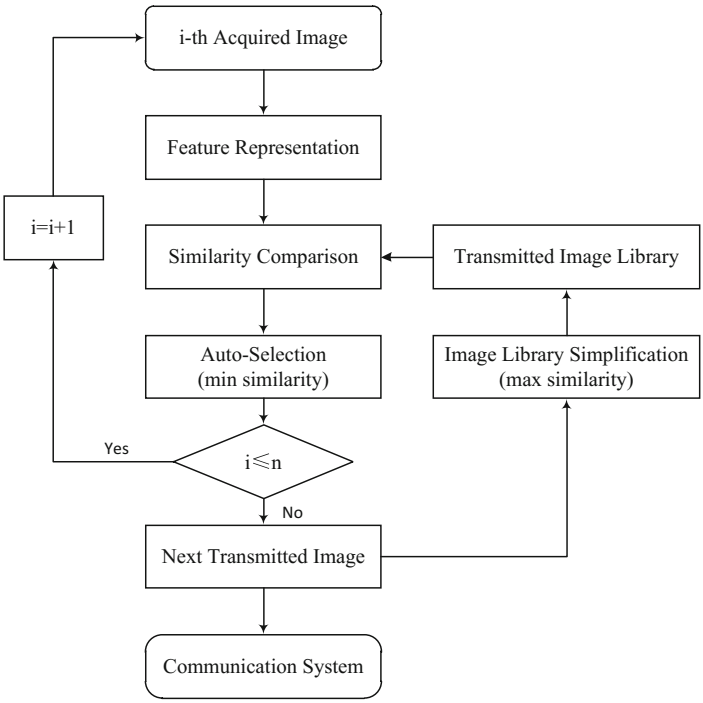


Fig. 3. The flowchart of our proposed bandwidth adaptive image communication via similarity based auto-selection.

3 Coefficient Diagram Method

In order to evaluate the performance of our proposed image communication strategy, a small section of the famous movie named 'Gravity', which contains 2575 frames of earth observing images, is selected in this experiment. The contrast of the brightness is tuned down to simulate low illumination images. In addition, salt & pepper noises are added to simulate space noises on the images.

In this experiment, the image acquiring rate is set as 30 frames per second. Assume that each image can be effectively compressed to 128 kbits (can be compressed according to JPEG2000). The communication rate is set as 32 kbits. In the proposed image selection algorithm, the maximum size of transmitted image library *Tm* is set to 10. Totally 36 images are selected by the proposed algorithm for transmission. The average correlation coefficient of the current transmitted image to previous transmitted image is adopted for quantitative evaluation. Generally, smaller average correlation coefficient indicates that current acquired image contains more distinguishing information to previous transmitted images. The instant-acquiring strategy, in which the next image is acquired when the transmission of current image is finished, is adopted for comparison. As shown in Fig. 4, the average correlation coefficient of the proposed algorithm is smaller than that of instant-acquiring strategy, indicating that the images selected by the proposed algorithm presents more information than that by the instant-acquiring strategy.

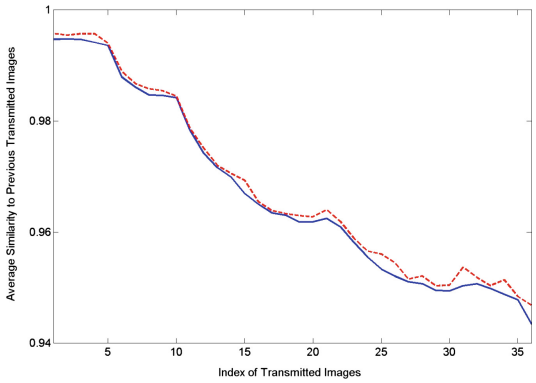


Fig. 4. The comparison of our proposed method and instant-acquiring strategy on the average similarity to previous transmitted images. The red dotted line indicates the instant-acquiring strategy, the other indicates ours. (Color figure online)

In order to visually evaluate the performance of the proposed algorithm, the image sequences are down-sampled five times to make it to be quick change scenarios. As a result, an image sequence consists of 575 images is simulated as camera acquired images. All the other parameters are set as that in previous experiment. The image sequence selected by our proposed algorithm for transmission is shown in left column of Fig. 5. For comparison purpose, the instant acquiring image sequence to be transmitted, where the next image is acquired when the transmission of current image is

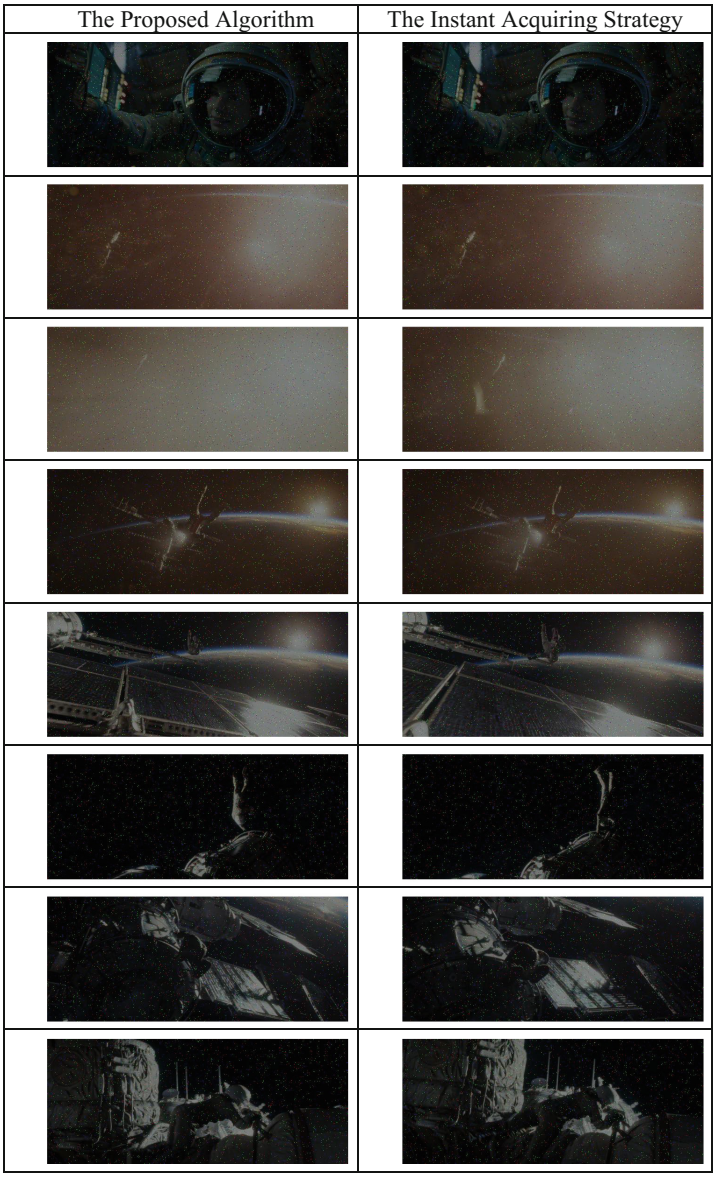


Fig. 5. The selected image sequence to be transmitted by different algorithms for the synthetic scene.

finished, is shown in the right column of Fig. 5. It is observed that the results of the proposed algorithm are visually similar to that of the instant acquiring strategy. However, slight different can be found in the two sequence. For example, the fifth image obtained in the proposed image contains more macroscopic information than that by the instant acquiring strategy.

4 Conclusion

In this paper, a novel onboard bandwidth adaptive image communication strategy via similarity based auto-selection is proposed. The algorithm is adaptive to the communication bandwidth and imaging system, and the most distinguishing image to previous transmitted images, instead of instantly acquired image is selected for transmission. The onboard transmitted image library can be updated and simplified to reduce the storage requirement and save computation time. The proposed algorithm is very useful for remote sensing image acquiring with limited communication bandwidth, such as small-satellite, deep-space probe, UAV, and etc.

References

1. PhoneSat Flight Demonstrations. http://www.nasa.gov
2. QB50 Twitter. https://www.qb50.eu
3. Li, Q.H., Ren, G.Q., Wu, Q.Z.: High-resolution aerial image compression based on technology of the rate pre-allocation. Opto-Electron. Eng. **39**, 118–124 (2012)
4. Gong, J., Zheng, C., Tian, J., Wu, D.: An image-sequence compressing algorithm based on homography transformation for unmanned aerial vehicle. In: International Symposium on Intelligence Information Processing and Trusted Computing, pp. 37–40. IEEE Press, Huanggang (2010)
5. Zhang, N., Jin, L., Wu, Y., Zhang, K.: An improved fast SPIHT image compression algorithm for aerial applications. J. Multimedia **6**, 494–501 (2011)
6. Li, J., Jin, L.X., Li, G.N., Wang, W.H.: Application of ADV212 to the large field of view multi-spectral TDICCD space camera. Spectrosc. Spectral Anal. **32**, 1700–1707 (2012)
7. Hayden, D.S., Chien, S., Thompson, D.R., Castano, R.: Using onboard clustering to summarize remotely sensed imagery. Intell. Syst. IEEE **25**, 86–91 (2010)
8. Sarmah, S., Bhattacharyya, D.K.: A grid-density based technique for finding clusters in satellite image. Pattern Recogn. Lett. **33**, 589–604 (2012)
9. Wu, J., Christensen, H.I., Rehg, J.M.: Visual place categorization: problem, dataset, and algorithm. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4763–4770. IEEE Press, Saint Louis (2009)
10. Wu, J., Rehg, J.M.: Centrist: a visual descriptor for scene categorization. IEEE Trans. Pattern Anal. Mach. Intell. **33**, 1489–1501 (2011)

Hyperspectral Image Fusion Based on Multistage Guided Filter

Wenqian Dong, Song Xiao(✉), Jiahui Qu, and Lizhao Li

State Key Laboratory of ISN, Xidian University, Xi'an 710071, China
xiaosong@mail.xidian.edu.cn

Abstract. In this paper, a novel multistage guided filter based hyperspectral (HS) pansharpening algorithm is presented. The intensity component (INT) from interpolated HS image is generated using adaptive IHS method and the optimization equation is solved to get weight vector at first. Then, the paper proposes a multistage guided filter strategy to extract the spatial detail by using panchromatic image and INT image as a guidance image respectively, which is different from the traditional methods. In this way, the detail information can be obtained in a consistent manner. The obtained spatial details information is finally added into the interpolated HS image to generate fused HS image. Experimental results based on different remote sensing images indicate that the presented approach improves performance in the spatial and spectral aspects.

Keywords: Hyperspectral (HS) image · Panchromatic (PAN) image
Image fusion · Guided filter

1 Introduction

Hyperspectral image fusion aims to integrate the hyperspectral and panchromatic to generate a single fused HS image. Many excellent approaches were presented in the existing literature [1–7]. Component substitution (CS) which contains algorithms such as Gram-Schmidt (GS) and GS Adaptive (GSA) [3] is one class of hyperspectral pansharpening methods. Although the CS method has high fidelity of the spatial information and lower computation complexity, it can generate the spectral distortion. Another technique known as Smoothing Filter based Intensity Modulation (SFIM) [4] belongs to multiresolution analysis (MRA). Although MRA algorithms ensure spectral consistency as well as temporal coherence, they have more complicated implementation and larger calculated amount in comparison with component substitution. Matrix factorization methods and hybrid methods are two other classes of hyperspectral pansharpening methods which have received attention over the past decade. Matrix factorization methods include the algorithms, for example, Coupled Nonnegative Matrix Factorization (CNMF) [5]. Matrix factorization methods can achieve good performance, but it has high computational costs. Hybrid methods contain algorithm such as guided filter PCA (GFPCA). The GFPCA method can preserve the spectral information well and has lower spectral distortion. However, the GFPCA method products lots of blurs because the spatial information is added not enough.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 381–387, 2018.
https://doi.org/10.1007/978-3-319-78130-3_39

In proposed approach, the detail information of HS and PAN are synchronously considered. Since the extracted spatial details are obtained from both the PAN and HS images, the spatial details of the fused HS image are injected adequately. Experiments indicate that the presented approach acquires outstanding fusion quality.

2 Proposed Method

Recently, edge-preserving filters especially the guided filter [6] have been a popular research topic in image processing. The guided filter can avoid ringing artifacts. In this paper, the guided filter is introduced to fuse PAN image and HS image. In a small window, guided filter assumes the output image is linearly related to the guidance image. The linear coefficients are obtained by solving the least squares optimization formula.

As shown in Fig. 1, it summarizes the main processes of the presented fusion approach. First, the intensity component (INT) is got by adaptive IHS method. Then, we solve optimization equation to get the weight vector. Next, the details are extracted by using multistage guided filter. The guidance images are the PAN and INT images respectively. The final fused HS can be composed of interpolated HS and extracted spatial details.

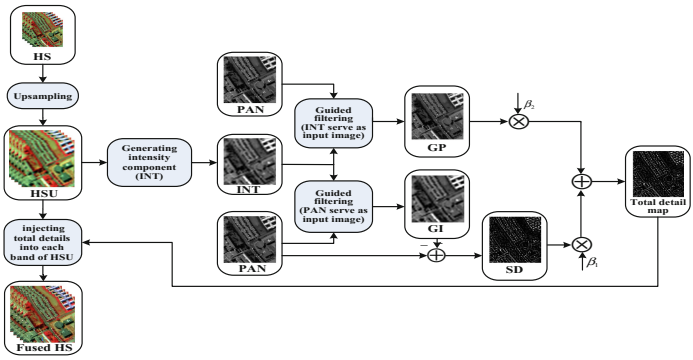


Fig. 1. Flowchart of hyperspectral image fusion scheme with guided filter.

2.1 Generating the Synthetic Intensity Component (INT) Adaptively

The low resolution HS and PAN have different dimension. Up-sampling is applied to the HS image first. The interpolated HS has the same size as the PAN. The upsampled HS image is denoted as HSU.

For the HSU image, the intensity component is used for extracting the spatial detail information. The intensity component (INT) which represents the approximate detail layer of HS image can be defined as the linear equation below.

$$INT = \sum_{i=1}^{m} \lambda_i HSU_i \tag{1}$$

here λ_i denotes the weight, m represents number of the HSU image bands. λ_i measures proportional spectral overlap between each band of HSU (HSU_{i}) and PAN. In AIHS algorithm [7], the INT is obtained by solving an optimization equation which is shown in Eq. (2).

$$\min_{\lambda_1,\ldots,\lambda_m} \left\| PAN - \sum_{\mathrm{i}=1}^{\mathrm{m}} \lambda_{\mathrm{i}} HSU_{\mathrm{i}} \right\|^2 \qquad s.t. \lambda_1 \geq 0, \ldots, \lambda_m \geq 0. \tag{2}$$

Equation (2) is the linear ridge regression model, and we solve this optimization equation to get its solution which is given by

$$\lambda = (HSU^T \times HSU)^{-1} \times HSU^T \times PAN \tag{3}$$

where $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ represents weight vector, $()^T$ represents the transposition operator. In Eq. (3), the PAN image and the interpolated HS image (HSU) are all denoted by the matrices. And one row of the matric denotes one band of the interpolated HS image. One column represents the spectral vector of a pixel. $HSU \in R^{n \times m}$, $PAN \in R^{n \times 1}$ represent the HSU and PAN images, respectively. m represents number of the HSU image bands. n denotes number of pixels in each band of HSU.

2.2 Extracting the Detail Information

In this step, the purpose is to extract the detail information which will be injected to the HS image. Instead of estimating the difference between PAN and INT, guided filter is used for extracting spatial details which is the PAN redundant part than that of the upsampled HS image. Guided filter is also used for capturing detail information serving as supplementary information from the PAN.

(1) This step is used for obtaining spatial information from PAN image by guided filtering. Guidance image is INT component, and the input image is PAN image.

$$GI = F(PAN, INT, \gamma_1, \varepsilon_1) \tag{4}$$

where F is the guided filter function, GI is the guided filtered output image, γ_1 and ε_1 are the size and the regularization parameter, respectively. Here, a relatively smaller window size and a relatively smaller regularization parameter are adopted. GI has the details of HS image since the guided filter transfer details from guidance component INT to output image GI. Thus output image GI is subtracted from the PAN image to get the details difference value between HS and PAN.

$$SD = PAN - GI \tag{5}$$

where SD represents the spatial information difference value. This process can avoid adding too much redundant spatial detail to HS image.

(2) For purpose of obtaining a sufficient amount of extracted details and getting the consistency in extracted details, Similarly, the guided filter is also applied to capturing the detail information which serves as supplementary information. The guidance image is PAN image, the input image is INT image.

$$GP = F(INT, PAN, \gamma_2, \varepsilon_2) \tag{6}$$

where GP represents the guided filtered output, γ_2 and ε_2 are the size and the regularization parameter, respectively. Here, a relatively larger window size and a relatively larger regularization parameter are adopted. GP has spatial information of PAN image. We have to weigh the spectral proportion of SD and GP to determine the benefits that they will bring.

(3) As shown in Fig. 1, the total spatial details can be obtained

$$T_s = \beta_1 SD + \beta_2 GP \tag{7}$$

where, T_s is the total spatial information. The injection gains β_1 and β_2 $(0 < \beta_1, \beta_2 < 1)$ determine the quantity of the added spatial detail information. They impact the fusion results directly. SD which denotes the spatial details difference between the HS and PAN images should be injected more, while GP which serves as supplementary information should be injected less. Thus β_1 is close to 1, whereas β_2 is a smaller number which is approximate to 0.

2.3 Obtaining the Fused Image

The obtained total details are finally added into each band of HSU. The procedure of obtaining the fused HS image can be expressed as

$$F_i = \mathrm{HSU}_i + T_s \tag{8}$$

where, F represents the fused image, F_i represents the ith band in F.

3 Experimental Results and Analysis

Generally, the fusion quality of a hyperspectral pansharpening method can be measured through subjective evaluation and objective evaluation. Several indexes are adopted to assess the fusion quality of each fusion method. Cross correlation (CC) is utilized to judge the degree of spatial distortion correctly. Spectral angle mapper (SAM) index evaluates spectral information similarity. Root mean squared error (RMSE) and Erreur

relative globale adimensionnelle de synthèse (ERGAS) are global indexes. The more similar the final fusion image with the original image, the smaller the RMSE and ERGAS will be. GS, GSA, GFPCA, CNMF and SFIM algorithms are compared for spatial and spectral quality. The window sizes and the regularization parameters of the guided filters are set to $\gamma_1 = 15$, $\gamma_2 = 58$, and $\varepsilon_1 = \varepsilon_2 = 10^{-6}$. β_1 and β_2 are the parameters which determine the quantity of spatial details. In order to understand the influence of parameters β_1 and β_2 ($0 \leq \beta_1, \beta_2 \leq 1$), the proposed method is performed on the Reflective Optics System Imaging (ROSIS) dataset with different (β_1, β_2) settings. β_2 is first given a random value 0.05. The proposed method obtains the best performance when β_1 equals to 0.8. Then β_1 is set to 0.8. The proposed method achieves the optimal results when $\beta_2 = 0.02$. Therefore, $\beta_1 = 0.8, \beta_2 = 0.02$ are set to be the default parameters of the proposed method.

3.1 Experimental Results and Analysis

The experiment is performed on semi-synthetic dataset generated from ROSIS. ROSIS play a huge role in geoscience research. The dataset is denoted as PaviaU dataset, which main applications include geology, vegetation, exploration, thematic mapping and hydrology. These hyperspectral images used in our experiment consist of 103 bands and they are obtained in the range of 0.4–0.9 μm. The obtained simulated PAN image pixel size is 205×175 and the HS image pixel size is 41×35.

The fused HS images obtained by the GS, GSA, GFPCA, CNMF, SFIM and the presented methods are displayed in Fig. 2(b)–(g). As shown in Fig. 2, GS can add sufficient spatial details and texture clear. But the color of the roof is too light. Spectral distortion of GSA and CNMF is clearly visible (Fig. 2(c), (e)) as colors do not match the reference HS image. The result of the CNMF method is slightly sharp. Figure 2(d) shows that the GFPCA method has a worst performance in spatial aspects, since the spatial information is added not sufficient and generates lots of blurs. The SFIM method (Fig. 2(f)) and the proposed method (Fig. 2(g)) produce images whose colors better match reference HS image. The proposed method produces an image whose objects are clearer than those of the others, particularly for the scene containing large roof areas.

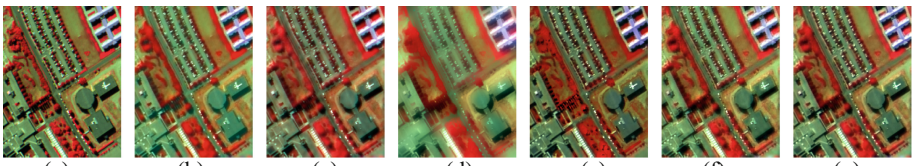


Fig. 2. Visual comparison of different methods. (a) Reference HS image. (b) GS. (c) GSA. (d) GFPCA. (e) CNMF. (f) SFIM. (g) Proposed.

Table 1 shows the objective quantitative assessment of different fusion methods. The proposed algorithm yields the best fusion performance. The CC, RMSE, and ERGAS indexes rank first. These objective indicators prove once again that the presented algorithm provides a better result than compared six approaches in spectral aspect and spatial aspect.

Table 1. Quality metrics for Fig. 2.

Indexes	GS	GSA	GFPCA	CNMF	SFIM	Proposed
CC	0.9062	0.9146	0.7977	0.8743	0.9207	**0.9270**
SAM	7.0372	7.2163	9.1437	8.3256	**6.7222**	6.7760
RMSE	354.7219	317.5397	494.6062	381.7975	309.6550	**300.0954**
ERGAS	4.8246	4.2080	7.0291	5.0050	4.0244	**3.7859**

4 Conclusions

A novel multistage guided filter based hyperspectral pansharpening algorithm is presented. A widely used improved adaptive IHS (IAIHS) method is utilized to calculate the intensity component (INT) which represents the approximate detail layer of HS image. Then, multistage guided filter scheme is utilized for getting the spatial information by using PAN image and INT image as a guidance image respectively, which is effective and simple. Experiments performed on the ROSIS dataset and the proposed hyperspectral image fusion method is compared with six excellent image fusion algorithms. The quantitative measures and displayed results clearly outlines that the presented approach is a more outstanding approach in improving the spatial performance of HS image while maintaining its spectral resolution.

Acknowledgments. This work was supported by NSFC (No. 61372069), National Defense Pre-research Foundation, SRF for ROCS, SEM (JY0600090102), “111” project (No. B08038) and the Fundamental Research Funds for the Central Universities.

References

1. Tu, T.M., Su, S.C., Shyu, H.C., Huang, P.S.: A new look at IHS-like image fusion methods. Inf. Fusion **2**, 177–186 (2001)
2. Shah, V.P., Younan, N., King, R.L.: An efficient pan-sharpening method via combined adaptive PCA approach and contourlets. IEEE Trans. Geosci. Remote Sens. **56**, 1323–1335 (2008)
3. Aiazzi, B., Baronti, S., Selva, M.: Improving component substitution pansharpening through multivariate regression of MS+Pan data. IEEE Trans. Geosci. Remote Sens. **45**, 3230–3239 (2007)
4. Liu, J.G.: Smoothing filter based intensity modulation: a spectral preserve image fusion technique for improving spatial details. Int. J. Remote Sens. **21**, 3461–3472 (2000)

5. Yokoya, N., Yairi, T., Iwasaki, A.: Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. IEEE Trans. Geosci. Remote Sens. **50**, 528–537 (2012)
6. He, K., Sun, J., Tang, X.: Guided image filtering. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1397–1409 (2013)
7. Rahmani, S., Strait, M., Merkurjev, D., Moeller, M., Wittman, T.: An adaptive IHS Pan-sharpening method. IEEE Geosci. Remote Sens. Lett. **7**, 746–750 (2010)



Adaptive Long-Term Object Tracking for Real-Time Visual Search in Multimedia Surveillance

Lei Li, Song Xiao(✉), and Feiyang Tan

State Key Laboratory of Integrated Service Networks,
Xidian University, No. 2, South Taibai Street, Xi'an 710071, China
xiaosong@mail.xidian.edu.cn

Abstract. Mobile visual search attracts an increasing attention in multimedia communication recent years which demands higher performance with less delay. An efficient moving object tracking algorithm based on the correlation filter framework is proposed in this paper aiming at accrute real-time mobile visual search in multimedia surveillance. We employ complementary features including HOG and color attributes for the target appearance representation, and the optimization to solve for both the kernelized correlation filter and the adaptive target response jointly is adopted to counter fast motion, motion blur and occlusion. Besides, a separate scale pyramid filter is applied for accurate scale estimation while being computationally efficient. Moreover, a compressed support vector machine (CSVM) serving as a modifier is employed with a compressed sensing matrix to further boost performance. Finally, extensive experimental results on large-scale benchmark datasets validate that the proposed algorithm outperforms state-of-the-art methods in terms of efficiency, accuracy, and robustness.

Keywords: Video object tracking · Correlation filter
Compressed sensing · Mobile visual search

1 Introduction

Mobile visual search is one of the challenging applications of Multimedia communication, which is exploited in several applications such as military visual guidance, robot vision navigation, security monitoring, human-machine interaction, etc. With the marvelously huge number of visual resources as well as the rapidly developing hand-held terminals, mobile visual search has become pervasive and maturely developed over the past decade. Video object tracking algorithms have important practical value in the field of mobile visual search, which has become a topic of increasing interest in this field [1,2]. Although much progress has been made in recent years [3], it is still a challenging task due to complicated interfering factors in the multimedia videos like partial and full

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 388–397, 2018.
https://doi.org/10.1007/978-3-319-78130-3_40

occlusion, shape deformation, large scale variations, fast motion, and background clutters, to name a few.

Recent years, correlation filters are introduced into visual tracking for their attractive performance both in accuracy and efficiency. Despite the speed and performance advantages of these algorithms, there still exists several drawbacks in the traditional correlation filter tracking framework. For instance, these methods generate circularly shifted image patches to approximate actual translations, which become unreliable in many realistic tracking scenarios. Besides, the selection of the target response in the traditional correlation filter based trackers is intimately related to the assumed motion model. While this model is strict and prone to drift, some trackers usually make it difficult to locate the object precisely.

In this paper, we propose an adaptive long-term object tracking method for real-time visual search in multimedia surveillance. We reformulate the optimization to solve for both the filter and target response jointly with multi-feature representation to tackle the problem of fast motion, motion blur and partial occlusion. Besides, a scale pyramid filter is adopted to counter variety of the target size during tracking while computationally efficient. In addition, a discriminative classifier called CSVM is introduced to conduct the long-term processing on the fly. The proposed algorithm is finally evaluated on popular OTB100 benchmark [4] datasets.

The paper is organized as follows: First, the overall flow chart of the proposed algorithm is given in Sect. 2. Then we follow up with detailed description of all components in the algorithm in Sect. 3. Next, experiments on popular visual tracking benchmark are performed in Sect. 4, and finally we conclude in Sect. 5.

2 Proposed Algorithm Architecture

In this section, we present the framework of the proposed algorithm which can be overviewed in Fig. 1. We decompose the object tracking task in multimedia videos into translation, scale estimation, and re-detection of the target. The upper half of Fig. 1 represents the initialization of learning adaptive kernelized filter tracking model (denoted as KCF_{AT}) and CSVM re-detection model. A KCF_{AT} tracker based on HOG and color attributes features is adopted to account for the target representation and motion. Meanwhile, we collect training samples in an image search area with large responses. A CSVM classifier is then constructed by facilitating efficient random projection from the image feature space to a low-dimensional compressed subspace. In the lower part of diagram, when a new frame comes, we use the trained KCF_{AT} model to locate the target center, and a separately scale pyramid correlation filter is applied at the target center to estimate suitable target scale. When the tracking result of KCF_{AT} is judged as non-target by CSVM model, the target center will be renewed by re-detection mechanism. Finally, features in the new target are extracted to update KCF_{AT} and CSVM model individually.

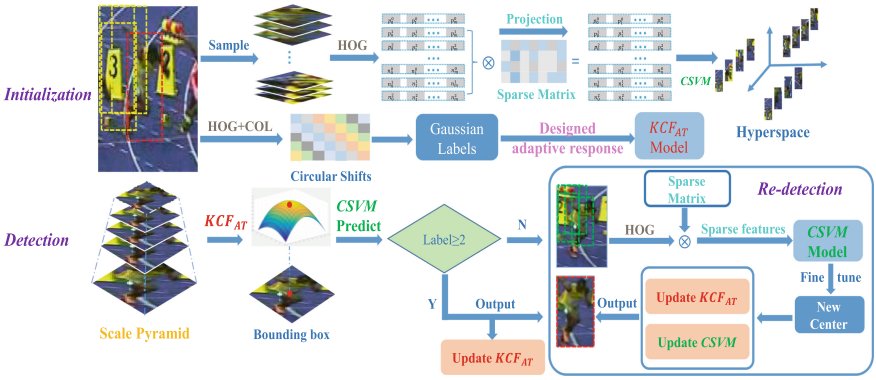


Fig. 1. Framework of the proposed tracking algorithm.

3 Main Components

3.1 Response-Adaptive KCF$_{AT}$ Tracker

KCF tracker [5] proposed by Henriques et al. trained the filter w using an image patch x of size $W \times H$. Relying on the cyclic property, all the circular shifts of $\mathbf{x}_{w,h}, (w,h) \in \{0,\ldots,W-1\} \times \{0,\ldots,H-1\}$ are generated as training samples with Gaussian function labels $y(w,h)$. The goal of training is to find a function $f(z) = \mathbf{w}^T z$ that minimizes the squared error over samples $\mathbf{x}_{w,h}$ and their regression targets $y(w,h)$,

$$w = \underset{w}{argmin} \sum_{w,h} |\varphi\left(x_{w,h}\right) \cdot w - y\left(w,h\right)|^2 + \lambda ||w||^2. \tag{1}$$

Where φ represents the mapping to the Hilbert space, and λ is a parameter for the regularization term. Motivated by [6], we reformulate the peak values of the response map $\mathbf{y}$ in (1) based on not only its original appearance model $\widetilde{\mathbf{x}}$ but also a prior motion model $\mathbf{y_0}$ to build a joint optimization problem:

$$\underset{w,y}{minimize}\ ||\widetilde{\mathbf{x}}w - y||_2^2 + \lambda_1 ||w||_2^2 + \lambda_2 ||y - y_0||_2^2. \tag{2}$$

As shown in Fig. 2, at first a filter model w is learned by solving Eq. 1, and a fixed number of translations are sampled using a standard Kalman filter to learn the prior $\mathbf{y_0}$ in the next frame. Using the fast Fourier transformation to compute the correlation, the object function can be minimized as $\mathbf{w} = \sum_{w,h} \alpha(w,h) \emptyset(\mathbf{x}_{w,h})$ and α contains all the $\alpha(w,h)$ coefficients which is defined by

$$\alpha = \frac{\left(\frac{\lambda_2}{\lambda_1}\widetilde{\mathbf{x}} \odot \widetilde{\mathbf{x}}^* + \lambda_2\right) \odot y_0}{\frac{\lambda_2}{\lambda_1{}^2}\left(\widetilde{\mathbf{x}} \odot \widetilde{\mathbf{x}}^* \odot \widetilde{\mathbf{x}} \odot \widetilde{\mathbf{x}}^*\right) + \frac{1+2\lambda_2}{\lambda_1}\left(\widetilde{\mathbf{x}} \odot \widetilde{\mathbf{x}}^*\right) + (1+\lambda_2)}. \tag{3}$$

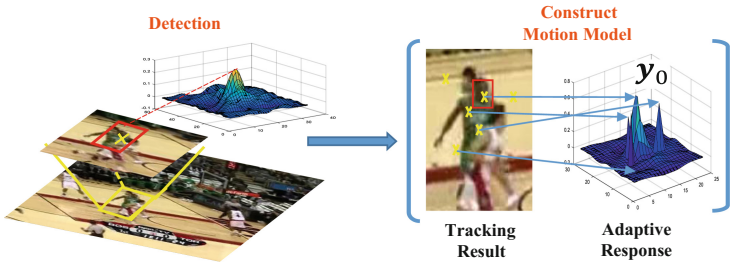


Fig. 2. Construction of a prior for the target response.

Where $*$ donates the complex conjugate. During tracking, a patch z with the same size of x is cropped out in the new frame. The response map in the detection step is then computed as:

$$y = \frac{1}{\lambda_1}\mathbf{z}^* \odot \widetilde{\mathbf{x}} \odot \alpha. \tag{4}$$

3.2 Fast Scale Pyramid

For scale estimation, we extract image blocks of different sizes to learn a separate scale filter. Let $M \times N$ be the target size in the current frame and S be the size of the scale filter ($\mathrm{S} = \{a^n | n = [-(S-1)/2], \ldots, [(S-1)/2]\}$). For each $s \in S$, we extract a candidate image path J_n of size $sM \times sN$ centered around the target position in the previous frame. Unlike [7], affine transform technique is then adopted to resize all patches back to the original size of the target. Finally, the optimal scale $\mathbf{s}$ can be obtained by

$$s = \underset{s}{argmax}\left(\max\left(y_s\right)\right). \tag{5}$$

Here y_s donates the correlation response of the learned appearance model to J_n.

3.3 Compressed Support Vector Machine

To alleviate unrecoverable drift caused by the mislocalized target position, a discriminative classifier called CSVM is built to rectify the target position serving as a modifier. Assume we are given a set of labeled samples $\{(x_i, y_i)\}$, our aim is to learn a model which formulated as follows:

$$min\frac{1}{2}\left(w^T w\right) + \lambda\sum_i \xi_i$$

$$\text{Subject to: } y_i\left(W^T\phi\left(x_i\right) + b\right) \geq 1 - \xi_i, \xi_i \geq 0. \tag{6}$$

Where x_i is a sample feature, y_i represents its corresponding label, and ϕ donates the kernel function.

Training of CSVM: In the initial frame, the training set $\{(x_i, y_i) \,|i = 1, \ldots, n\}$ is generated by scanning window that scans in a target search area. Their labels are determined by the ratio of overlap area with the target. These samples which are labeled with 1, 2, and 3 donate background, the target that maybe occluded or suffered from deformation, and the target of interest respectively. After preprocessing, we compress these samples using a random Gaussian matrix to project data from the high-dimensional image space $x \in R^m$ to a lower-dimensional subspace $v \in R^n$. This matrix $R \in R^{n \times m}$ where $r_{ij} \sim N(0, 1)$ is defined as:

$$r_{ij} = \sqrt{p} \times \begin{cases} 0 \rightarrow probability\left(1 - \frac{1}{p}\right) \\ 1, -1 \rightarrow probability \frac{1}{2p}, \frac{1}{2p} respectively. \end{cases} \tag{7}$$

This matrix is very sparse and satisfies the restricted isometry property [8,9], thus keeping computational load of constructing the classification hyperplane at a low level while preserving almost all the information of the original feature. During training, cross validation and grid search are employed to find the best parameters of CSVM.

Prediction of the target position: After the prediction of the target position and scale, we extract its feature and predict the corresponding label according to the trained CSVM model. If the label equals to 2 or 3, it indicates the predicted center is the target center most likely. However, while labeled with 1, we try to search some better candidate image patches that may contain the actual target of interest. As the diagram 3 illustrates, suppose that features $\delta = \{f_u^v | f_u \in R^m\}$ are extracted from v image patches, and their corresponding weight $w_t \, (t = 1, \ldots, v)$ is proportional to the center distance between the candidate image patch and the target center predicted in previous frame. Compressed by the same compressed sensing matrix, their confidence level of identifying as the target are donated as:

$$L = \left\{ l_v = W_v^u \cdot \phi \left(R^{n \times m} * f_u^v \right) | f_u^v \in \delta, m \gg n \right\}. \tag{8}$$

We formula $Flag = \sum (l_t \geq 2 | t = 1, \ldots, v)$ that indicates the number of the predicted labels which equal to 2 or 3. When this $Flag$ exceeds a threshold which is set empirically, their corresponding features will be added into the training set to update CSVM parameters. Finally, the maximum confidence c_t from the

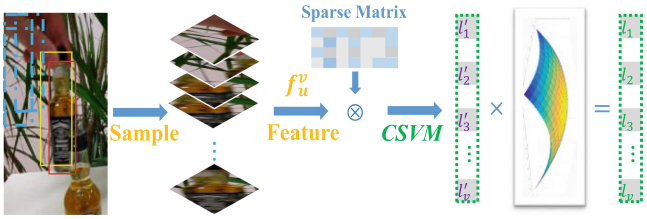


Fig. 3. Prediction phase of our CSVM.

best candidate patch will return its coordinate p_t as the new target center. c_t is formulated as follows:

$$c_t = \underset{t}{argmax}(l_t * w_t | t = 1, \ldots, v). \tag{9}$$

4 Experimental Evaluation

4.1 Experimental Setup and Methodology

In the proposed algorithm, λ_1 and λ_2 in KCF_{AT} are set to 10^{-3} and 10^{-5} respectively, and 7 translations are used to form the target response y_0. The learning rate β is set to 0.01, and a 33-level scale filter is adopted for scale estimation. Besides, we set $p = m/4$ in compress matrix, and n = 50 for lower-dimensional space. We use the same parameter values for all the sequences. All the experiments are implemented on the same platform in MATLAB on an Intel(R) Core(TM) i7-4790 CPU (3.60 GHz) PC with 8 GB memory.

4.2 Quantitative Evaluation

The proposed algorithm is evaluated on OTB100 benchmark [4] with state-of-the-art trackers including LCT [10], KCF [5], MEEM [11], STC [12], and MUSTer [13], etc. We randomly selected 70 video sequences to carry out experiments. Figure 4 illustrates that our approach performs well against the existing state-of-the-art algorithms, and always outperforms some famous trackers such as KCF [5] and TLD [14] (which have been put into practical application) by a large margin. Besides, a more detailed algorithmic comparison chart based on the performance over six tracking challenging attributes is presented in Fig. 5, which indicates the proposed algorithm achieves outstanding performance under these complex video tracking conditions.

4.3 Qualitative Evaluation

In order to further highlight the superior performance of the algorithm, we intuitively compare our algorithm with three state-of-the-art methods including MEEM [11], STC [12] and KCF [5] by contrasting tracking results on challenging sequences in Fig. 6.

Experimental results: As illustrated in Fig. 6, the MEEM tracker can mitigate tracking drift problem (Basketball and Bolt), but it is susceptible to changes in illumination (Coke) and can't account for scale variation (CarScale). The STC tracker is able to estimate the target scale changes (Basketball and CarScale), but it does not perform well when camera perspective changes (BlurBody) or in the presence of fast motion (Deer). The KCF tracker performs well in handling significant illumination change and deformation (Coke). However, it easily drifts when the target experiences fast motion as well as motion blur (Deer) and does

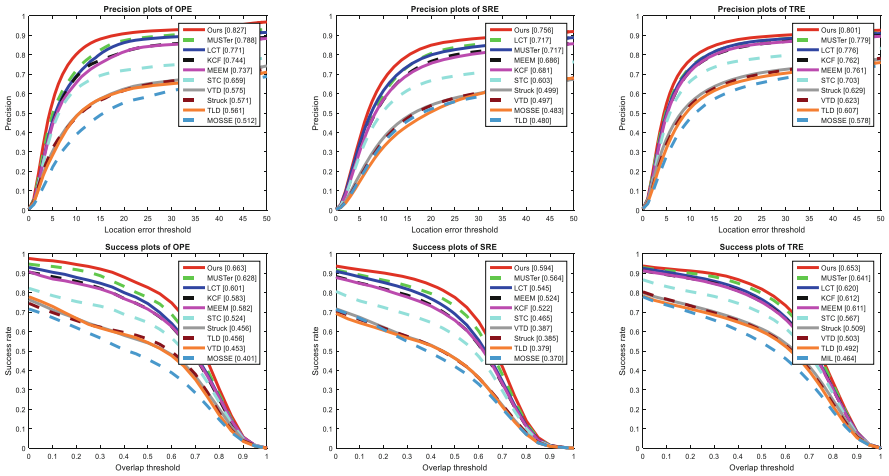


Fig. 4. Distance precision and overlap success plots over 100 benchmark sequences using one-pass evaluation (OPE), temporal robustness evaluation (TRE) and spatial robustness evaluation (SRE). The legend contains the area-under-the-curve score for each track.

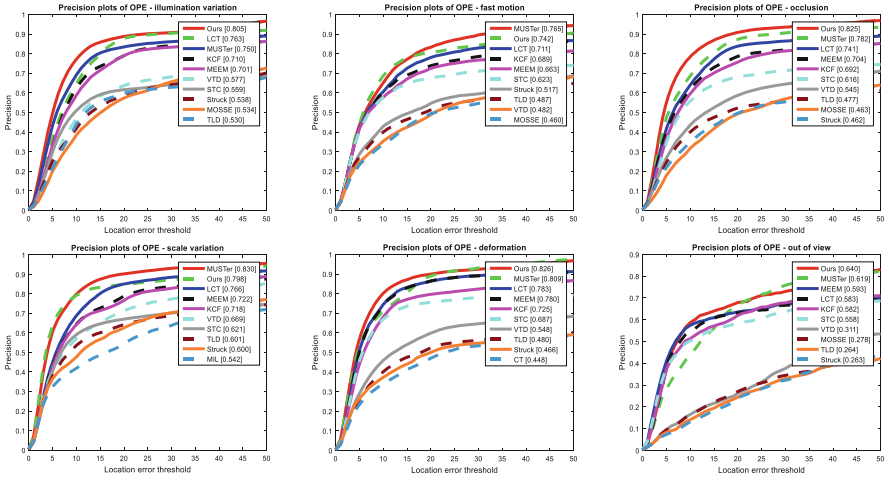


Fig. 5. Distance precision plots over six tracking challenges.

not re-detect targets in the case of tracking failure (BlurBody). Overall, our algorithm is superior to these algorithms in terms of illumination variation, fast motion, motion blur, heavy occlusion, scale variation, deformation, and out of view.

Performance analysis of our algorithm: Favorable improvement has been achieved by the proposed algorithm, which mainly attributed to four reasons.

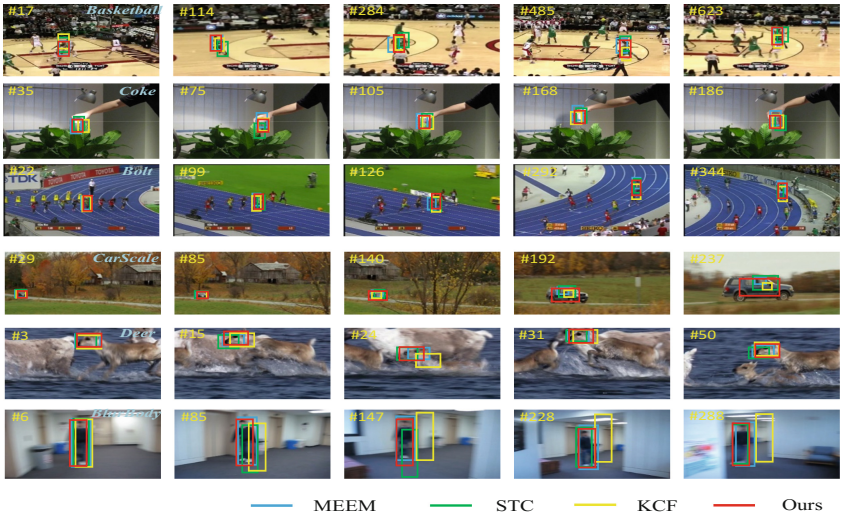


Fig. 6. The bounding box of four tracking algorithms.

First, we use HOG and color-naming features to construct more discriminative target representation model which are less sensitive to deformation and partial occlusion (*Basketball* and *Coke*). Second, we construct a response-adaptive motion model by means of using actual translation measurements of correlation, which makes the detection step of the algorithm more reliable in fast motion, motion blur and partial occlusion (*Deer* and *BlurBody*). Third, a fast scale pyramid filter model is introduced to adapt to the scale change of the target (*CarScale*), thus avoiding adding unnecessary background information into model update. Finally, we train our CSVM model to rectify undesirable tracking results, keep our tracker away from unrecoverable drift, and re-detect the target after it runs out of view.

It's worth noting that hierarchical features learned from deep convolutional neural networks have been shown to be effective for many computer vision tasks such as image classification and recognition in recent years. Wang and Yeung [15] firstly proposed an autoencoder network to learn representative features for generic object tracking task. Since then, a variety of video target tracking algorithms based on deep learning methods have been proposed, and achieved good performance advantages. Considering the tracking speed and adaptability, this paper focuses on the study of traditional representation-model based target tracking algorithm. We recommend applying our proposal to realistic mobile visual search task. More details about recent object tracking methods based on deep learning can be referred to [16].

4.4 Speed Evaluation

A good tracker should not only have high tracking accuracy, but also achieve real-time performance and can be put into practical application. We evaluate the speed of our algorithm by average time consumption comparing with some algorithms on benchmark datasets. Figure 7 illustrates our tracker can reach 30.6 frames per sec (an average speed which tested in [4] with 100 video sequences), so real-time performance can be achieved under various complex conditions in the multimedia video object tracking task.

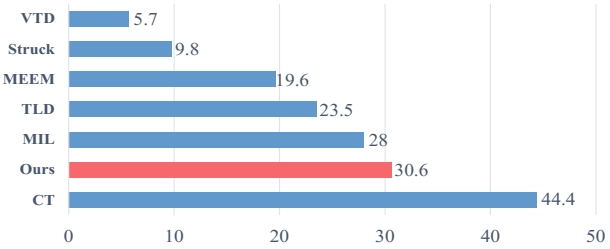


Fig. 7. Processed frames/sec on several popular trackers.

5 Conclusion

In this paper, we present a real-time correlation filter based video tracking method, which can perform long-term tracking of targets under complex conditions in multimedia videos. We introduce a target adaptive-response motion model as well as an accurate scale estimation scheme to improve tracking precision while computationally efficient. Besides, a target re-detection method called CSVM is proposed to further enhance the tracking performance by re-detecting the target when necessary. Numerous experiments with popular algorithms on challenging datasets demonstrate that the proposed tracking method performs well in terms of accuracy, robustness, and speed. In the future, we will focus on the application of this algorithm in the field of multimedia surveillance, especially in mobile visual search.

Acknowledgments. This work was supported by NSFC (No. 61372069), "111" project (No. B08038) and the Fundamental Research Funds for the Central Universities.

References

1. Fayed, S., Youssef, S.M., El-Helw, A., et al.: Adaptive compressive sensing for target tracking within wireless visual sensor networks-based surveillance applications. J. Multimed. Tools Appl. **75**(11), 6347–6371 (2016)
2. Nandhini, S.A., Radha, S.: Compressed sensing based object detection and tracking system using measurement selection process for wireless visual sensor networks. In: IEEE International Conference on Wireless Communications, Signal Processing and Networking, pp. 1117–1122. IEEE (2016)
3. Smeulders, A.W., Chu, D.M., Cucchiara, R., et al.: Visual tracking: an experimental survey. J. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1442–1468 (2013)
4. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. J. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1834–1848 (2015)
5. Henriques, J.F., Caseiro, R., Martins, P., et al.: High-speed tracking with kernelized correlation filters. J. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 583–596 (2014)
6. Bibi, A., Mueller, M., Ghanem, B.: Target response adaptation for correlation filter tracking. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 419–433. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_25
7. Danelljan, M., Häger, G., Khan, F.S., et al.: Accurate scale estimation for robust visual tracking. In: British Machine Vision Conference, pp. 65.1–65.11 (2014)
8. Achlioptas, D.: Database-friendly random projections: Johnson-Lindenstrauss with binary coins. J. Comput. Syst. Sci. **66**(4), 671–687 (2003)
9. Zhang, K., Zhang, L., Yang, M.H.: Real-time compressive tracking. In: IEEE European Conference on Computer Vision, pp. 864–877. IEEE (2012)
10. Ma, C., Yang, X., Zhang, C., et al.: Long-term correlation tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5388–5396. IEEE (2015)
11. Zhang, J., Ma, S., Sclaroff, S.: MEEM: robust tracking via multiple experts using entropy minimization. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 188–203. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_13
12. Zhang, K., Zhang, L., Liu, Q., Zhang, D., Yang, M.H.: Fast visual tracking via dense spatio-temporal context learning. In: IEEE Proceedings of European Conference on Computer Vision, pp. 127–141. IEEE (2014)
13. Hong, Z., Chen, Z., et al.: MUlti-Store Tracker (MUSTer): a cognitive psychology inspired approach to object tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 749–758. IEEE (2015)
14. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. J. IEEE Trans. Pattern Anal. Mach. Intell. **34**(7), 1409–1422 (2012)
15. Wang, N., Yeung, D.Y.: Learning a deep compact image representation for visual tracking. J. Adv. Neural Inf. Process. Syst. **26**, 809–817 (2013)
16. Hu, D., Zhou, X., Yu, X., et al.: Study on deep learning and its application in visual tracking. In: International Conference on Broadband and Wireless Computing, Communication and Applications, pp. 240–246. IEEE (2015)



Multi-objective Optimization and Design of LEO Satellite Broadband Networks

Dawei Yan[1(✉)], Cong Liu[2], Peng You[1], and Shaowei Yong[1]

[1] Institute of Electronic Science and Engineering,
National University of Defense Technology, Changsha 410073, Hunan, China
yandawei-1986@163.com
[2] College of Mathematics and Information Science,
Baoji University of Arts and Sciences, Baoji 721013, Shaanxi, China

Abstract. In this paper, a multi-objective optimization design methodology of low Earth orbit (LEO) satellite broadband network is proposed. Firstly, the quality of service (QoS) metrics and network stability factor (NFS) are established for the LEO satellite network. Then, a multi-objective optimization model for LEO satellite network design is constructed based on the constraints of permanent Inter-satellite Link (ISL) and QoS metrics. In addition, a constrained multi-objective optimization algorithm is designed by integrating NSGA-II with constraints handling. Finally, the simulated results validate the effectiveness of the proposed method and show that this approach provides a new idea for the optimization deployment of the network architecture.

Keywords: LEO satellite broadband network · Quality of service
Network stability · Cost-efficient · NSGA-II

1 Introduction

Satellite networks are becoming an attractive approach for worldwide wireless communication services. They not only provide global coverage, but they are also capable of consistently sustaining high bandwidth level [1]. Geostationary Earth orbit (GEO) satellites suffer from high propagation delay, which is not suitable for most applications. Therefore, the main emphasis of research has shifted to development of non-geostationary (NGEO) networks, especially LEO satellite networks, due to unique capabilities in terms of extensive geographic coverage, low end-to-end delay, and high-date rate Internet-based applications and so on [2, 3].

As in LEO satellite networks, countless trades exist between system performance, system cost, and each of the design parameters. Thus, a method is needed to enable a greater search of the tradespace and explore design options that might not otherwise be considered during the conceptual design phase [4]. Suzuki et al. proposed a mobile satellite communication system with stable ISL [5]. However, due to the limits of its assumption, the system could not guarantee QoS. Jilla and Miller presented a multi-objective, multidisciplinary design optimization methodology for distributed satellite system based on QoS metrics [4]. But it lacked the consideration of establishing ISL. Although the LEO satellite network design methodology can be found in

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 398–408, 2018.
https://doi.org/10.1007/978-3-319-78130-3_41

many literatures, there are only a handful of studies that concentrate on both the QoS and satellite networking. A systematic study on the design of LEO satellite broadband network is still inadequate. Thus, this problem is still an open-ended issue.

In this article, the research develops and applies a network cost-efficient and stability for LEO satellite networks optimization design with the QoS and permanent ISL. Moreover, a new constrained multi-objective evolutionary algorithm based on integrating NSGA-II [6] with constraints handling is tailored to search the nonlinear and non-convex tradespace of the LEO satellite broadband network.

The rest of this paper is organized as follows. The QoS metrics and network stability factor are defined in Sect. 2. Section 3 gives the LEO satellite broadband network composition. In Sect. 4, the optimization model and solution algorithm is proposed. Section 5 shows the simulation results and analysis. Conclusions of the paper are drawn in Sect. 6.

2 Equivalent Reformulations

2.1 QoS Metrics

The QoS is essentially originated from the contradictions between the quality and the number of services in communication systems. However, it has not been defined explicitly yet. We present the QoS metrics of the LEO satellite broadband network as the effectiveness and fairness based on availability of service of the satellite network as shown in Table 1.

Table 1. The QoS metrics for LEO satellite broadband network.

Capability	Equivalent design parameter
Effectiveness	Signal-to-noise ratio
	Information rate
	Bit error rate
Fairness	Coverage percentage
	Degree of network connectivity

Effectiveness which reflects service quality can be described with respect to the following factors:

Signal-to-noise ratio: which is the ability of a system to isolate and identify the information signals from different interfering signals or noise source.
Information rate: which is used to represent the minimum amount of information for an individual user within the LEO satellite broadband network.
Bit error rate: BER refers to the quality of the information being transmitted through the LEO satellite broadband network.

Fairness based on user requirement and the system state is expressed in two aspects:

Network coverage: the percentage of service coverage that the network can provide broadband communication service for users in the satellite network. For LEO satellite broadband networks, this is quantified by coverage percentage.
Network connectivity: which characterizes the probability that communication links are established between different users in the network. For the broadband mission, this is measured by the degree of network connectivity.

2.2 Satellite Network Stability Model

There are four ISLs for each satellite in the LEO satellite network. In the case of intra-orbit ISL, the relative location between the satellites on the same orbit is almost unchanged. On the contrary, the distance, azimuth and elevation angles of inter-orbit ISL are dynamic. The distance and azimuth angle variation of the inter-orbit ISLs determine the stability of the LEO satellite broadband network directly. As a consequence, satellite network stability factor (Ψ) is defined as the linear combination of distance rate and angular rate of the ISL. Ψ is expressed as

$$\Psi = \alpha \dot{\Delta d} + \beta \dot{\Delta \psi} \tag{1}$$

where α and β are the coefficient satisfying the equation $\alpha + \beta = 1$. It can be adjusted according to the difficulty of establishing the ISLs. It is notice that the network stability factor is smaller, the satellite network is more stable.

Distance rate of ISL $\dot{\Delta d}$ is defined as [7]

$$\dot{\Delta d} \triangleq E\left(\left|\frac{\partial \Delta d}{\partial t}\right|\right) \tag{2}$$

where Δd is the relative distance of the ISL, $E(\bullet)$ denotes averaging method.

Similarly, angular rate of ISL $\dot{\Delta \psi}$ is defined as [7]

$$\dot{\Delta \psi} \triangleq E\left(\left|\frac{\partial \psi_{AB}}{\partial t}\right|\right) \tag{3}$$

where ψ_{AB} is the azimuth angle of the ISL.

3 LEO Satellite Broadband Network Model

3.1 Constellation Model

Network Coverage. Walker Constellations [8] are the homogeneous symmetric type of satellite constellations with prominent advantages of the global and latitudinal coverage. Coverage percentage (PC) is used to evaluate coverage performance of the satellite constellation. The numerical simulation method for PC can be defined as:

$$\eta_{Coverage} = \frac{100}{N_{tol}} \sum_{i=1}^{N_{tol}} L\,\mathrm{cov}_i \tag{4}$$

In Eq. (4), $L\,\mathrm{cov}_i \in \{0, 1\}$, where any grid covered by one or more satellites during simulation time is expressed as $L\,\mathrm{cov}_i = 1$, otherwise, $L\,\mathrm{cov}_i = 0$. N_{tol} denotes the total number of simulation steps.

Network Connectivity. Establishing the permanent ISL between satellites in the adjacent orbit could eliminate impacts of topological dynamics of the constellation. The condition of permanent ISL between two satellites is given as [9].

$$\frac{r}{R_{eq}} \geq \sqrt{1 + \frac{(1 - \varepsilon_{\min})^2}{1 - \varepsilon_{\min}^2}} \tag{5}$$

where, r denotes orbital radius of the satellites, R_{eq} is the equivalent Earth radius by considering the effects of topography and atmospheric, $\varepsilon_{\min}$ is the minimum x′-axis of sat2 coordinate.

Based on the condition of permanent ISL, degree of network connectivity (DONC) is defined as:

$$\eta_{DONC} = \frac{100}{N_{isl} L} \sum_{i=1}^{N_s - 1} \sum_{j=i+1}^{N_s} \sum_{k=1}^{L} v_{isl}^{i,j,k} \tag{6}$$

where the ISL connectivity $v_{isl}^{i,j,k} \in \{0, 1\}$ is evaluated between the ith and jth satellite at the kth sample instant. The number of ISLs is $N_{isl} = 2N_s$ for a constellation of N_s satellites.

3.2 Link Budget Module

It is assumed that uncoded BPSK signals and coherent demodulation for a satellite network, bit error rate (BER) of coherent reception can be calculated by

$$p_e = \frac{1}{2}\left(1 - erf\sqrt{E_b/N_0}\right) \tag{7}$$

Obviously, the corresponding SNR threshold E_b/N_0 can be obtained from Eq. (7) for a specified BER in the LEO satellite broadband network.

The total downlink data rata R of the satellite to the grid can be calculated based on the link budget equation as:

$$R = \frac{P_{Sat} G_{Sat} G_r L_D L_{RM}}{SNRkT} \tag{8}$$

where P_{Sat} denotes the transmitted power of the satellite, G_{Sat} denotes the gain of satellite transmitting antenna, G_r is gain of the user antenna, L_{RM} is link margin for the rain loss, T is the user terminal system noise temperature, k denotes the Boltzmann's constant, SNR is the required signal-to noise ratio according to Eq. (7), and L_D is the propagation loss of down link.

Then the gain G_{Sat} of the satellite antenna is

$$G_{Sat} = \eta_{Sat} \frac{4\pi A_{Sat} f_c^2}{c^2} \tag{9}$$

where A_{Sat} is the equivalent area of the satellite downlink antenna, η_{Sat} denotes the antenna efficiency, and f_c is the downlink transmission frequency.

Based on Eqs. (8) and (9), the satellite capacity C_{sat} can be defined as

$$C_{sat} = \frac{R\eta_{MAE}}{R_{user}} \tag{10}$$

where η_{MAE} is the multi-access scheme efficiency, and R_{user} denotes the individual user information rate (i.e. T1 link, which is defined as the information rate of at least of 1.544 Mbps in broadband communication by the Federal Communication Commission).

Finally, the maximum capacity of LEO satellite broadband network is the total number of the users who can satisfied by T1 link.

4 Optimization Model and Algorithm

4.1 Optimization Variables

The optimization variables contain seven parameters as given in Table 2. These optimization variables are the key independent design parameters which can be controlled by multi-objective optimization algorithm. Note that the design space may include thousands of different LEO satellite networks, thus it is necessary to explore optimal solution of constellation design.

Table 2. Optimization variables of LEO satellite broadband network.

Design variable	Symbol	Uints
Number of orbital planes	P	–
Number of satellite per orbital planes	S	–
Phase factor	F	–
Orbit height	h	km
Inclination	i	deg
Equivalent area of satellite downlink antenna	A_{Sat}	m^2
Power of payload	P_{Sat}	w

4.2 Objective Functions and Constraints

The aim of this paper is to achieve the network architecture with optimal cost-efficient ratio and strong stability under the constraints that the QoS metrics and permanent visibility are satisfied. The objective functions and constraints of the network optimization design can be mathematically defined as

$$\begin{aligned} &\text{Objective:} && Max\ \Phi = \frac{N_{grid} C_{Sat}}{N_{Sat} A_{Sat} P_{Sat}} \\ &&& Min\ \Psi = \alpha \Delta d + \beta \psi \\ &\text{Constraints :} && \text{Subject to} \end{aligned}$$

$$\begin{cases} SNR \geq SNR_0 \\ BER \leq BER_0 \\ R \geq R_0 \\ \eta_{Coverage} = 100\% \\ \eta_{DoNC} = 100\% \\ \frac{r}{R_{eq}} \geq \sqrt{1 + \frac{(1-\varepsilon_{\min})^2}{1-\varepsilon_{\min}^2}} \end{cases} \tag{11}$$

where Φ is network capacity per cost which is defined as the overall capacity of the LEO satellite broadband network divided by potential network cost including the number of satellites, the equivalent area of the antenna, the total power of satellite, etc. Ψ is satellite network stability factor, which is smaller, the satellite network is more stable. (SNR_0, BER_0, R_0, η_{DoNC}, $\eta_{Coverage}$) is element of QoS metrics and h/R is the condition of permanent visibility. For any user in the network, the signal to noise ratio $SNR \geq SNR_0$, bit error rate $BER \leq BER_0$, and information rate $R \geq R_0$ should be met. And for the overall network, coverage percentage and degree of network connectivity should meet requirements (i.e. $\eta_{Coverage} = 100\%$ and $\eta_{DoNC} = 100\%$).

4.3 Constrained Multi-objective Optimization Algorithm

As shown in Eq. (11), LEO satellite network optimization design has a large search space which is non-convex, nonlinear, and discontinuous. However, EA is essentially an unconstrained optimization algorithm. It is necessary to integrate evolutionary algorithm with the constraint handling approaches to form a constrained evolutionary algorithm. Therefore, a constraint handling method referred to as Infeasible Sorting (IS) is proposed for our specific issues, which means that we need to analyze the infeasible solutions in each generation. We call this: improved NSGA-II based on Infeasible Sorting (NSGA-II-IS).

Infeasible Sorting Operator: All solutions in each generation are sorted in ascending order on the basis of the value of the constraint violation, respectively. Constraint violation value of 0 is allotted to the solutions that do not violate a constraint and an ascending order based on the sorted list is allotted to the other solutions that violate a constraint. In the sorted list solution with least constraint violation value gets rank 1 and the solutions with the same value get the same rank. This sorting process is

repeated for all the constraints. Then constraint violation level for each solution is the sum of the ranks which are obtained for all the constraints.

The procedure of sovling the problem (11) by improved NSGA-II

```
Initialize
  P                                          #population Size
  M                              #number of Optimization Objectives
  N_G>1                                    #number of Generations
  0<α<1                          #proportion of infeasible solution
  i=1
Calculate
  P_inf=α*P                    #number of infeasible solutions in P
  P_f=P-P_inf                    #number of feasible solutions in P
  pop_1=Initialize()                    #Initialize the population
  Evaluate(pop_1)                         #Constraint handling
Repeat
  Child pop_i-1=Evolve(pop_1)#do selection,crossover and mutation
  Evaluate(child pop_i-1)                  #constraint handling
  (S_f,S_inf)=Split(pop_i-1+pop_i-1)      #divide into a feasible
                              set (S_f) and an infeasible set (S_inf)
  F_s=ε-dominated sort[S_f(M)]
  Crowding distance assignment(F_s)
  F_inf=ε-dominated sort[S_inf(M+1)]
  Crowding distance assignment(F_inf)
  pop_i= S_f(1,P_f)+S_inf(1,P_inf)          #generate a new population
  i=i+1
Until i=N_G
```

5 Simulations and Results

Simulations are implemented to confirm the effectiveness of the proposed optimization design methodology. All the computations are done in Matlab and some of the results are analyzed by STK (Satellite Tool Kit). To establish a baseline for analysis and comparison, the interval of the latitude and longitude on the entire Earth's surface is 6° for grid resolution and the simulation period is 24-h. The parameter settings for the NSGA-II-IS are outlined in Table 3.

Table 3. Parameters for NSGA-II-IS

Parameter	Value
Population size	100
Generations	100
Crossover probability	0.9
Mutation probability	0.1
Crossover distribution index	15
Mutation distribution index	20
Proportion of infeasible solutions	0.2

Considering the effect of complex terrain on Earth surface, the minimum elevation angle should be larger than 10° to ensure the quality of satellite communication [10]. Therefore, the approximate Pareto fronts found by NSGA-II-IS for LEO satellite network are illustrated in Fig. 1.

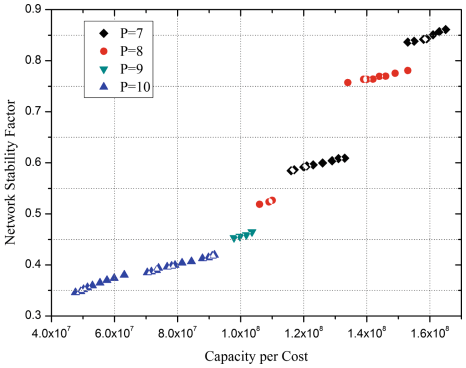


Fig. 1. Tradeoff between network capacity per cost and stability

From Fig. 1, it can be seen that system requirement conflict between network cost-efficient ratio and stability: the improvement of the cost-efficient ratio is always accompanied by the decrease of the network stability (the network stability factor is smaller, the satellite network is more stable) with the number of orbital planes in the constellations increasing, while the less the cost-efficiency ratio of network will be. At the same time, when the number of orbital planes is fixed, the cost-efficiency ratio of the network reduces with the increase of the number of satellites in the constellation.

This series of non-dominated solutions achieves a better tradeoff between the cost-efficiency ratio and stability of the LEO satellite broadband network. It is worth noting that the existence of characteristic 'steep' has been found in the tradeoff curve since the total number of satellites in constellations increases discontinuously in the optimization process.

Base on the Pareto solutions, the constellation parameter set (P = 7, S = 9, F = 1, h = 1125.2 km, and i = 68 deg) can be easily selected for the basic design of the LEO satellite broadband network. This constellation gives the best network capacity per cost. However, these parameters result in the least stability of the network. Pareto solutions given out in this article lets the decision maker to choose the best tradeoff between the network cost-efficient and stability according to the system design requirements.

The LEO network with the best network capacity per cost is taken for example. In the view of the users, the coverage multiplicity of the LEO satellite network is shown in Fig. 2. The coverage multiplicity is defined as the number of satellites in visual range of each user. It is important to determine the availability of the network. The contour in Fig. 2 illustrates that the constellation (63/7/1) could achieve the global seamless coverage. The area covered by 2 satellites takes up more than 90% of the earth, and 84% of the area between north latitude 30°–70° with a dense population can be covered

by 3 satellites. There are also at least 2 satellites in view of users in low latitude area and most of the polar region.

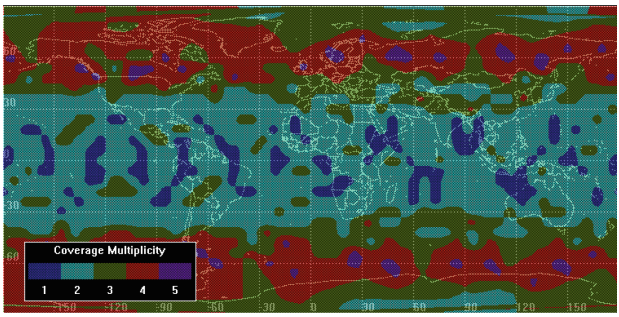


Fig. 2. Coverage multiplicity of the LEO network (7/9/1)

Figure 3 shows the variation in distance, azimuth and elevation of inter-orbit ISL for the constellation parameters (63/7/1) selected for the LEO satellite broadband network. The solid line shows the distance variation of adjacent satellites in adjacent orbital planes. The broken line shows the azimuth variation for the inter-orbit ISL. From Fig. 3, it is obvious that the inter-orbit ISL is permanent during the whole system running with this constellation parameter set.

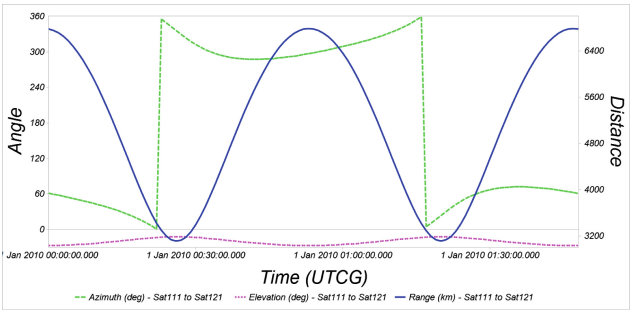


Fig. 3. The distance, azimuth and elevation angles of Inter-Orbit ISL

Figure 4 shows the average elevation angle versus the latitude of the ground station for single-satellite visibility. It is obvious that average elevation is greater than 25° between 70° south latitude and 70° north latitude. It will be better to provide communication services for users at different terrain. Even though in the polar region the average elevation is over 10°.

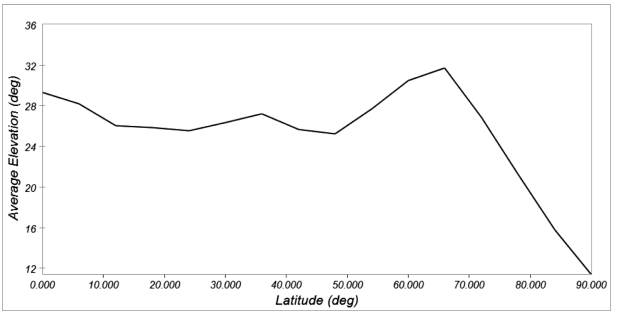


Fig. 4. Average elevation of the ground users at different latitudes

6 Conclusions

In this paper, a multi-objective optimization design method of LEO satellite broadband network is developed based on QoS and the stability of LEO satellite network. Based on in-depth analysis of the satellite network model, the optimization variables of the LEO satellite network design are chosen. In addition, a multi-objective model with constraints on the basis of the network cost-efficient and stability is prospoed. To solve this nonlinear, non-convex, and discontinuous hybrid optimization problem, a constrained multi-objective optimization algorithm is designed based on integrating NSGA-II with constraints handling. For global LEO satellite broadband network, a set of Pareto solutions of constellation design are found to achieve the efficient tradeoff between the network cost-efficient and stability for the decision maker. The proposed optimization design methodology can be used as guidelines for optimization deployment of LEO satellite network.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grants 61601487.

References

1. Ekici, E., Akyildiz, I.F., Bender, M.D.: A multicast routing algorithm for LEO satellite IP networks. IEEE/ACM Trans. Netw. **10**(2), 183–193 (2002)
2. Jamalipour, A.: Broad-band satellite networks-the global IT bridge. Proc. IEEE **1**(89), 88–104 (2001)
3. Taleb, T., Kato, N., Nemoto, Y.: Recent trends in IP/NGEO satellite communication systems: transport, routing, and mobility management. IEEE Wirel. Commun. Mag. **12**(5), 63–69 (2005)
4. Jilla, C.D., Miller, D.W.: Multi-objective, multidisciplinary design optimization methodology for distributed satellite systems. J. Spacecr. Rockets **41**(1), 39–50 (2004)
5. Suzuki, R., Motoyoshi, S., Yasuda, Y.: A study of constellation for LEO satellite network. In: 22nd AIAA International Communications Satellite Systems Conference, Monterey, California, pp. 1–10 (2004)

6. Deb, K., Agrawal, S., Pratap, A., Meyarivan, T.: A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In: Schoenauer, M., Deb, K., Rudolph, G., Yao, X., Lutton, E., Merelo, J.J., Schwefel, H.-P. (eds.) PPSN 2000. LNCS, vol. 1917, pp. 849–858. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-45356-3_83
7. Zhang, J.: Space-Based Mobile Communication Network. National Defense Industry Press, Beijing (2011)
8. Walker, J.G.: Some circular orbit patterns providing continuous whole earth coverage. J. Br. Interplanet. **24**(1), 369–384 (1971)
9. Wang, Z., Wang, P., Gu, X., Guo, Q.: Research on design of permanent inter-satellite-links in satellite networks. J. Commun. China **27**(8), 129–133 (2006)
10. Restrepo, J., Maral, G., Massy, G., Velez, J.: Evaluation of traffic performances in non-geo satellite constellation for mixed services (mobile and fixed). In: 20th AIAA International Communication Satellite Systems Conference, pp. 1–8 (2002)



Routing Protocol Design for LEO Satellite Networks with Uni-directional Links

Juan Wei(✉), Lihong Wang, and Shixuan Feng

School of Communication Engineering, Xidian University, Xi'an 710071, Shaanxi, People's Republic of China
weijuan@xidian.edu.cn

Abstract. Most of the existing routing protocols designed for LEO satellite networks assume that links are bi-directional symmetric links, which is in contradiction with appearance of Uni-directional link. Those protocols may decrease their performance or at worst cannot establish any path under Uni-directional links existence. In this paper, we will present an efficient routing protocol for LEO satellite networks with Uni-directional links. However, using Uni-directional link introduces new problems such as detection of link disconnection. In order to solve these problems, we employ the group strategy and parallel computing. Simulation results show that more packets could be delivered than Datagram Routing Algorithm (DRA) with pruning Uni-directional links. It is also shown that average end-to-end delay can be decreased in some extent by using group strategy and parallel computing.

Keywords: Routing protocol · Uni-directional links · LEO satellite networks Group strategy · Parallel computing

1 Introduction

Most routing protocol designs and their implementations of LEO satellite networks assume that every link is bi-directional and symmetrical [1, 2]. This assumption does not hold true with the emerging of Uni-directional links (UDLs) in satellite networks with different transmission power, bandwidth, memory and external noise source. UDLs may be as ubiquitous as bi-directional links (BDLs) in future satellite networks. Moreover, BDLs in reality are not symmetrical, since the bandwidth available in both directions differs, and it can be advantageous to split a BDL into two UDLs for the Quality of Service routing. As conventional routing protocols cannot be applied directly in the satellite networks with UDLs, alternative routing schemes are needed.

Compared with internet and other networks, satellite networks have some special features such as dynamic topology, high speed movement and so on. So the current routing solutions for these networks with UDLs [3, 4] cannot be used for satellite networks directly. Hu and Liu [5] developed a modified A-DSR routing protocol for satellite networks with UDLs. However, DSR protocol is on-demand routing protocol for Ad hoc network, which mainly apply to low mobility, small amount of data transmission [6]. But for high mobility satellite networks with UDLs, A-DSR degrades the network performance and increases the routing load.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 409–416, 2018.
https://doi.org/10.1007/978-3-319-78130-3_42

In satellite networks, LEO (Low Earth Orbit) satellite networks have much more advantages such as low earth orbit [7], small satellite-earth link delay, simple terminals, spectrum multiplexing, high link bandwidth and so on. As a result, LEO satellite networks are becoming more and more popular in the research of satellite systems.

In this paper, based on group strategy and parallel computing, we will propose an efficient routing protocol for LEO satellite networks with UDLs to enhance satellite networks performance. The simulation experiment will be realized from two aspects. The first one establishes links including UDLs with the new protocol. The second one tends to prune all UDLs from components of links in link state database. Then, links consisting of only BDLs are established. For this case, we choose the typical on-board DRA [8].

DRA generates shortest distance paths between source and destination by taking advantage of the regular mesh mode topology in LEO satellite networks. Special strategies are used to reduce the network congestions and failures of links. Because the protocol does not need to know the topology information of the whole network, additional overhead caused by the renewal of topology is avoided.

The rest of the paper is structured as follows. In Sect. 2 we describe network model. The UDLs link detection and routing calculation are presented in detail in Sect. 3. In Sect. 4, we present the results of a simulation study in which the new protocol is compared to DRA. Conclusions are presented in Sect. 5.

2 Network Model

We assume the LEO networks are composed of N separate orbit planes, and each has M satellites. The orbit planes and the satellites in a plane are evenly distributed. The following definitions are given.

2.1 Neighbor Nodes (NNs)

The geographical location of a satellite S is given by [lons, lats] indicating the longitude and latitude of the location of S respectively. Each satellite is connected by ISLs (Inter-Satellite-Links) to four satellites: two neighboring satellites in the same orbit (intra-orbit neighbors) and two neighboring satellites in adjacent orbits, one satellite in one left-neighbor orbit and one satellite in one right-neighbor orbit.

2.2 Logical Location (LL)

We suppose that earth surface is covered by a series of satellite nodes with unique LL. The LL is fixed and can be shown as $\langle p, s \rangle$, where p represents the orbit plane, $p = 0, 1, \bigwedge N - 1$ and s represents location of satellite node in orbit plane, $s = 0, 1, \bigwedge, M - 1$.

2.3 Group and Group Head (GH) Selection

All nodes in satellite networks are divided into groups, thus a large network can be virtually partitioned into several sub-networks. We can divide the networks into different groups on the basis of the relation of relative movement between the all nodes in satellite networks.

In general, we select the lighter load node (the node of southern hemisphere near over the Polar Regions) and shorter effective link length (the node at the center of orbit plane) as GH in each group.

In each group, the GH maintains its group members and intra-group link topology information. Group members that connect to adjacent groups are GHs.

3 UDLs Routing Protocol Design

The new UDLs routing protocol proposed in this paper includes UDLs detection, routing calculation and packet transmission.

3.1 UDLs Detection

UDLs detection is the most important premise for satellite networks routing protocol design. According to relative location relationship among satellites, we firstly divide satellite networks into groups. Then, in every group, each satellite sends detection packet to NNs. Based on the parallel computing and UDLs detection method, link state of satellite networks in each group could be detected in parallel. The detailed steps of UDLs detection method are as follows.

(1) Let node A and its neighbor nodes B, C, D and E be shown in Fig. 1. Node A sends Hello detection packets to all NNs, and NNs will reply response in detection time T_{DT}. The discriminant rule between neighbor links states are:

$$P_{LS} = \begin{cases} 1 & T_{DT} \leq T_{DTT} \\ 0 & T_{DT} > T_{DTT} \end{cases} \tag{1}$$

Where P_{LS} is link state. T_{DTT} is time threshold which satisfies

$$T_{DTT} > 2T_{LTD} \tag{2}$$

Where T_{LTD} is link delay. For node B and D, if $P_{LS} = 1$, we can determine that the links A to B and A to D are BDLs, marked as $flag = 0$.

(2) For node C, If $P_{LS} = 0$ and node C can sends packets to one of its NNs except node A, we consider that the link between node A and node C is an UDL, marked as $flag = 1$, which show the link direction is $A \rightarrow C$. Otherwise node C is a failure satellite.

(3) For node E, if $P_{LS} = 0$ and node A can receive the packet from E, we can judge the link between node E and A is an UDL, marked as $flag = -1$, which show the link direction is $E \rightarrow A$.

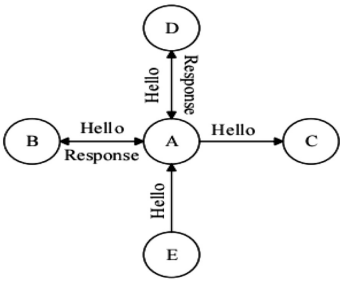


Fig. 1 An example of a uni-directional link

(4) For each node, the steps (1), (2) and (3) are repeated until all nodes links are detected completely.
(5) Each node transfers its link state to GH by effective adjacent link. If the adjacent satellites are more than one, then we compare these satellites' distance between themselves and GH by determine the logical location. We select the shortest distance to GH as transmission direction.
(6) GH incorporates redundant link state, and obtains intra-group final link state detection results.
(7) GHs in every group exchange link state information and then global network link state database can be established.
(8) GH calculates routing table for intra-group all nodes according to global network link state database, and then distributes to the corresponding nodes.

3.2 Routing Calculation

In order to calculate routing, we select the shortest route method between two points in an interconnected network proposed by Nicholson, and modify the method according to the characters of satellite networks. We want to find the shortest route from source node s to terminating node t in satellite networks in which the global network link state is known.

For all network nodes i, the necessary initialization is

$$S(i) = d_{si} \quad T(i) = d_{it} \quad P(i) = s \quad Q(i) = t \tag{3}$$

Where d_{ij} is the distance between the nodes i and j in the direction from node i to node j. If there is no connection or UDLs reverse direction connection between i and j, set $d_{ij} = \infty$. For example in Fig. 1, $d_{AE} = \infty$.

$S(i)$ is the current minimum distance from the source node s to the node i.
$T(i)$ is the current minimum distance from i to the terminating node t.
$P(i)$ is the node preceding i in the current optimal route from s to i.
$Q(i)$ is the point succeeding i in the current optimal route from i to t.

The detailed steps are as follows.

(1) Find the least distance routes out of node s and into node t with the adjacent connected nodes. In order to save precious satellite networks resources, we present different routing calculation scheme to obtain the least distance routes depending on whether source node and terminal node are in intra-group or inter-group.
If a routing session starts, the source node s first sends a route request packet to its GH, if both the source node s and the destination node t belong to the same group, then routing calculation can be conducted in intra-group by GH. But if there cannot establish any path under Uni-directional links existence in intra-group. We need extend the scope into adjacent groups. If s and t are in different groups, we need initiate the inter-cluster route construction. The routes can be calculated by GHs from each group, respectively.
(2) Extend these routes to their adjacent connected points.
   (a) If on any iteration the next shortest route starts from s rather than t, then we proceed as follows: For each route starting at s of next shortest length, examine all connections from its end node m. If there is a connection from s to the node i via the node m, it will be of length $S(m) + d_{mi}$, and this replaces the existing value of $S(i)$ when it is shorter. Then replacement occurs $P(i)$ is set to m.
   (b) If on any iteration the shortest route starts from t rather than s, then we proceed exactly as described above substituting t, $(m) + d_{im}, T(i), Q(i)$.
   (c) The above two steps (a) and (b) given by Nicholson, we consider another situation that we find the two equal least distance routes from out of point s and into point t in step (1), on any iteration the next shortest route starts both from s and t, we parallel proceed as described in (a) substituting s, $S(m) + d_{mi}$, $S(i)$, $P(i)$, and t,$T(m) + d_{im}, T(i), Q(i)$.
(3) Check to see if the shortest route has been found. If on any iteration the least value of $S(i) + T(i)$ (representing a complete route) is at least as small as the sum of the lengths of the shortest route out of s and the shortest route out of t. That is $Min(S(i) + T(i)) \leq MinS(i) + MinT(i)$, then a shortest complete route has been found. The nodes on the shortest route can be traced back through the values of $P(i)$ and $Q(i)$.
(4) Repeat cyclically until the condition in the step (3) is satisfied.
(5) Update corresponding satellite routing table with the newest route information.

3.3 Transmission

During routing calculation, we take into account a variety of possible factors of link states such as UDLs, BDLs, failure satellite nodes and unavailable links. So discarding of data packet could be avoided tremendously. When each satellite node receive packet, it will check its router table according to current time. And the source node can obtain appropriate routing path to the destination node. Using the new protocol, all packets can be transmitted correctly in accordance with the shortest delay path.

4 Performance Evaluation

In this section, we will evaluate the performance of our new protocol. For comparison purposes, we conduct simulation experiments from two aspects. The first one establishes links including UDLs with the new protocol. The second one tends to prune all UDLs from components of links in link state database. Then, links consisting of only BDLs are established. For this case, we choose the typical on-board DRA.

We build the satellite networks model with UDLs which is constructed by 20 nodes. These satellites are divided into 5 groups according to the orbit planes as shown in Fig. 2. Based on the rule of GH selection, nodes 2, 6, 10, 14 and 18 will become the GHs and other nodes in corresponding orbit plane will join relative group. Network performance is evaluated with packet delivery fraction and average end-to-end delay, which are the most important metrics for best-effort traffic. UDLs ratio is defined as the ratio of the number of UDLs to the number of total links in network topology. In order to ensure the feasibility DRA with pruning UDLs, we assume the UDLs ratio is from 0 to 0.5.

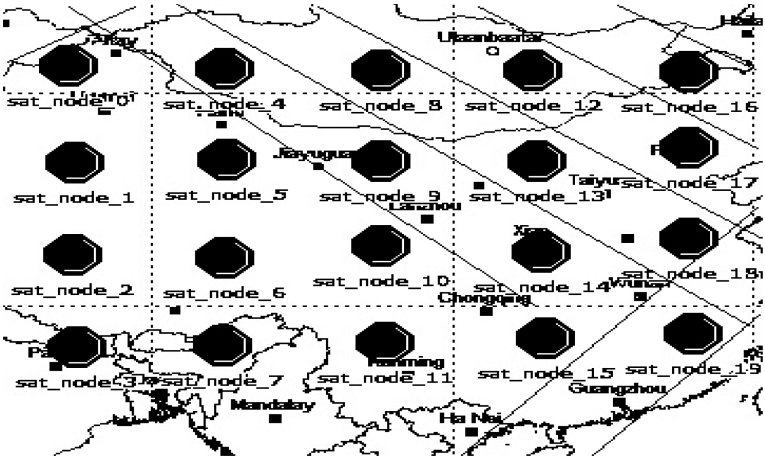


Fig. 2 LEO satellite networks model

Figure 3 shows packet delivery fraction of DRA in pruning UDLs and the new protocol with UDLs in different UDLs ratio. We can observe that the new protocol outperforms DRA as UDLs ratio increases. Moreover, the performance of the new protocol degrades slower and smoother than that of DRA with pruning UDLs, which indicates its durability in LEO satellite networks with UDLs environment.

Figure 4 shows average end-to-end delay for the new protocol and DRA in different UDLs ratio. It is clear that the average end-to-end delay of the new protocol is a little worse than that of DRA. However, in LEO satellite networks with UDLs, because of the link detection, maintenance and network load by control overhead, the delay is larger than that of only BDLs. But we adopt group and parallel computing in link detection and routing calculation to lessen average end-to-end delay in some extent.

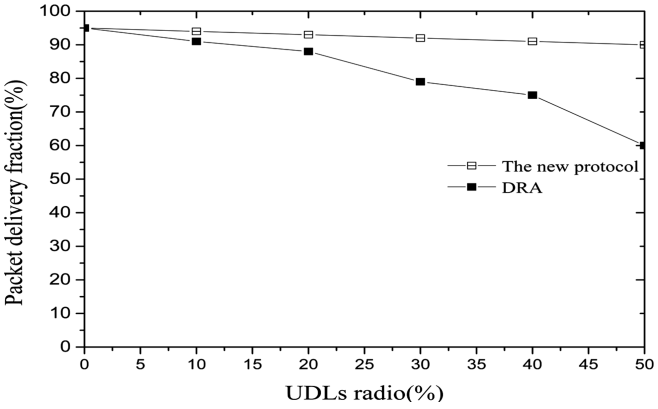


Fig. 3 Packet delivery fraction of the new protocol vs DRA

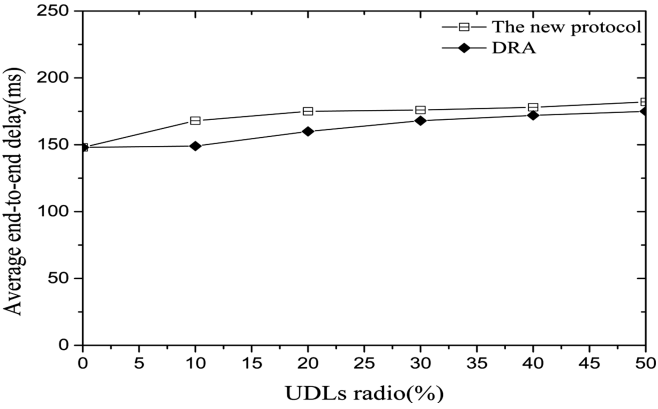


Fig. 4 Average end-to-end delay of the new protocol vs DRA

5 Conclusion

In this paper, we proposed a new routing protocol for LEO satellite networks with UDLs. The new protocol adopted group and parallel computing in link detection and routing calculation to lessen average end-to-end delay. Since only GHs maintain group link topology information, the overhead of exchanging route table of each node is reduced. The simulation results showed that the new routing protocol including UDLs could ensure high packet delivery ratio. Although pruning UDLs can avoid the problem of UDLs, they establish a longer path or fewer numbers of paths, which leads to weaker network connectivity and lower packet delivery ratio than protocols using unidirectional links. Moreover, different routing calculation scheme were used by judgement whether source node and terminal node are in intra-group or inter-group, which saved precious LEO satellite networks resources.

Acknowledgments. This study is supported by grants from the National Natural Science Foundation of China (Grants No: 51375385, 51675425), 111 Project (B08038)

References

1. Wu, Z., Jin, F., Luo, J., et al.: A graph-based satellite handover framework for LEO satellite communication networks. J. IEEE Commun. Lett. **20**, 1547–1548 (2016)
2. Gao, Z., Guo, Q., Na, Z.: Novel optimized routing algorithm for LEO satellite IP networks. J. Syst. Eng. Electron. **22**, 917–925 (2011)
3. Lu, Y., Zhao, Y., Sun, F., et al.: Routing techniques on satellite networks. J. Softw. **25**, 1085–1100 (2014). (in Chinese)
4. Riviou, K., Kouroupetroglou, G., Oikonomidis, N.: A network of peers and practices for addressing learner variability: UDLnet. J. Stud. Health Technol Inform. **217**, 32–39 (2015)
5. Hu, H., Liu, Y.: An efficient A-DSR routing algorithm for asymmetric space information network. In: 2010 IEEE International Conference on Information Theory and Information Security (ICITIS), pp. 880–883. IEEE Press, Beijing (2010)
6. Liu, Z., Li, W., Long, H., et al.: The Simulation and analysis of energy-related DSR in Opnet. In: Proceeding of the 27th Chinese Control Conference, pp. 275–279. IEEE Press, Kunming (2008)
7. Jia, X., Lv, T., He, F., et al.: Collaborative data downloading by using inter-satellite links in LEO satellite networks. J. IEEE Trans. Wirel. Commun. **16**, 1523–1524 (2017)
8. Liu, H., Sun, F., Yang, Z., et al.: A novel distributed routing algorithm for LEO satellite network. In: 2012 International Conference on Industrial Control and Electronics Engineering, pp. 37–40. IEEE Press, Xi'an (2012)



Measuring Quality of Experience of Novel 360-Degree Streaming Video During Stalling

Wenjie Zou(✉) and Fuzheng Yang

State Key Laboratory of Integrated Services Networks, Xidian University, Taibai Road. 2, Xi'an 710071, China
wjzou@xidian.edu.cn, fzhyang@mail.xidian.edu.cn

Abstract. With the rapid growth of novel 360-degree video streaming applications, there has been a strong demand of quality-of-experience (QoE) measurement. For the novel experience provided by the 360-degree video, there still lack of corresponding research and usable evaluation model. In this paper, the user's QoE during the stalling event for the 360-degree streaming video viewed on the Virtual Reality (VR) Head-Mounted Display (HMD) is obtained for the first time by an elaborately designed subjective experiment. The impact of stalling event on the QoE of 360-degree video streaming is then investigated. We find that there is a significant difference between the QoE during the stalling for the 360-degree video streaming and the traditional video streaming, which indicates that the existing QoE evaluation models are no longer suitable for evaluating the QoE of 360-degree video streaming. This preliminary finding appeals for more researches on modeling the specific QoE of 360-degree video streaming.

Keywords: Virtual reality · 360-degree video streaming
Quality of experience · Stalling event

1 Introduction

In recent years, the virtual reality (VR) technology has had an explosive development. Corresponding products, such as the VR device, content, and software, are fast becoming an ordinary consumer item. It is reported that 4% of US consumers already own a VR device, which means that over 9 million US consumers already own a VR device [1]. Simultaneously, VR is predicted to have a global user base of more than 275 million by 2025 [2]. As one of the most important part of VR content besides the VR games, the VR video, a.k.a., spherical/360-degree video, also has had a fast development in the past few years. For instants, major multimedia streaming service providers, such as YouTube [3] and Facebook [4], have already launched the 360-degree video channels in 2015 to supporting 360-degree video streaming for VR devices and offered a huge number of 360-degree video streaming online. The VR devices, such as the GearVR and HTC Vive, also provide lots of 360-degree streaming videos to the consumers.

In contrast to the traditional video streaming, the 360-degree video streaming can record all 360-degrees of a scene, simultaneously. However, this type of video cannot be displayed with traditional method. Typically, users view the 360-degree videos on the Head-Mounted Display (HMD). Viewers can rotate a 360 video's perspective to

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 417–424, 2018.
https://doi.org/10.1007/978-3-319-78130-3_43

watch the immersive 360 space from different angles. Coupled with the high field of view (FOV) provided by the HMD, the 360-degree video streaming can provide a relatively immersive experience, which is significantly different with the experience provided by the traditional video streaming. Commonly, when viewers watch the 360-degree video streaming on a HMD, only part of the 360-degree scene can be seen at a specific moment, which means that only part of the whole image is needed to be display. For instants, the prevailing level of FOV on the VR HMD is about 110 × 110 degree. Assuming that the 360-degree video streaming is also displayed with these display FOVs. The display area is only approximate 1/5 of the whole sphere surface. By contrast to the traditional streaming video, the 360-degree video streaming usually needs a higher resolution, such as 4K (3840 × 1920), and thus a higher coding bitrate to ensure the video quality.

For these 360-degree video streaming services, HTTP transport is usually employed to ensure a reliable delivery of video streams through TCP and enable the user to play incompletely downloaded video by using the progressive download technology [5]. Moreover, the TCP-based streaming liberates the video decoders from handling packet losses and prevents the QoE from the influence caused by packets loss. However, due to the channel throughput fluctuations, when the TCP throughput is lower than the playback rate and the buffered video is played out, the video playback will pause and wait for new video data. This unpleasant playback pauses, also called stalling, can greatly impact the user's quality of experience (QoE) [6–8], which describes an end user's holistic perception and satisfaction. Although some HTTP Adaptive Streaming (HAS) technologies have been proposed to mitigate the risk of playback interruptions by adapting the bitrate of the video stream to the network throughput, it is unable to completely get rid of the impairment of stalling in playback [9]. Hence, it is critical to evaluate the users' QoE for the HTTP-based 360-degree video streaming services, which can provide a constructive guide for service providers and network operators to optimize their services.

In the past decades, a number of researches targeted at the evaluation of the impairment of rebuffering on users' QoE of HTTP-based video streaming service, as summarized in [10]. However, to the best of our knowledge, there is still lack of a research focusing on the QoE evaluation for the 360-degree streaming video, especially on the impairment of stalling. In fact, how much the influence of the novel viewing manner of 360-degree video streaming in a VR system on the overall QoE is remain unclear. Therefore, it is quite meaningful to investigate the feature of QoE of 360-degree video streaming during the stalling to solve the problem mentioned.

In this paper, we focus on investigating the impact of stalling on the QoE of 360-degree video streaming and try to find out whether this impact is different on the QoE of 360-degree video streaming and that of traditional video streaming. The viewers' QoE during the stalling event are obtained for the first time by an elaborately designed subjective experiment. Based on these subjective QoE, we find that the variation of QoE during the stalling for the 360-degree video and traditional video is significantly different. The QoE of 360-degree video streaming has its own specific characteristic. Hence, it is suggested to propose new specific QoE evaluation model for the 360-degree video streaming, but not just utilize the existing QoE evaluation model to assess the QoE of 360-degree video streaming, directly.

2 Subjective Experiment for QoE of 360-Degree Video Streaming

To the best of our knowledge, current publicly available database are dedicated to the traditional video. Just a few of researches focus on the video presentation quality of the 360-degree video. However, there is still lack of a 360-degree video streaming database for investigating the impact of stalling event on the 360-degree video streaming viewed on the HMD. Here, we develop a dedicated database to study the impact of stalling event. By contrast, the 360-degree video streaming is not only viewed with the novel manner on the HMD, but also viewed with the traditional manner on the PC monitor. The QoE during the stalling can be obtained by the elaborately designed subjective experiment.

2.1 Test Material and Experiment Design

A video database of 12 pristine high-quality 360-degree videos (I420 format, 30 fps, Equirectangular projection, monoscopic vision [11]) of size 4K (3840 × 1920) were selected to cover diverse content, including humans, plants, natural scenes, and architectures. The length of each video was 20 s. A screenshot from each video is included in Fig. 1. Using aforementioned sequences as the source, the videos were encoded into two different quality levels with x264 encoder, the detail of the setting is listed in Table 1. Here, we encoded the video with constant QP to make the video quality as close as possible for the different video contents.

For obtaining the QoE during the stalling event, the encoded videos were displayed with different length of stalling, as shown in Table 1. The rated results were utilized to represent the users' QoE at the end of stalling event. In this way, the subjective QoE at the discrete-time point during the stalling event can be obtained.

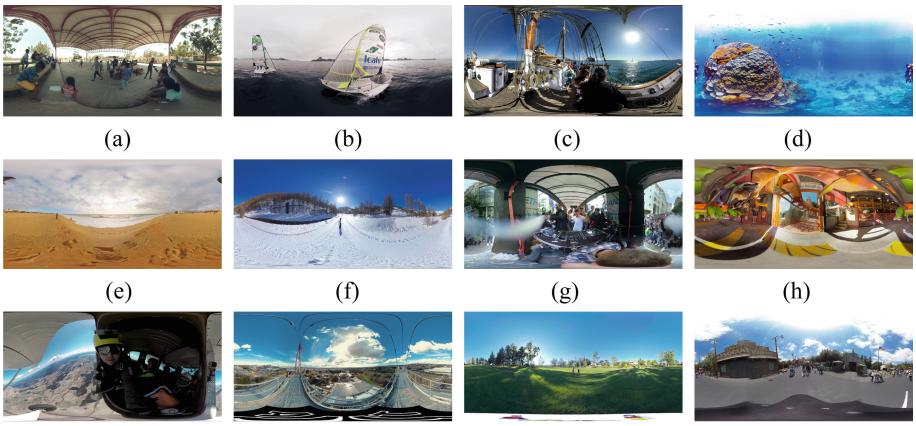


Fig. 1. Test sequences. (a) Playground. (b) Sailboard. (c) Sailing. (d) Cay. (e) Sand beach. (f) Skiing. (g) Carnival Parade. (h) Roller coaster. (i) Skydiving. (j) Stroll. (k) Campus. (l) Lawn.

Table 1. Setting of experiment.

Name	QP	Playback (s)	Stalling (s)	Resume playback (s)
Playground	20	18	0, 0.1, 0.2	2
Sailboard	20	18	0.5, 1	2
Sailing	20	18	2, 4	2
Cay	20	18	6, 10	2
Sand beach	20	18	15, 20	2
Skiing	20	18	30, 50	2
Carnival parade	32	18	0, 0.1, 0.2	2
Roller coaster	32	18	0.5, 1	2
Skydiving	32	18	2, 4	2
Stroll	32	18	6, 10	2
Campus	32	18	15, 20	2
Lawn	32	18	30, 50	2

2.2 Experimental Procedures and Data Screening

A total number of 25 non-expert subjects participated in this experiment, including 13 males and 12 females aged between 21 and 30 years with a mean of 24. All of them reported normal or correct-to-normal sight and pasted the eyesight check/color blindness test.

The subjective testing experiment was setup as a normal indoor home settings with ordinary illumination level, with no reflecting ceiling walls and floors [12]. A vogue HMD and a 32 inch LCD monitor with a native resolution of 3840 $\times$ 2160 pixels were utilized as the display devices. The HMD has a screen with the native resolution of 2160 $\times$ 1200 pixels and a 110 degree horizontal FOV. The video playback was supported by a high performance server. Moreover, a uniform 360-degree video player with the projection type of the sphere was developed to display the videos on both monitor and HMD. For the convenience of rating, a scoring tool was developed to let the subject use the controller to rate scores within the HMD. The study adopted a single-stimulus quality scoring strategy. Each subject viewed the videos on the monitor and HMD, respectively. The order of the video sequences was random.

During the viewing of 360-degree video, the subject should change the viewpoint by rotating his/her head when viewing on the HMD. However, the direction of the 360-degree video displayed on the monitor was set to be constant to simulate the manner that the traditional video displayed. The display area on the 4K monitor was alterable. It was set to allow the video to be displayed with 1:1 in the area and same with the region displayed on the HMD. The distance from the viewer to the monitor was set to be 3 times of the height of display area as specified in [12]. Before the test on the HMD, the subject must adjust both the distance between lenses (IPD) and the distance from the user's eyes to the lenses (lens-to-eye) to calibrate for the clearest view. Between each two video sequences, there was an interval, which depended on the duration subject took to rate the score. He/she could rate his/her overall QoE in the Absolute Category Rating (ACR) 5-point scale (corresponding to the experience of

"excellent," "good," "fair," "poor," and "bad" from 5 to 1 point) during the interval. Before the formal test, the subjects were asked to rate a few example videos to get familiar with the scoring scale and the scoring tool.

After the subjective tests, the credibility of assessment results was checked using the linear Pearson correlation coefficient (LPCC) suggested by ITU-T P.913 [12], and no subject was discarded. The QoE of each video sequence was measured in terms of the Mean Opinion Score (MOS) [12].

3 Objective QoE During the Stalling Event for 360-Degree Videos

The main objective of this subjective experiment is to investigate the impact of stalling event on the QoE of 360-degree video streaming. Here, the QoE of 360-degree video during the stalling is firstly investigated based on the subjective scores rated on the HMD. The relationship between the duration of stalling and the QoE is shown in Fig. 2. It can be observed that the QoE drop increases with the increase in the duration of stalling. It is interesting that the QoE of 360-degree video during the stalling becomes gradually steady after a quick decrease of QoE during a few seconds at the beginning of stalling. After the steady stage, the QoE continues to drop quickly.

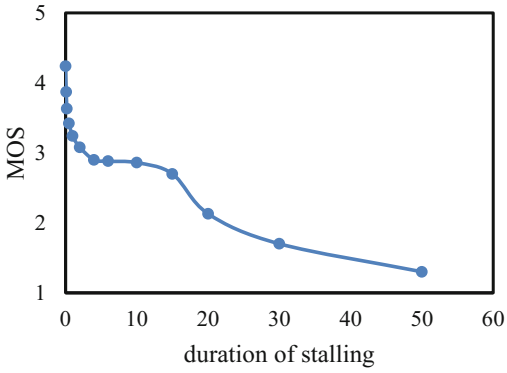


Fig. 2. The relationship between the duration of stalling event and the QoE for the high quality video viewed on the HMD.

By contrast, a significantly different phenomenon can be observed on QoE for the 360-degree video streaming viewed with traditional manner, as shown in Fig. 3. The QoE drop conforms to the common exponential decay of QoE [13–15] for traditional streaming during the stalling.

One explanation of this difference may be that the special manner of viewing for the 360-degree video streaming is significantly different with that for the traditional video streaming. More specifically, for the common manner of viewing for traditional video streaming, the information of the freezing picture is quite limited. The viewer has to

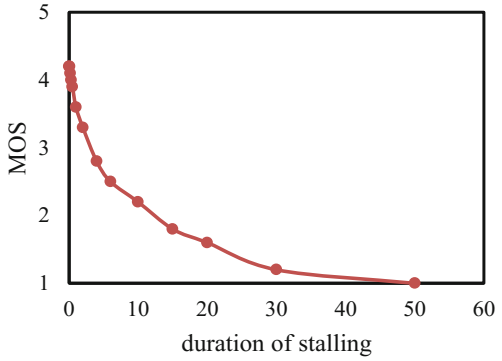
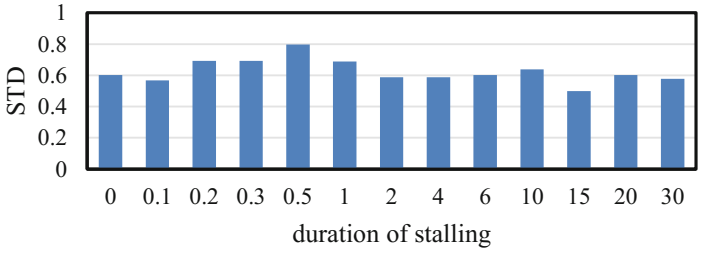
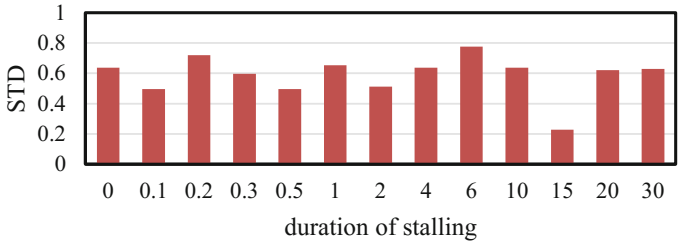


Fig. 3. The relationship between the duration of stalling event and the QoE for the high quality video viewed on the monitor.

just get through the whole stalling event. By contrast, the viewer can watch the other part of the 360-degree scene of the freezing 360-degree video streaming. During this period, the dissatisfaction caused by the stalling for the 360-degree video streaming is quite smaller than that for the traditional video, as the differences among the MOS for



(a)



(b)

Fig. 4. The relationship between the STD of the individuals' rated results and corresponding duration of stalling event. (a). Video viewed on the HMD with the novel manner of viewing. (b). Video viewed on the monitor with the traditional manner of viewing.

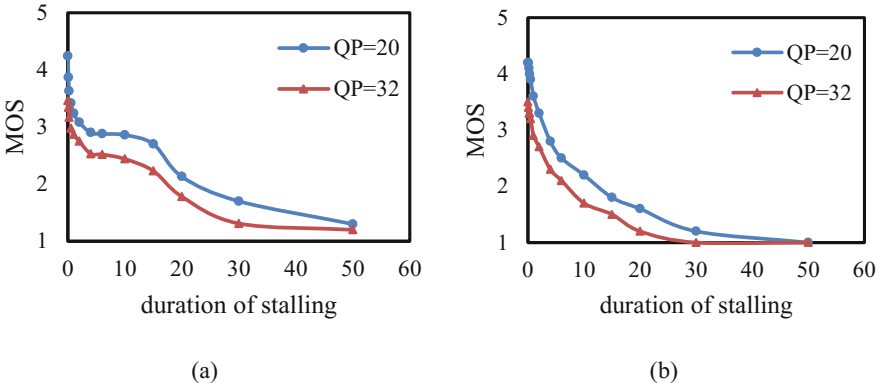


Fig. 5. The relationship between the QoE and the duration of stalling event. (a). Video viewed on the HMD with the novel manner of viewing. (b). Video viewed on the monitor with the traditional manner of viewing.

the videos rated on the HMD and monitor in Figs. 2 and 3. After finishing watching the freezing 360-degree scene, the QoE continuous to decrease relatively fast. From a general view, the rate of QoE drop for the traditional manner of viewing is significantly higher than that for the manner of viewing for 360-degree video streaming frozen the same length of time.

For the deviation between the individuals, Fig. 4 shows the standard deviation (STD) of the rated results of individuals for each point plotted on the Figs. 2 and 3. It can be seen that the values of STD are small, which indicated that although there exists a slight difference among the individuals' judgments for the QoE of the video interrupted by the same stalling event, the overall trends of the QoE drop are the quite similar.

Moreover, for the video encoded with different quality levels, i.e., QP = 20 and QP = 32, the higher the quality level of the video is, the faster the QoE drops, as shown in Fig. 5. It is noteworthy that this phenomenon can be observed from the QoE drop in both manners of viewing. One explanation of this phenomenon may be that there is a higher viewer expectation when the video quality is high, and thus the interruption caused by stalling make them feel more frustrated.

4 Conclusion

In this paper, a subjective experiment was conducted for the novel 360-degree video streaming on a vogue VR HMD and a 4K monitor. According to the experiment results, it is found that the variation of QoE during the stalling for the 360-degree video and traditional video is significantly different. A relatively stable QoE can be maintained after a fast QoE drop during a few seconds at the beginning of stalling. This phenomenon is considered caused by the special manner of viewing of 360-degree video streaming. This finding reveals that the existing QoE evaluation models for the

traditional video streaming are no longer suitable to be utilized to evaluate the QoE of 360-degree streaming videos, which motivates us to conduct more subjective experiments for modeling the specific QoE of 360-degree streaming videos.

Acknowledgments. This work was supported by the National Science Foundation of China under Grant 61601349, 61571337, and by the 111 Project B08038.

References

1. Digital Democracy Survey: A multi-generational view of consumer technology, media and telecom trends, Deloite's Technology, Media and Tele-communications (TMT) practice, March 23, 2016. https://www2.deloite.com/content/dam/Deloite/glob-al/Documents/Technology-Media-Telecommunications/gx-tmt-deloite-democracy-survey.pdf
2. The Future of Consumer VR and Its Impact on Video Viewing, 2016–2025: The Diffusion Group, January 6, 2016. http://tdgresearch.com/report/the-future-of-consumer-vr-and-its-impact-on-video-viewing-2016-2025/
3. Youtube. https://www.youtube.com
4. Facebook. https://www.facebook.com
5. Adobe. Delivery: Progressive download. Video Technology Center, http://www.adobe.com/devnet/video/progressive.html
6. Krishnan, S.S., Sitaraman, R.K.: Video stream quality impacts viewer behavior: inferring causality using quasi-experimental designs. IEEE/ACM Trans. Netw. **21**(6), 2001–2014 (2013)
7. Definition of Quality of Experience (QoE). Rec. TD 109rev2 (PLEN/12) ITU-T, Geneva, Switzerland (2007)
8. Song, J., Yang, F., Zhou, Y., et al.: QoE Evaluation of Multimedia Services Based on Audiovisual Quality and User Interest. IEEE Trans. Multimedia **18**(3), 444–457 (2016)
9. Yao, J., Kanhere, Salil S., Hossain, I., Hassan, M.: Empirical Evaluation of HTTP Adaptive Streaming under Vehicular Mobility. In: Domingo-Pascual, J., Manzoni, P., Palazzo, S., Pont, A., Scoglio, C. (eds.) Networking 2011. LNCS, vol. 6640, pp. 92–105. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20757-0_8
10. Seufert, M., Egger, S., Slanina, M., Zinner, T., Hobfeld, T., Tran-Gia, P.: A survey on quality of experience of HTTP adaptive streaming. IEEE Commun. Surv. Tutor. **17**(1), 469–492 (2015)
11. Howard, I.P., Rogers, B.J.: Binocular Vision and Stereopsis, pp. 31–33. Oxford University Press, USA (1995)
12. ITU-T P. 913: Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in Any Environment (2014)
13. Fiedler, M., Hossfeld, T., Tran-Gia, P.: A generic quantitative relationship between quality of experience and quality of service. IEEE Netw. **24**(2), 36–41 (2010)
14. Hoßfeld, T., Seufert, M., Hirth, M., Zinner, T., Tran-Gia, P., Schatz, R.: Quantification of YouTube QoE via crowdsourcing. In: IEEE International Symposium on Multimedia (ISM). IEEE, pp. 494–499 (2011)
15. Hoßfeld, T., Schatz, R., Biersack, E., Plissonneau, L.: Internet video delivery in YouTube: from traffic measurements to quality of experience. In: Biersack, E., Callegari, C., Matijasevic, M. (eds.) Data Traffic Monitoring and Analysis. LNCS, vol. 7754, pp. 264–301. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36784-7_11



A Realizable and Efficient Approach for Building Software Defined Satellite Network

Wei Ma(✉), Leifang Hui, Hanwen Sun, and Zhe Zhao

China Academy of Space Technology, No. 504, East Chang'an Road, Xi'an, People's Republic of China
mawei_mail@sina.com

Abstract. As a part of the 10 key enabling technologies for 5G, SDN and NFV bring great influence and inspiration to satellite network architecture. SDN/NFV-enabled satellite network building method in transparent satellite system has been researched; however less attention has been paid to the regenerative process satellite system. In addition, satellite system is a typical restricted system, the theoretical research results and successful technology solutions of the ground network are difficult to be applied directly to the satellite system. In this paper, we study how to build a software defined satellite network based on regenerative process satellite system, how to support the software definition of satellite nodes, and put forward a realizable and efficient approach for building software defined satellite network, on this basis, the application scenario of the software definition satellite network is prospected.

Keywords: SDN · NFV · Software defined satellite network

1 Introduction

Satellite system is a typical restricted system (weight, size, power dissipation, component selection, etc.) with long research and development cycle, high-cost and high-risk. Moreover satellite system has broad prospects for development which integrated with ground network can bring huge economical and market efficiency. Diversified business and application requirement, fast evolution of technology make the weakness of traditional satellite system (i.e. absence of programmable ability, function difficult to be modified in life cycle, ability difficult to be extended, etc.) be amplified time and again. Furthermore, single satellite render services for multiple different systems simultaneously turn into subsequent trend of development due to the shortage of space frequency and orbit place resource.

So far, software defined network (SDN) and network function virtualization (NFV) are parts of the 10 key enabling technologies for 5G [1], such thoughts like decoupling the control and data planes of the network equipment and logically centralizes the network intelligence, network slice, etc. will produce far-reaching effects on the development of satellite network. At the same time we should realize that the special position of satellite system in country level, specific feature of space

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 425–432, 2018.
https://doi.org/10.1007/978-3-319-78130-3_44

environment and constraint condition make it difficult or unsuited for relevant theoretical achievement and technical solutions in ground network to apply to satellite network directly. In terms of such background, this paper launches relevant research work in software defined satellite network architecture, node implementation structure and application scenarios.

2 Network Architecture

Based on transparent satellite, SDN/NFV-enabled satellite network towards 5G can be constructed through the introduction of SDN/NFV technologies in the satellite ground segment domain, a number of use cases in different scenarios can be developed quickly in this way [2].

Different from above pattern, this paper focus on software defined satellite network architecture and node implementation structure based on regenerative process satellite. In this situation, software defined satellite network architecture use classical hierarchical architecture which includes data plane, control plane and application plane, as Fig. 1 illustrated.

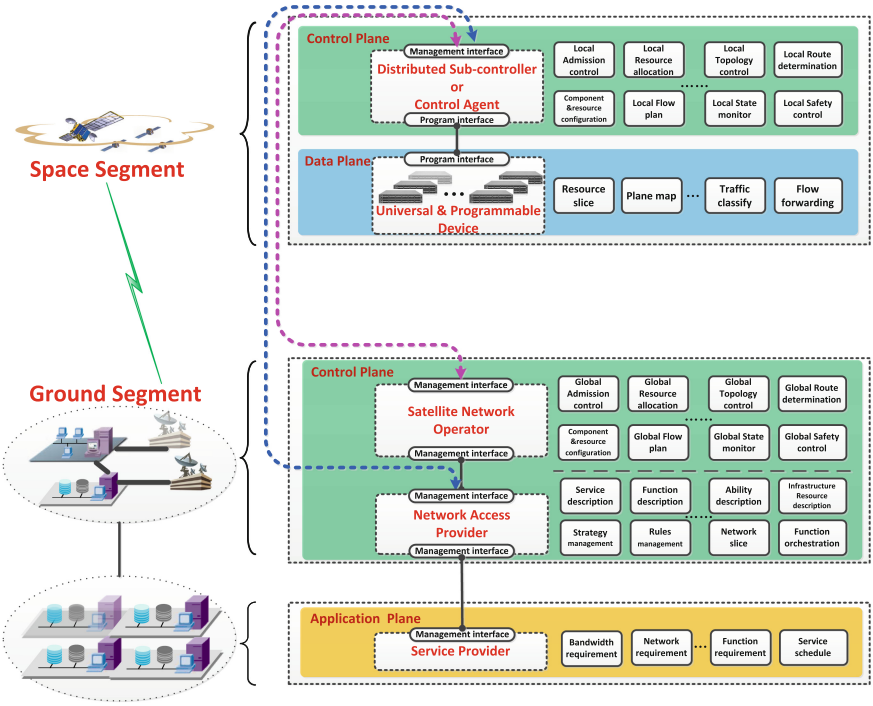


Fig. 1. Software defined satellite network architecture

Distinguish from centralized control mode in ground SDN, software defined satellite network choose dynamic hierarchical control mode, in which the function of control plane should be realized by satellite and ground infrastructure together out from considerations as follows:

- Radio link is the principal link of satellite network, big propagation delay time will be introduced by long distance and multi-hop, processing requirement of strong real-time service is difficult to be satisfied, besides link transmission bandwidth, process capacity of ground control center etc. should turn into the ability bottleneck of whole system.
- As a significant part of national information infrastructure satellite network play a key role in national strategy, it is necessary to consider the self-operation and management requirement for satellite network carefully independent of ground control.
- As a public infrastructure for military, commercial and civilians, different service providers have explicit requirement for private network, custom service and independent control relatively, it is difficult to satisfy different appliance requirement in single central control mode.

The functions of different plane in software defined satellite network are described as follows.

2.1 Data Plane

Data plane consists of universal programmable equipment, which offer standardized program interface to control plane and fulfill resource slice (computation, logic, storage, etc.), field extraction, virtual data plane map, traffic classification, flow forward, etc. process following the dispatch rules of control plane.

2.2 Control Plane

Control plane make management domain plan combined with topological optimization control strategy, which select suitable satellite and deploy local domain controller (usually every satellite need deployment of at least one controller or control agent). Satellite node controller in charge of user management, access control, resource allocation, topology control, path computation, state monitor, safety control, etc. functions of local domain, moreover receive and follow the control strategy and process rules dispatched by Satellite Network Operator or Network Access Provider, accomplish network resource allocation, virtual data plane map, traffic classification, flow forward, etc. process.

Ground Satellite Network Operator act as the central controller, which can master global state information of different domain and implement control to disperse satellite node controller in diverse domain, accomplish maintenance, management, deployment and operation of satellite network infrastructure.

Ground Network Access Provider is responsible for management and operation of satellite network access element (satellite terminal, gateway, etc.), besides accomplish service description, function description, ability description and resource requirement description map according to particular mission demand of different service provider,

plan matched network slice, strategy, rules and function orchestration, etc. and complete configuration of corresponding controller on the basis of mapping result, thus provide bulk transport resource for diverse service providers.

2.3 Application Plane

Application plane towards different service providers, service provider sell corresponding services and equipment to diverse user group and fulfill bandwidth, network, function, etc. requirement mapping according to the ability demand to render specified service, then submit it to Network Access Provider to handle.

3 Node Implementation Structure

The core thought of SDN is decouple control plane of network equipment with data plane, only keep forward function at the bottom of network hardware and carry out central control on the top, then achieve the programmable ability of network application and function, namely so-called software defined. NFV is redefinition to the architecture of net element equipment, which separate network service from the tight coupling relation with special hardware and position [3]. NFV and SDN have strong complementarity, despite both can be reconciled, not interdependence, that is to say NFV can be deployed separately independent of SDN [4]. The comparison of SDN and NFV architecture as follows (Fig. 2):

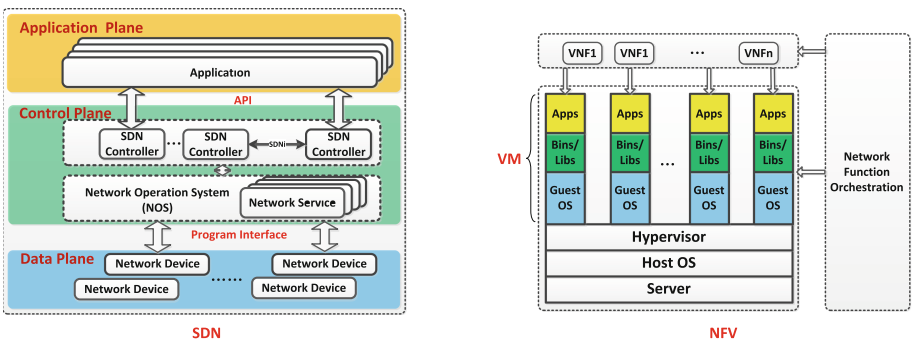


Fig. 2. Comparison of SDN and NFV architecture

The hardware foundation of SDN is standardize programmable network equipment (i.e. router, switch, storage device, etc.), on this basis decoupling of control plane and data plane can be achieved, meanwhile standardization of interface between different planes and central control is stressed [5].

The hardware foundation of NFV is high performance server, based on virtualization technology multiple VM should be activated simultaneously; specific VM function can be customized and orchestrated by diverse service provider, thus network function software-enabled can be realized [6]. Network isolation and the addition, alteration and migration of network functions can be achieved conveniently in this way.

Satellite system is a representative circumscribed system, there is severely restriction for the weightiness, volume, power dissipation, component application, space single particle effect resist ability, etc. of space network equipment, equipment on board is difficult to reach the ability level of ground high performance server. Software on board usually use embedded real-time operation system, utilization of virtualization software such like Hypervisor, VMWare, etc. on board is still immature. Although the impertinent way of NFV is incapable to copy in satellite network, the design thought of NFV should be used for reference. The network equipment of SDN is usually been bring about dependent on specific switch chip, NPU, etc. compared with which more FPGA, CPU is adopted in satellite equipment. At present satellite network equipment is on the evolutionary development way toward standardization, programmable direction, but in consideration of the distinguishingly position of satellite network and safety factor, direct application of standards such as Openflow, etc. on board may be restricted.

In the implementation structural design of satellite node, above-mentioned reality constraint condition should be considered fully, to guarantee the scheme realizable; On the other hand the fast evolution and development requirement of processing technologies on board should be took into account, to make the node implement structure extendable, which is capable of supporting follow-up upgrade and technical smooth transition. The satellite node implement structure proposed in this paper as Fig. 3.

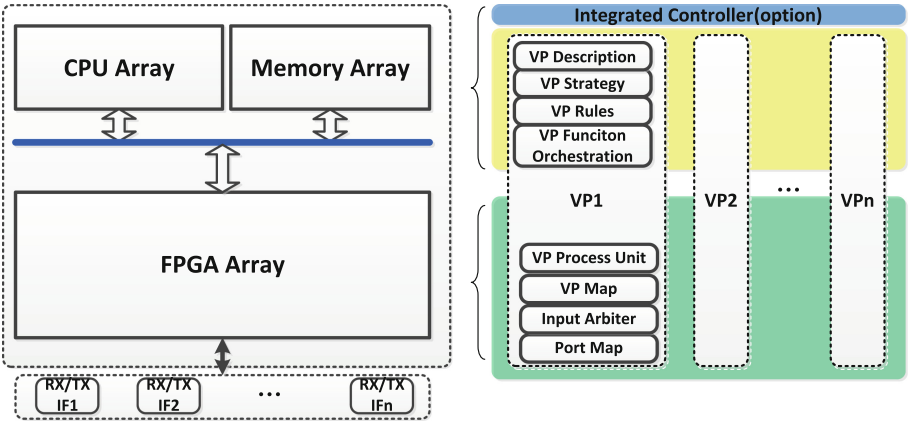


Fig. 3. Satellite node implementation structure

Universal, programmable design thought is adopted in satellite node, kernel process function is achieved through the usage of FPGA array, CPU array and storage array, etc. In the aspect of network equipment resource slice, NFV thought is used for reference and virtual plane (VP) [7] is introduced to support resource slice (computation, storage, logic, etc.), custom-made for process rules and function in multi-plane. The function of equipment data plane is mainly realized on account of FPGA array, the function of equipment control plane is mainly achieved in view of CPU array, the interface definition between control plane and data plane is standard and programmable which draw lessons from SDN. Network controller can be designed independently, also can be integrated in specified network equipment of satellite.

In the aspect of technical compatibility from start to finish, constraint bottleneck mainly lies in the maturity of correlated techniques on board and control granularity can be implemented. Subject to the handling capacity of satellite equipment and technical maturity of virtualization software, the application effect similar to VM can be achieved for satellite node by using VP pattern, and with the unceasing development of process technology on board VP pattern can transition to VM progressively. The granularity of resource vertical segmentation can be developed from component level to complete resource sharing and segmented by software step by step. NOS on board should be developed toward universalization, standardization and platform progressively. Aforesaid technology evolution and development requirement can be satisfied through the use of satellite node structure proposed in this paper.

4 Application Scenarios

Software defined satellite network architecture changed the function deployment way in traditional satellite system revolutionarily, coefficient of utilization of satellite resource can be promoted effectively, apply styles of satellite system should be enriched greatly. Typical software defined satellite network application scenarios can be predicted include:

Scenario 1- network resource sharing and segmentation according to need. Network resource is capable of segmentation based on VP technology, multiple VP should be in motion simultaneously and independently. Same network facility can render services for military, commercial and civil users at the same time, in addition the isolation of different networks with distinct security level can be guaranteed. Such application scenario as follows (Fig. 4):

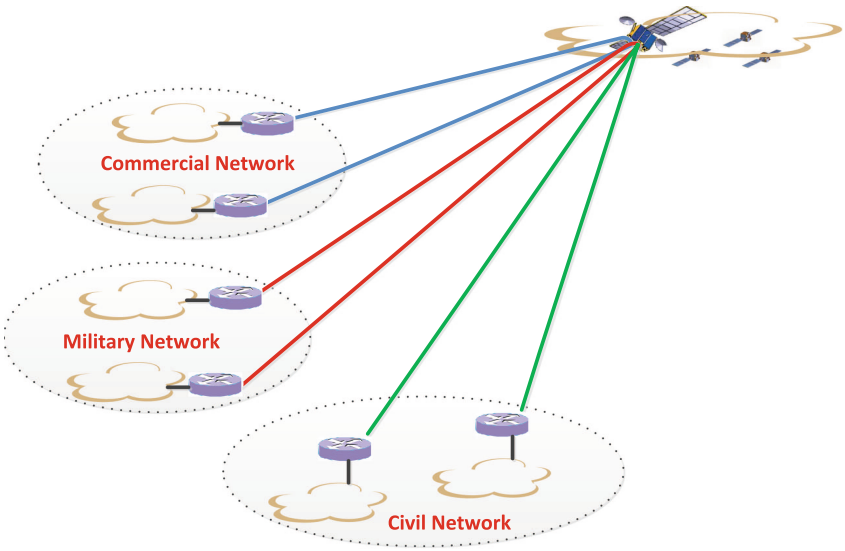


Fig. 4. Scenario 1- network resource sharing and segmentation according to need

Scenario 2- network function deployment and orchestration according to need. The functional network element of satellite node can be custom made and instantiation by users according to need, multiple diverse functional network elements can be instantiated simultaneously in the same satellite, thus variegated network functions should be realized. One pluripotent satellite can be achieved in this way, thereby demand pressure of space frequency, orbit space resource should be relieved to a great extent. Such application scenario as follows (Fig. 5):

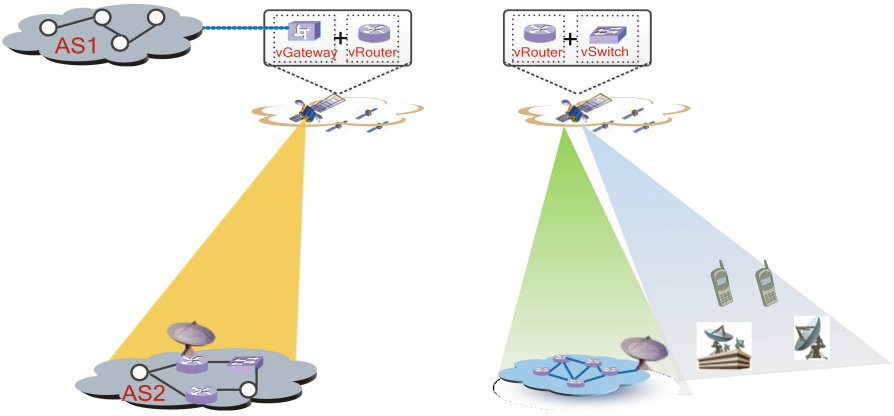


Fig. 5. Scenario 2- network function deployment and orchestration according to need

Scenario 3- network function copy and migration according to need. Users can choose different satellite to backup, copy and migrate needful network functions conveniently. Such application scenario as follows (Fig. 6):

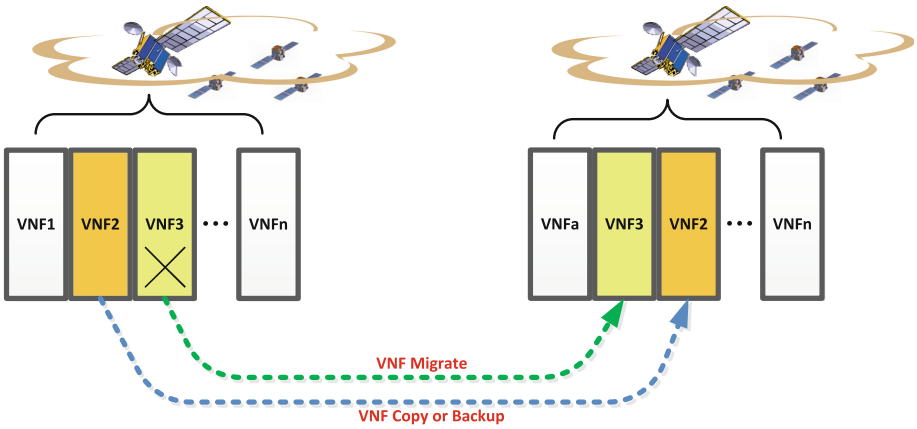


Fig. 6. Scenario 3- network function copy and migration according to need

5 Conclusion

By building a software defined satellite network, satellite system can adapt to rapid technological evolution, meet diverse and personalized task needs, and make valuable satellite resources use efficiently in its whole life cycle. The approach of building software defined satellite network proposed in this paper takes fully into account the capacity status of satellite system and the future technology development trend, and makes it possible to build a software defined satellite network progressively. With the development of satellite processing technology and the continuous upgrading of capabilities, the degree of virtualization of resources and the customization granularity of the function will be greatly improved in the future.

References

1. Akyildiz, I.F., Nie, S., Lin, S.-C., Chandrasekaran, M.: 5G roadmap: 10 key enabling technologies. J. Comput. Networks **106**, 17–48 (2016)
2. Ferrús, R., Koumaras, H., Sallent, O., Agapiou, G., Rasheed, T., Kourtis, M.-A., Boustie, C., Gélard, P., Ahmed, T.: SDN/NFV-enabled satellite communications networks: opportunities. Scenar. Chall. J. Phys. Commun. **18**, 95–112 (2016)
3. Baldonia, G., Melitab, M., Micalizzia, S., Ramettaa, C., Schembraa, G., Vassalloa, A.: Video broadcasting services over SDN-NFV enabled networks: a prototype. J. Procedia Comput. Sci. **98**, 560–565 (2016)
4. Schaller, S., Hood, D.: Software defined networking architecture standardization. J. Comput. Stand. Interfaces **54**, 197–202 (2017)
5. Masoudi, R., Ghaffari, A.: Software defined networks: a survey. J. Network Comput. Appl. **67**, 1–25 (2016)
6. Lin, Q., Qi, Z., Jiewei, W., Dong, Y., Guan, H.: Optimizing virtual machines using hybrid virtualization. J. Syst. Softw. **85**, 2593–2603 (2012)
7. Liao, Y., Yin, D., Gao, L.: Network virtualization substrate with parallelized data plane. J. Comput. Commun. **34**, 1549–1558 (2011)



Proactive Edge Caching Strategy Based on Mobility Prediction in Dense Small Cell Networks

Tianming Song[1], Heli Zhang[1(✉)], Xi Li[1], Chunsheng Zhu[2], and Hong Ji[1]

[1] Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing, People's Republic of China
{stm,zhangheli,lixi}@bupt.edu.cn, jihong@bupt.edu.cn
[2] Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada
chunsheng.tom.zhu@gmail.com

Abstract. Content caching on the edge of dense networks is an emerging and critical technology to support the thirst for content of mobile users. However, the user mobility imposes additional difficulties on caching design because a moving user may not be able to obtain a whole file from a single small base station (SBS). In this paper, we consider the content caching problem in a dense small cell networks. We proposed a proactive mobility-aware (PMA) caching strategy combined with users mobility prediction. In this strategy, we first predict future locations of mobile users, then optimally cache content at SBSs where mobile users are most likely to access in the future. We evaluate efficacy through simulation and compare our strategy against a commonly used cache strategy. Simulation results of our caching strategy show a considerable improvement on the users cache hit ratio.

Keywords: Caching strategy · Dense small cell networks
Mobility-aware caching

1 Introduction

In future fifth generation (5G) networks, the mobile data traffic will grow at a compound annual growth rate (CAGR) of 47 percent in the next 5 years, reaching 49.0 exabytes per month by 2021 [1]. In order to support mobile users in such wireless traffic, content caching at the network edge (i.e., SBSs, user devices etc.) has been proposed as a cost-efficient solution to offload cellular/backhaul traffic, which also satisfies users' QoE and alleviating network congestion [2,3].

Some prior arts make good contributions on edge caching schemes. In [4], an architecture based on distributed caching of the content in SBSs was presented. Due to the short distances between mobile users and cache-enabled SBSs, the

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 433–442, 2018.
https://doi.org/10.1007/978-3-319-78130-3_45

transmission of files can be done very efficiently. A geographical caching approach, aiming to maximize the probability of finding the requested files at the SBSs, has been established [5]. In [6], the problem of distributed caching in dense wireless small cell networks (SCNs) is investigated using mean field games (MFGs). The problem of content placement for caching at the wireless edge with the goal to maximize the energy efficiency has also been studied [7].

Since the movement of users change the topology of network, the optimal solution of caching placement changes over the time. Hence, one of the research direction of edge caching is to take account of user mobility in caching strategy. In [8], Keshavarzian et al. proposed to model users' mobility as a discrete Markov chain and derived a polynomial time approximation solution for a situation which the information about the mobility pattern of mobile users in the wireless cell does not exist. Based on the periodic property of user mobility pattern, Wang et al. presented specific design examples for mobility-aware caching content placement, including caching at BSs and caching at UTs [9]. However, due to the dense deployment of the SBSs in 5G cellular networks, a travelling user will cross the SBSs more frequently and may not be able to obtain a complete file from one single SBS. A similar situation has been analysed in [10]. A whole file was divided into several equal-size chunks and a strategy was proposed to manage each SBS's caching content, but the design only considered the scenario of one user, with a assumption that the user path was known.

In this paper, we proposed a Proactive Mobility-Aware (PMA) caching strategy specifically adapted to a dynamic environment of dense small cell networks. In our strategy, we firstly predict the mobility of users. Then, based on the content request of users, the macro base station (MBS) proactively sends content to the SBSs which are adjoining to users' future locations. Thus our solution maximizes the cache hit ratio. We evaluate our approaches and compare them with a common caching placement strategy, using a real user mobility traces dataset. The simulation results show our caching strategy achieved a remarkable cache hit ratio improvement.

The rest of the paper is organized as follows. In Sect. 2, we describe our scenario and the problem formulation. Section 3 introduces the proposed PMA caching strategy for the caching problem. Section 4 provides performance evaluations of PMA, and Sect. 5 concludes the paper.

2 System Model and Problem Formulation

In this section, we introduce the system model and formulate the caching placement optimization problem.

2.1 System Model

We consider the dense depolyments of a single MBS with a set of SBSs $\mathcal{C}$ and a set of mobile users $\mathcal{U}$. We define $\mathcal{F}$ as the set of equal-sized popular files. As for each $f \in \mathcal{F}$, we assume the popularity distribution is known a priori.

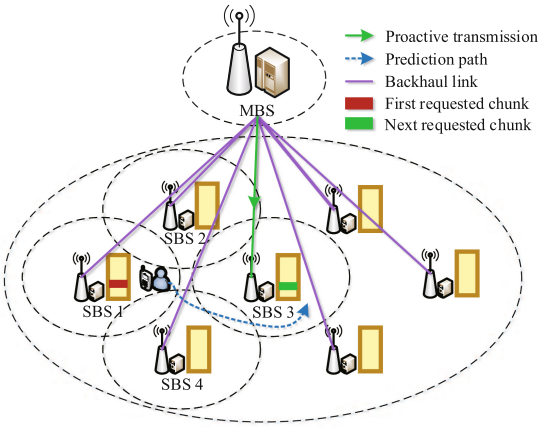


Fig. 1. A sample of cache-enabled SBS deployment with MBS

Futhermore, we use $Z(f)$ to denote the probability that users may request file f. Thus, $\sum_{f \in \mathcal{F}} Z(f) = 1$. As is illustrated in Fig. 1, after predicting the user's path from SBS1 to the SBS 3, the MBS can proactively transmit the next chunk of user request file to the cache storage of SBS 3.

2.2 Mobility Model

For users' mobility prediction, we divide the time into discrete time slots and model the users' movement as a derived mobility Markov chain called n-MMC in [11], which has a higher prediction accuracy than the traditional mobility Markov chain. In n-MMC, the users common position is first clustered into a set of locations $\mathcal{P}$ by DJ-Cluster clustering method. Then, the transition matrix $\mathbf{M}$ of the n-MMC is calculated through analysis of user's trajectory.

We consider a single time slot τ and use $l_u^\tau (l_u^\tau \in \mathcal{P})$ to represent the location of user u in slot τ. Through the user's n previous locations $l_u^\tau, l_u^{\tau-1}, \cdots l_n^{\tau-n+1}$ and the transition matrix $\mathbf{M}$, we can obtain the probability of the user u transferring to location $p (p \in \mathcal{P})$ at slot $\tau + 1$, which is denoted by $\phi_{u,p}^{\tau+1}$.

In dense deployment networks, a few base stations are accessible to the user at the same time. We use a binary variable $\gamma_{p,c}$ to denote the access ability, and $\gamma_{p,c} = 1$ means that location p is in the transmission range of SBS c. The access ability can be obtained through calculating the distance between the location p and SBS c.

2.3 Chunk Request Probability

To facilitate the analysis, we assume the transmission rate for each user is the same. In particular, we consider that each file is composed of k continuous chunks of the same size and one user is able to receive only one chunk within one time

slot. We denote the chunks of file f by $f_1, f_2, \cdots f_k$. To focus on media streaming service, we assume the user always receives successively from f_1 to f_k.

We use a binary variable $R^{\tau}_{u,f,m}$ to indicate the file request at the time of slot τ, where $R^{\tau}_{u,f,m} = 1$ if the chunk f_m has been requested by user u. Besides, we use $\varphi^{\tau+1}_{u,f,m}$ to denote the probability of the user u requesting chunk f_m at slot $\tau + 1$. Therefore, according to the previous hypothesis, if $\exists f' \in \mathcal{F}$ which makes $R^{\tau}_{u,f,m} = 1$, which means that the user u requests the last chunk of file f' within the present slot τ, and then

$$\varphi^{\tau+1}_{u,f,m} = \begin{cases} Z(f) & m = 1 \\ 0 & m \neq 1 \end{cases} \tag{1}$$

and otherwise, if $\exists f' \in \mathcal{F}$ which makes $R^{\tau}_{u,f',m'} = 1 (m' \neq k)$, then

$$\varphi^{\tau+1}_{u,f,m} = \begin{cases} 1 & f = f', m = m' + 1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

2.4 Backhaul and Storage Restrictions

In the deployment of dense network, the service performance of a SBS is limited by insufficient backhaul. We consider each SBS adopts wireless backhaul link, so that the SBSs which are out of service occupy no wireless backhaul resources. For simplicity, we assume the MBS is able to transmit to SBSs B chunks through all backhaul link within one time slot. Once a user's request chunk fails to match the cached chunks of the accessible SBSs of the user, the backhaul link of accessible SBS will be occupied in order for user to acquire files from MBS.

We use $CS(c)(\forall c \in \mathcal{C})$ to denote the cache capacity of each SBS with the size of chunk as the unit. Let a binary variable $\chi^{\tau}_{c,f,m}$ indicate the cache placement, where $\chi^{\tau}_{c,f,m} = 1$ if the chunk f_m has been placed at the cache storage of SBS c in time slot τ. It is easy to see that for $\forall c \in \mathcal{C}$,

$$\sum_{f \in \mathcal{F}} \sum_{m=1}^{k} \chi^{\tau+1}_{c,f,m} \leq CS(c) \tag{3}$$

2.5 Problem Formulation

According to the description above, one unit of backhaul will be occupied for each user if the cache fails. Let Bt^{τ} represent the total backhaul occupied for users whose caches fail, which can be mathematically written as

$$Bt^{\tau} = \sum_{u \in \mathcal{U}} \mathbf{I}(\sum_{c \in \mathcal{C}} \sum_{f \in \mathcal{F}} \sum_{m=1}^{k} R^{\tau}_{u,f,m} \cdot \gamma_{l^{\tau}_u,c} \cdot \chi^{\tau}_{c,f,m} = 0) \tag{4}$$

where $\mathbf{I}$ is the indicator function and is equal to one if the argument be true and zero if be false. Besides, we use Br^{τ} to denote the surplus backhaul which can be used to replace the cache chunks in SBS at slot τ. Then we can express Br^{τ} as

$$Br^{\tau} = \sum_{c \in \mathcal{C}} \sum_{f \in \mathcal{F}} \sum_{m=1}^{k} \chi_{c,f,m}^{\tau} \oplus \chi_{c,f,m}^{\tau+1} \tag{5}$$

where $\chi_{c,f,m}^{\tau} \oplus \chi_{c,f,m}^{\tau+1} = 1$ means the cache state of chunk f_m in SBS c changes between slot τ and $\tau + 1$.

We use $E_u^{\tau+1}(\chi)$ to denote the mathematical expectation of cache hit ratio at slot $\tau + 1$ of user u, so we have

$$E_u^{\tau+1}(\chi) = \sum_{p \in \mathcal{P}} \phi_{u,p}^{\tau+1} \sum_{c \in \mathcal{C}} \gamma_{p,c} \cdot \sum_{f \in \mathcal{F}} \sum_{m=1}^{k} \varphi_{u,f,m}^{\tau+1} \cdot \chi_{c,f,m}^{\tau+1} \tag{6}$$

In order to denote the cache hit ratio of all users, we define a value $E_{total}^{\tau+1}(\chi)$ which is the sum of $E_u^{\tau+1}(\chi)$ of all users. then we have

$$E_{total}^{\tau+1}(\chi) = \sum_{u \in \mathcal{U}} E_u^{\tau+1}(\chi) = \sum_{u \in \mathcal{U}} \sum_{p \in \mathcal{P}} \phi_{u,p}^{\tau+1} \sum_{c \in \mathcal{C}} \gamma_{p,c} \cdot \sum_{f \in \mathcal{F}} \sum_{m=1}^{k} \varphi_{u,f,m}^{\tau+1} \cdot \chi_{c,f,m}^{\tau+1} \tag{7}$$

It is obvious that $E_{total}^{\tau+1}(\chi)$ is proportional to the cache hit ratio of all users at slot $\tau + 1$, so the caching placement optimization problem is expressed as

$$\text{Maximize} \quad E_{total}^{\tau+1}(\chi) \tag{8}$$

Subject to

$$\sum_{f \in \mathcal{F}} \sum_{m=1}^{k} \chi_{c,f,m}^{\tau+1} \leq CS(c) \qquad \forall c \in \mathcal{C} \tag{9}$$

$$Br^{\tau} + Bt^{\tau} \leq B \tag{10}$$

3 Proactive Mobility-Aware (PMA) Caching

In this section, we first introduce the concept of the cache content weight at each SBS. Then we discuss our strategy which solves the caching placement problem in a two-step procedure.

3.1 SBS Content Weight

It is not hard to find that $\gamma_{p,c}$ is a constant in the Eq. (6). Therefore, to facilitate the description, we denote the probability that user u is able to reach SBS c at slot $\tau + 1$ by $\psi_{u,c}^{\tau+1}$. The probability can be obtained according to $\phi_{u,p}^{\tau+1}$ and $\gamma_{p,c}$, hence we have

$$\psi_{u,c}^{\tau+1} = \sum_{p \in \mathcal{P}} \phi_{u,p}^{\tau+1} \cdot \gamma_{p,c} \tag{11}$$

Therefore, (8)–(10) can be rewritten as

$$\text{Maximize} \quad \mathrm{E}_{\mathrm{total}}^{\tau+1}(\chi) = \sum_{\mathrm{u}\in\mathcal{U}} \sum_{\mathrm{c}\in\mathcal{C}} \sum_{\mathrm{f}\in\mathcal{F}} \sum_{\mathrm{m}=1}^{\mathrm{k}} \psi_{\mathrm{u,c}}^{\tau+1} \cdot \varphi_{\mathrm{u,f,m}}^{\tau+1} \cdot \chi_{\mathrm{c,f,m}}^{\tau+1} \tag{12}$$

Subject to

$$\sum_{f\in\mathcal{F}} \sum_{m=1}^{k} \chi_{c,f,m}^{\tau+1} \leq CS(c) \qquad \forall c \in \mathcal{C} \tag{13}$$

$$Br^{\tau} + Bt^{\tau} \leq B \tag{14}$$

It is obvious that for a certain chunk f_m, if the cache of SBS c is increased by f_m, i.e. $\chi_{c,f,m}^{\tau} = 0, \chi_{c,f,m}^{\tau+1} = 1$, then $E_{total}^{\tau+1}(\chi)$ will be increased by $\sum_{u\in\mathcal{U}} \psi_{u,c}^{\tau+1} \cdot \varphi_{u,f,m}^{\tau+1}$, and if the cache of SBS c is decreased by f_m, i.e. $\chi_{c,f,m}^{\tau} = 1, \chi_{c,f,m}^{\tau+1} = 0$, then $E_{total}^{\tau+1}(\chi)$ will be decreased by $\sum_{u\in\mathcal{U}} \psi_{u,c}^{\tau+1} \cdot \varphi_{u,f,m}^{\tau+1}$. Therefore, we introduce the concept of content weight and we denote the content weight by a Matrix $\mathbf{\Omega}_c^{\tau+1}$. Specifically, $\mathbf{\Omega}_c^{\tau+1}(f,m)$ denotes the content weight of chunk f_m of SBS c, then we have

$$\mathbf{\Omega}_c^{\tau+1}(f,m) = \sum_{u\in\mathcal{U}} \psi_{u,c}^{\tau+1} \cdot \varphi_{u,f,m}^{\tau+1} \tag{15}$$

Hense, each SBS has a content weight matrix $\mathbf{\Omega}^{\tau+1}$, and our goal is to maximize the sum of content weights of all cached chunks in all SBSs.

3.2 Proactive Mobility-Aware Caching Strategy

Our caching strategy consists of two steps. The first step is to consider the requested chunks which the SBS cache fails to hit in present slot τ. These chunks will be transmitted to the SBSs that are accessible to the cache-failed users through backhaul link. Meanwhile, these users will obtain these chunks from these SBSs. $E_{total}^{\tau+1}(\chi)$ can be optimized by deciding whether to store these transferred chunks or not and if store, the cache of which SBS will be used.

First, we calculate the content weight matrix $\mathbf{\Omega}^{\tau+1}$ of all chunks of all files in each SBS. Then we find the set of all users whose requests the cache fails to hit and denoted as $\mathcal{U}_{fail}^{\tau}$, so we have

$$\mathcal{U}_{fail}^{\tau} = \left\{ u \left| \sum_{c\in\mathcal{C}} \sum_{f\in\mathcal{F}} \sum_{m=1}^{k} R_{u,f,m}^{\tau} \cdot \gamma_{l_u^{\tau},c} \cdot \chi_{c,f,m}^{\tau} = 0, u \in \mathcal{U} \right. \right\} \tag{16}$$

For each user u, $\mathcal{C}_u^{\tau}$ is used to denote the set of all base stations that are accessible to the user at present. Then we have

$$\mathcal{C}_u^{\tau} = \left\{ c \left| \gamma_{l_u^{\tau},c} = 1, c \in \mathcal{C} \right. \right\} \tag{17}$$

The weight of currently requested chunk and the minimum weight of all chunks that have been stored in cache of SBS c is found, with the weight difference calculated. The user is guided to choose the SBS whose weight difference is the maximum, and this requested chunk will replace the chunk whose weight is the minimum.

Algorithm 1. First step of PMA caching strategy

Require:
a)A set of cache fail users $\mathcal{U}_{fail}^{\tau}$
b)All sets of reachable SBSs of user locations $\mathcal{C}_u^{\tau}(u \in \mathcal{U})$
c)All the content weight matrix $\mathbf{\Omega}_c^{\tau+1}(c \in \mathcal{C})$;
for each $u \in \mathcal{U}_{fail}^{\tau}$ **do**
  Read the user request f_m
  for each $c \in \mathcal{C}_u^{\tau}$ **do**
    Find the minimal-weight cached chunk f_m
    Calculate $\Delta_c = \mathbf{\Omega}_c^{\tau+1}(f, m) - \mathbf{\Omega}_c^{\tau+1}(f', m')$
  end for
  Find the maximal weight difference $\Delta_{\max}$ and the SBS c'
  if $\Delta_{\max} > 0$ **then**
    Replace the minimal-weight cached chunk with f_m in c'
  end if
end for
return

The second step is to optimize cache content from each SBS with the surplus backhaul Br^{τ}.

Algorithm 2. Second step of PMA caching strategy

Require:
a)Residual backhaul Br^{τ}
b)All content weight matrix $\mathbf{\Omega}_c^{\tau+1}(\forall c \in \mathcal{C})$
for each $c \in \mathcal{C}$ and $\mathbf{\Omega}^{\tau+1}$ **do**
  Find the minimal-weight cached chunk $f'_{m'}$
  Find the maximal-weight uncached chunk f_m
  Calculate the weight difference $\lambda_c = \mathbf{\Omega}_c^{\tau+1}(f, m) - \mathbf{\Omega}_c^{\tau+1}(f', m')$
end for
while $Br^{\tau} > 0$ **do**
  Find the maximal weight difference $\lambda_{\max}$ and the SBS c'
  Replace the minimal-weight cached chunk with the maximal-weight uncached chunk in c'
  Fresh the weight difference $\lambda_{c'}$ of c'
  Reduce the residual backhaul $Br^{\tau} \leftarrow Br^{\tau} - 1$
end while
return

In this step, for each SBS c, the difference between the maximum weight of chunks that are not cached and the minimum weight of chunks that are cached in $\mathbf{\Omega}_c^{\tau+1}$ is calculated and denoted as λ_c. Then, the SBS c' with the maximum $\lambda_{c'}$ is known to us. The min-weight chunk is then replaced by the max-weight chunk. After the replacement, $\lambda_{c'}$ is updated. The replacement will not halt until Br^{τ} is exhausted.

By performing these two-step algorithms in turn, the cache placement will get the best optimization.

4 Performance Evaluation

4.1 Scenario Design

In this section, we evaluate the performance of proposed caching strategy. We use the human mobility data collected by GPS positioning devices from five different sites [12]. We choose the KAIST mobility pattern data set in this project, in which 92 users' movement trace data is recorded in the KAIST campus during one day. The time interval between two consecutive sampling point is 30 s. Therefore, we set the time slots length equal to 30 s and set the simulation duration equal to 1000 time slots. In order to give consideration to both the accuracy and complexity, we use 2-MMC to predict the mobility of users. In the aspect of small cell deployment, we generate the SBSs with a equidistant arrangement on the map of the KAIST campus. Refer to the description of outdoor deployment scenario in dense network [13], we set the distance of two adjoining SBSs to 40 m and the coverage of each SBS to 100 m. After considered the expected low cache device cost of the SBSs in dense networks, we set the caching storage capacity of each SBS to the size of one file, that is, the size of k chunk. For User's preference for contents, we assumed that the file popularity follows a Zipf distribution with a Zipf parameter α. We set the total number of files in the file library equal to 100.

4.2 Simulation Results

In order to assess the efficiency of our cache strategy, we compare the cache hit ratio performance of our caching strategy with the MPV strategy, for which each SBS stores the most popular contents [14]. In brief, the MPV strategy retrieves popularity information by keeping track of the number of requests to a content and evicts the least frequently requested contents.

Figure 2 depicts the cache hit ratio as a function of the Zipf parameter α for the PMA and the MPV caching strategies. As expected, larger value of the Zipf parameter α implies the requests from mobile users are more concentrated on the popular files, which makes the cache hit ratio of both caching schemes higher. Thanks to the proactive caching strategy based on user mobility prediction, we see that PMA outperforms MPV by 40% up to 90%.

Figure 3 shows the performance gain provided by PMA over MPV in different backhaul bandwidth. In order to better show the relationship between bandwidth

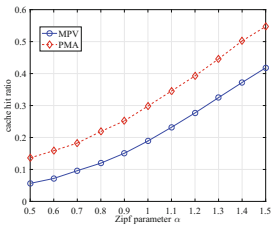


Fig. 2. Comparison of PMA and MPV: cache hit ratio vs Zipf parameter α; $B = 150, k = 3$.

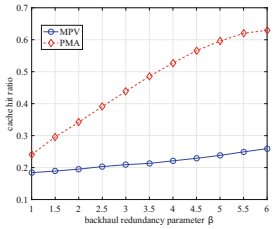


Fig. 3. Comparison of PMA and MPV: cache hit ratio vs Backhaul redundancy parameter β; $\alpha = 1, k = 3$.

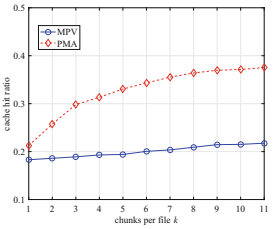


Fig. 4. Comparison of PMA and MPV: cache hit ratio vs Serving file size k; $B = 150, \alpha = 1$.

and performance, we define a backhaul redundancy parameter β, which is the ratio of the total backhaul and the number of mobile users. We observe that the performance of PMA is close to MPV when $\beta = 1$, which means the total backhaul equal to the total user access bandwidth. As β increases, the cache hit ratio is improved markedly. This is due to the way in which PMA effectively handle the remaining backhaul for proactive caching. However, with the rise of β, the growth of the cache hit ratio become slower, which indicates the excessive backhaul is less helpful in cache hit ratio improvement.

The influence of the size of a single popular file is shown in Fig. 4, where we plot the cache hit ratio against the number of chunks of a single file. As the figure shows, when the size of a serving file becomes larger, the performance of PMA caching strategy has also been improved, which means our proposed strategy runs better in the case of larger files.

5 Conclusion

In this paper, we have studied efficient edge caching in dense deployment scene as the users are moving between base stations. To solve this problem, we proposed a proactive mobility-aware (PMA) caching strategy. By predicting each user's future movement path using a n-MMC algorithm presented in [11], we adjust caching content placement in SBSs to fit each user's cache needs in the next moment. We have tested the PMA caching strategy with real user mobility pattern data in the campus of KAIST from the CRAWDAD dataset. The simulation result shows that the PMA strategy has a significant cache hit ratio improvement in different simulation conditions, with respect to MPV strategy in [14].

Acknowledgement. This paper is sponsored by the National Natural Science Foundation of China for the Youth (61501047) and Fundamental Research Funds for the

Central Universities of China (2017RC04), and the National Science and Technology Major Project of the Ministry of Science and Technology of China (2016ZX03001017).

References

1. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html
2. Boccardi, F., Heath, R.W., Lozano, A., Marzetta, T.L., Popovski, P.: Five disruptive technology directions for 5G. IEEE Commun. Mag. **52**(2), 74–80 (2014)
3. Bastug, E., Bennis, M., Debbah, M.: Living on the edge: the role of proactive caching in 5G wireless networks. IEEE Commun. Mag. **52**(8), 82–89 (2014)
4. Golrezaei, N., Molisch, A.F., Dimakis, A.G., Caire, G.: Femtocaching and device-to-device collaboration: a new architecture for wireless video distribution. IEEE Commun. Mag. **51**(4), 142–149 (2013)
5. Blaszczyszyn, B., Giovanidis, A.: Optimal geographic caching in cellular networks. In: 2015 IEEE International Conference on Communications (ICC), pp. 3358–3363, June 2015
6. Hamidouche, K., Saad, W., Debbah, M., Poor, H.V.: Mean-field games for distributed caching in ultra-dense small cell networks. In: 2016 American Control Conference (ACC), pp. 4699–4704, July 2016
7. Gabry, F., Bioglio, V., Land, I.: On energy-efficient edge caching in heterogeneous networks. IEEE J. Sel. Areas Commun. **34**(12), 3288–3298 (2016)
8. Keshavarzian, I., Zeinalpour-Yazdi, Z., Tadaion, A.: A clustered caching placement in heterogeneous small cell networks with user mobility. In: 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 421–426, December 2015
9. Wang, R., Peng, X., Zhang, J., Letaief, K.B.: Mobility-aware caching for content-centric wireless networks: modeling and methodology. IEEE Commun. Mag. **54**(8), 77–83 (2016)
10. Mahmood, A., Casetti, C., Chiasserini, C.F., Giaccone, P., Harri, J.: Mobility-aware edge caching for connected cars. In: 2016 12th Annual Conference on Wireless On-demand Network Systems and Services (WONS), pp. 1–8, January 2016
11. Gambs, S., Killijian, M.-O., Del Prado Cortez, M.N.: Next place prediction using mobility Markov chains. In: ACM Special Interest Group on Operating Systems (SIGOPS), Bern, Switzerland (2012)
12. Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S., Chong, S.: CRAWDAD dataset ncsu/mobilitymodels (v. 2009-07-23), July 2009. http://crawdad.org/ncsu/mobilitymodels/20090723
13. Kamel, M., Hamouda, W., Youssef, A.: Ultra-dense networks: a survey. IEEE Commun. Surv. Tutor. **18**(4), 2522–2545 (2016). Fourthquarter
14. Ahlehagh, H., Dey, S.: Video-aware scheduling and caching in the radio access network. IEEE/ACM Trans. Netw. **22**(5), 1444–1462 (2014)

The Minimum of GDOP for Satellite Navigation System

Xuetao Yang, Jianchao Du(✉), and Weibin Li

State Key Laboratory of ISN, Xidian University, Xi'an 710071, China
jcdu@xidian.edu.cn

Abstract. Geometric Dilution of Precision (GDOP) represents the magnification positioning error in satellite navigation, smaller GDOP represents high positioning accuracy. The analytic results of relevant methods to derive the minimum of GDOP by geocentric coordinate system cannot be achieved in reality, because it does not think fully about the satellite visibility constrains. This article employs the elevation angle to express the satellite visibility, and on basis of the local Cartesian coordinates coordinate system, the essay derives the minimum of GDOP with constraint of four satellites visible. Based on the information track of Beidou satellite navigation system, the minimum of GDOP is verified by using the STK simulation software.

Keywords: Satellite navigation · Geometric Dilution of Precision (GDOP)
Local cartesian coordinates coordinate system · Satellite elevation
Minimum

1 Introduction

In satellite navigation system, the Geometric of Precision (GDOP) is one of the indicators reflecting the positioning accuracy [1]. Under the same measurement error, the less the GDOP value, the higher the positioning accuracy will be [2]. The literature [3, 4] deduced that the theoretical minimum value of GDOP is $\sqrt{\frac{10}{n}}$ from a mathematical point of view. However, this minimum can not be achieved in reality. Literature [5] analyzes the minimum of GDOP in combination satellite navigation system is $\sqrt{\frac{8}{3}}$, but it has the same matter with literature [3, 4].

This paper transforms geocentric coordinate system into local Cartesian coordinates coordinate system, using the related properties of the matrix deduced the expression of satellite elevation, obtained the easily controlled satellite visibility constraint condition and found out the minimum value of GDOP at different elevation limits when there are four satellites is visible. According to the real-time acquisition of the BeiDou satellite TLE (Two-Line Element) and orbit information, this paper uses the Satellite Tool Kit (STK) to simulate and analyze the minimum value of GDOP at several major cities in the Asia Pacific region when the Beidou is on-orbit. The simulation results are in good agreement with the theoretical derivative results, which verifies the correctness of the derived results in this paper.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 443–447, 2018.
https://doi.org/10.1007/978-3-319-78130-3_46

2 Analysis of GDOP in Local Cartesian Coordinates Coordinate System

2.1 Matrix Transformation of the Minimum of GDOP

The positioning of the satellite navigation system requires knowledge of the three coordinate components (x, y, z) and the receiver clock error δ_t. So the satellite system to achieve positioning need at least four visible satellites for positioning. The coefficient matrix H of navigation positioning is often called observation matrix, as shown in Eq. (1) [6].

$$H = \begin{bmatrix} I_{x1} & I_{y1} & I_{z1} & 1 \\ I_{x2} & I_{y2} & I_{z2} & 1 \\ \cdots & \cdots & \cdots & \cdots \\ I_{xn} & I_{yn} & I_{zn} & 1 \end{bmatrix} \tag{1}$$

I_{xi}, I_{yi}, I_{zi} respectively represent the direction vector cosine between the receiver and the visible satellite. N is the number of visible satellites, satisfy $|I|_i = 1$, which is $I_i(i = 1, 2, \ldots, n)$ are located on the unit sphere. The Geometric Dilution of Precision is defined as Eq. (2).

$$GDOP = \sqrt{tr(H^T H)^{-1}} \tag{2}$$

$tr(H^T H)^{-1}$ express $(H^T H)^{-1}$ trace. $(H^T H)^{-1}$ denotes the weight coefficient matrix of the positioning error covariance matrix. For a very high order or a special structure matrix, the large matrix is classified as a small matrix by block method, so that the correlation calculation of the matrix is simplified. According to the idea of partitioning matrix, the matrix $H^T H$ in Eq. (2) can be divided as:

$$H^T H = \begin{bmatrix} C & U \\ U^T & D \end{bmatrix} \tag{3}$$

From the inverse method of the block matrix, we have:

$$(H^T H)^{-1} = \begin{bmatrix} (C - UD^{-1}U^T)^{-1} & -(C - UD^{-1}U^T)^{-1}UD^{-1} \\ -(D - U^T C^{-1}U)^{-1}U^T C^{-1} & (D - U^T C^{-1}U)^{-1} \end{bmatrix} \tag{4}$$

From the definition of the block matrix:

$$GDOP = \sqrt{tr(C - UD^{-1}U^T)^{-1} + tr(D - U^T C^{-1}U)^{-1}} \geq \sqrt{tr(C)^{-1} + tr(D)^{-1}} \tag{5}$$

H is a column full rank matrix, so $H^T H$ is a positive definite matrix. C, D is also. $UD^{-1}U^T$, $U^T C^{-1}U$ are all positive semidefinite matrices. The modulus is greater than or equal to 0. If and only if U = 0 can the equality of the above equation, that is:

$$\sum_{i=1}^{n} I_{xi}I_{zi} = \sum_{i=1}^{n} I_{xi} = \sum_{i=1}^{n} I_{yi}I_{zi} = \sum_{i=1}^{n} I_{yi} = 0 \tag{6}$$

At the same time, the minimum GDOP is obtained:

$$GDOP_{\min} = \sqrt{tr(C)^{-1} + \mathrm{tr}(D)^{-1}} \tag{7}$$

The minimum of GDOP is related to the vector cosine of the x, y, and z axes, so the problem of solving the minimum value of GDOP is transformed into finding (I_{xi}, I_{yi}, I_{zi})i $= 1, 2, \ldots$n, to satisfy the condition.

2.2 Detailed Expression of the Minimum of GDOP

Satellite elevation angle is greater than 0 when the receiver is on the Earth's surface. But the satellite positioning equation uses geocentric coordinate system. To the heart of O as the origin of coordinates, Z axis points to the Earth's Arctic, X axis points to an intersection of the reference meridian plane with the Earth's equator, and the x, y and z axes together form the right hand Cartesian coordinate system. A only contains the direction vector cosine of the x, y, z axis of the satellite at the receiver, but no satellite elevation. And local Cartesian coordinates coordinate system will be able to calculate the satellite elevation.

Local Cartesian coordinates coordinate system takes the user's location as the origin of the coordinates. The three axes are perpendicular to each other: ENU. According to coordinate transformation relationship, the parameters x, y, z of the geocentric solid coordinate system are transformed into the satellite elevation angle θ and azimuth angle ϕ. According to the rules, the unit direction vector cosine I_{xi}, I_{yi}, I_{zi} in Eq. (1) can be converted into $\cos\theta_n \cos\phi_n$, $\cos\theta_n \sin\phi_n$, $\sin\phi_n$. When n = 4, Eq. (6) can be described as:

$$\begin{cases} \cos\theta_1 \cos\phi_1 + \cos\theta_2 \cos\phi_2 + \cos\theta_3 \cos\phi_3 + \\ \cos\theta_4 \cos\phi_4 = 0 \quad \ldots\ldots f(1) \\ \cos\theta_1 \sin\phi_1 + \cos\theta_2 \sin\phi_2 + \cos\theta_3 \sin\phi_3 + \\ \cos\theta_4 \sin\phi_4 = 0 \quad \ldots\ldots f(2) \\ \cos\theta_1 \cos\phi_1 \sin\theta + \cos\theta_2 \cos\phi_2 \sin\theta + \\ \cos\theta_3 \cos\phi_3 \sin\theta + \cos\theta_4 \cos\phi_4 \sin\theta = 0 \ldots f(3) \\ \cos\theta_1 \sin\phi_1 \sin\theta + \cos\theta_2 \sin\phi_2 \sin\theta + \\ \cos\theta_3 \sin\phi_3 \sin\theta + \cos\theta_4 \sin\phi_4 \sin\theta = 0 \quad \ldots f(4) \end{cases} \tag{8}$$

Equation (8) shows that the vector sum of the y-axis and the x-axis in the visible satellite is zero, we have:

$$\begin{cases} \cos\theta_1 \cos\phi_1 + \cos\theta_2 \cos\phi_2 = \cos\theta_3 \cos\phi_3 + \cos\theta_4 \cos\phi_4 = 0 \quad \ldots f(5) \\ \cos\theta_1 \sin\phi_1 + \cos\theta_2 \sin\phi_2 = \cos\theta_3 \sin\phi_3 + \cos\theta_4 \sin\phi_4 = 0 \quad \ldots f(6) \end{cases} \tag{9}$$

$f(6)/f(5)$:

$$\begin{cases} \tan\phi_1 = \tan\phi_2 \\ \tan\phi_3 = \tan\phi_4 \end{cases} \tag{10}$$

The volume of the polyhedron with the unit length of the visible satellite and the receiver as the apex is inversely proportional to the GDOP value. So the smaller GDOP is satisfied:

$$\begin{cases} \cos\phi_1 \cos\phi_2 < 0 \\ \cos\phi_3 \cos\phi_4 < 0 \end{cases} \tag{11}$$

So $\phi_1 = \phi_2 + \pi$, $\phi_3 = \phi_4 + \pi$, substituting into Eq. (9). Set $\theta_1 = \theta_2$, $\theta_3 = \theta_4$. In this paper, we take a special solution of azimuth angle: $\phi_1 = 0$, $\phi_2 = \pi$, $\phi_3 = \frac{\pi}{2}$, $\phi_4 = \frac{3\pi}{2}$. Use φ_1 be the elevation angle of θ_1 and θ_2, and φ_2 be the elevation angle of θ_3 and θ_4. Substituted into Eq. (11) leads to:

$$GDOP_{\min} = \sqrt{\frac{1}{2\cos^2\varphi_1} + \frac{1}{2\cos^2\varphi_2} + \frac{\sin^2\varphi_1 + \sin^2\varphi_2 + 2}{2(\sin\varphi_1 - \sin\varphi_2)^2}} \tag{12}$$

Equation (12) shows the relationship between the elevation angle of the satellite and the minimum value of GDOP. Calculated by MATLAB can get the minimum of GDOP is 2.1201 when φ_1 is 7.83°, φ_2 is 82.53°. 2.1201 is 34.09% higher than the minimum value of GDOP obtained by the traditional calculation method.

3 Numerical Experiments

Satellites are gradually invisible with the elevation threshold increasing. The minimum value of GDOP varies with the satellite elevation threshold. When the satellite elevation range is smaller, the larger the GDOP value, and the lower the positioning accuracy. As the Beidou satellite system is currently mainly covering the Asia-Pacific region, this paper chooses several major cities in the Asia-Pacific region: Beijing, Chengdu, Singapore, Tokyo, Urumqi, Xi'an, analyzing the minimum value of GDOP in these cities changed with satellite elevation threshold φ in Table 1.

Table 1. The minimum value of GDOP in Asia-Pacific changed with φ

City	5°	10°	15°	20°
Beijing	2.249	3.341	3.594	6.101
Chengdu	3.840	4.199	4.281	9.967
Singapore	2.220	2.435	2.664	3.061
Tokyo	2.916	2.461	2.729	3.312
Urumqi	2.442	3.938	4.006	9.684
Xi'an	3.471	3.850	3.929	6.752

It can be seen in Table 1 that Beijing, Tokyo, Singapore with the satellite elevation threshold changes smaller, and it has a high positioning accuracy. The GDOP value of Chengdu, Urumqi, Xi'an and other areas is higher than Beijing and Tokyo. The minimum of GDOP increases slowly when the elevation threshold is less than 15°. The minimum of GDOP increases rapidly when the elevation threshold is higher than 15°. It main due to Beidou system is in the construction.

4 Conclusions

According to the orbit information of Beidou satellite navigation system, the minimum of GDOP together with the design matrix for four visible satellites have been derived theoretically by STK software in this paper. The minimum of GDOP is increasing with the increase of satellite elevation threshold, and greatly affected by it. The value of GDOP at some area is still large mainly due to the number of Compass satellite on orbit. The problem of the GDOP value in some areas is large can be solved after the completion global coverage of Beidou in 2020. At the same time, the combination of satellite navigation system is also one of the solutions to solve this problem, which is the following work need to focus on.

Acknowledgments. This work was supported by The Ningbo Natural Science Foundation (2014A610151).

References

1. Teng, Y.L., Wang, J.L.: A closed-form formula to calculate geometric dilution of precision (GDOP) for multi-GNSS constellations. GPS Solut. **20**(3), 331–339 (2016)
2. Teng, Y.L., Wang, J.L., Huang, Q.: Minimum of Geometric Dilution of Precision (GDOP) for five satellites with dual-GNSS constellations. Adv. Space Res. **56**(2), 229–236 (2015)
3. Xie, G.: Principles of GPS and Receiver Design, pp. 96–117. Publishing House of Electronics Industry, Beijing (2009)
4. Li, J.W., Li, Z.H., Zhou, W., Si, S.J.: Study on the minimum of GDOP in satellite navigation and its applications. Acta Geodaetica et Cartographica Sinica **40**(1), 85–88 (2011)
5. Teng, Y.L., Wang, J.L., Huang, Q.: Mathematical minimum of geometric dilution of precision (GDOP) for dual-GNSS constellations. Adv. Space Res. **57**(1), 183–188 (2016)
6. Xue, S.Q., Yang, Y.X.: Positioning configurations with the lowest GDOP and their classification. J. Geodesy **89**(1), 49–71 (2015)



A SDN Security Transmission Service Using Multi-connection and Time-Slot Scheduling

Shiqi Zhou[1,2(✉)], Rongheng Lin[1,2], Shuquan Fan[1,2], Hua Zou[1], and Budan Wu[1]

[1] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
Shiqi_zhou@126.com

[2] Science and Technology on Information Transmission and Dissemination Networks Laboratory, The 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang 050081, China

Abstract. SDN (Software-Defined Networking) provides a new approach to improve network security. In this paper, we design and implement a security service of transmission channels in SDN, which will efficiently lower the percentage of information leakage on the data plane of SDN networks. The core of the service is forwarding strategies based on multi-connection and time-slot scheduling, which functions on SDN controllers. According to the time-slot strategy, every data flow is transmitted on one path in a network at a certain moment, then in next time-slot it will be transmitted on another path, and so on. The multi-connection strategy allows to transmit data through multiple links at a same time. There is a client-side subsystem deployed in user terminals to process the data to send or receive. The processing of data including splitting and encrypting in application layer, coordinating with the forwarding strategies, reduces probability of the whole data flow being intercepted.

Keywords: SDN · Security · Multi-connection · Time-slot scheduling

1 Introduction

Security is always a focus of concern during all periods of networking development. Under the condition that some of switches in a network are under attack, attackers can easily intercept all information in the packets forwarded by the switches. SDN, an emerging network architecture, which decouples the network control and forwarding functions enabling the network control to become directly programmable and the underlying infrastructure to be abstracted for applications and network services [1], provides a new approach to deal with the situation and improve network security.

SDN controllers can change transmission path of data flow by distributing flow tables to OpenFlow switches, so we propose multi-connection and time-slot forwarding strategy, which functions on SDN controllers, to prevent all packets of a certain data flow from being intercepted. The SDN controller prepares multiple optional transmitting paths, and every data flow is transmitted on one path in the network at a certain moment. The time-slot strategy will transfer the data flow to another transmitting path

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 448–459, 2018.
https://doi.org/10.1007/978-3-319-78130-3_47

every time-slot, and by this analogy, packets of a data flow are forwarded by different switches in different time-slot. The multi-connection strategy allows to transmit data through multiple links at a same time. Every link adopts the time-slot strategy and does not interfere each other. The strategy secures the safety of the network and improves the efficiency of transmission.

The multi-connection and time-slot scheduling works on SDN controllers, and implement function on the OpenFlow switches, so it is transparent to users who send or receive data in the network. The client-side subsystem deploys in user terminals to process the data to send or received, and provides users with several splitting and encrypting choices to protect their data.

2 Related Work

SDN (Software Defined Networking) changes traditional way to design and manage networks. It has two defining characteristics: separation of the control plane and the data plane and that the control plane has an overall view of networks [2]. The SDN architecture provides capabilities of security monitoring, analysis and response, which can be used to enhance network security. The SDN controller can detect network status information and security-related data. Based on the analysis of these data, new security policy can be applied to the network in the form of flow rules. This approach can efficiently detect and contain network security threats.

In this paper we discuss data leakage which is a security challenge only involving the data layer of the network. For example, OpenFlow switches are under attack and flow rules expose, or packets forwarded by the switch are stolen. A number of proposals derived from multipath and time division have been made to solve this security threat and enhance network security [3].

In [4], time-based updates in SDN is put forward. The authors came to the conclusion that a time-based updating procedure is necessary to solve the packets loss problem during flow allocation [5]. There is another research proposed a method to improve QoS on SDN networks by scheduling packets forwarding [6]. It adopted routing strategy to implement packet scheduling based on network state.

3 Application Scenarios

In a SDN network, a user on a terminal host requires security transmission service to transmit files to another terminal host. As shown in Fig. 1, when the user with a certain security level need to send data, client-side subsystem deployed in the user's terminal sends a request to the management module to declaim itself to the system and gets respond of specified security policy corresponding to user's security level, which decides how to split and encrypt the data. Received the request, the management module will start the security transmission service on the SDN controllers, which acknowledge status of all switches in the network, calculate candidate paths, and distribute flow table entries to switches to implement multi-connection and time-slot scheduling.

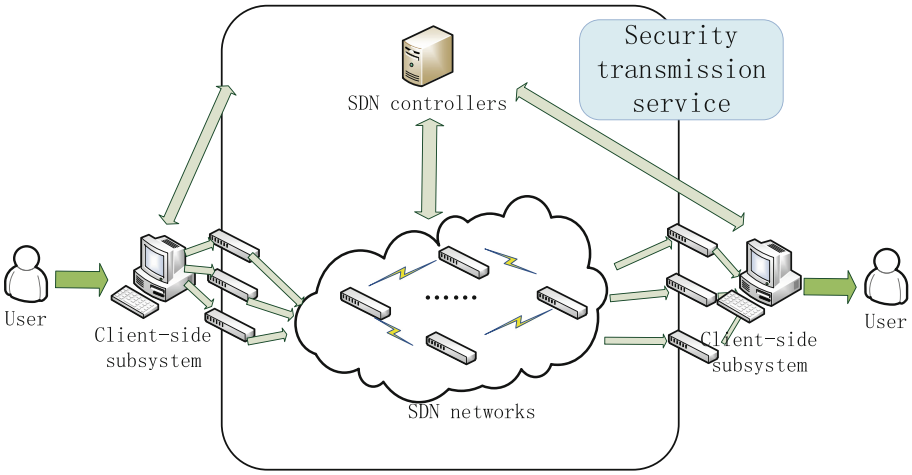


Fig. 1. Security transmission service deployed in a SDN network

As mentioned above, the security transmission service need to be deployed on both terminal hosts and web-side. On terminal hosts, the client-side subsystem processes data that users send or receive. On the web-side, SDN controllers are deployed on a server with the web-side security manager which provides the security transmission service. The switches in the bottom layer of the network should support OpenFlow Protocol.

In the SDN network, switches forwarding packets depend on their flow tables with multiple flow table entries and every data flow corresponds to a flow table entry. So multiple users can use the security transmission service at the same time and do not interfere with each other. Also, traditional forwarding (not based on OpenFlow protocol) is supported and not influenced by the security transmission service.

4 General Architecture

4.1 General Architecture

The security transmission service bases on SDN. SDN is an emerging architecture which decouples the network control and forwarding functions enabling the network control to become directly programmable and the underlying infrastructure to be abstracted for applications and network services. The SDN architecture is composed of application layer, control layer and infrastructure layer. As network intelligence is centralized in software-based SDN controllers and switches in the infrastructure layer are only in charge of forwarding, we can design modules cooperating with SDN controllers to provide security transmission service.

The security transmission service is provided by a client-side subsystem and a web-side security manager (as Fig. 2). In the web-side security manager, the forwarding layer is infrastructure layer of SDN, the controlling layer is based on SDN

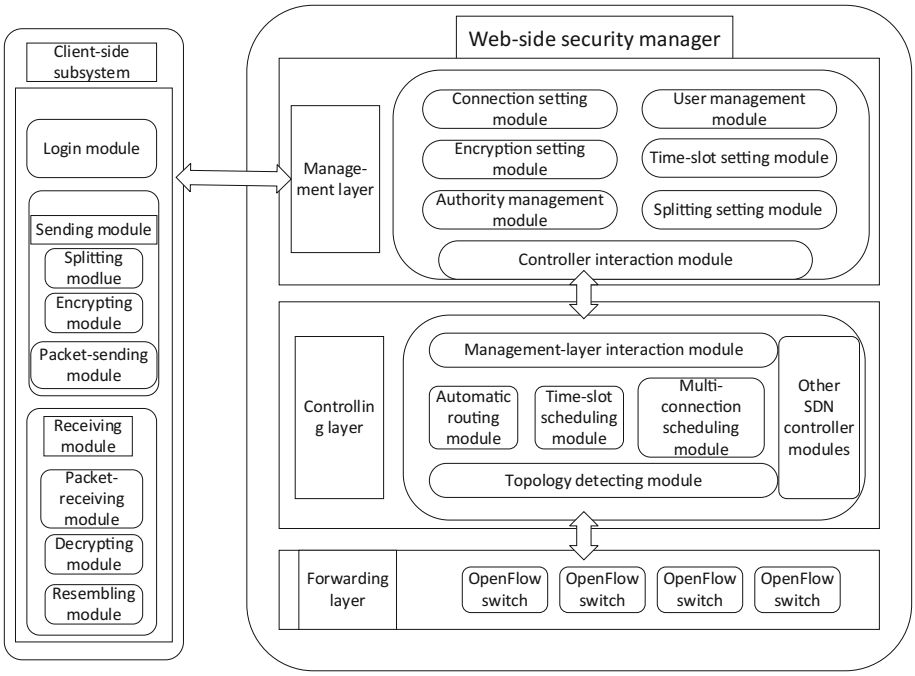


Fig. 2. General architecture of the security transmission service

controller layer and consisted of SDN controllers and modules to implement strategies, and the management layer uses API provided by SDN controller layer and practices as application layer.

4.2 The Client-Side Subsystem

The client-side subsystem processes users' data and sends them to multiple ports. As the left part in Fig. 2, it contains a login module, a sending module and a receiving module. The login module register user's identity information on the web-side manager and according to the security level of the user the manager adopts different scheduling strategy. The sending module has three submodules as splitting, encrypting and packet-sending, and correspondingly the receiving module is composed by assembling, decrypting and packet-receiving.

4.2.1 Splitting and Assembling

The splitting module gets splitting setting parameters from the web-side manager and splits data to be sent into requested fragments. There are multiple choice of splitting strategies. For example, the simplest way is to split the file into equal length. And we can also split the file according to its type. Take a case of a text file which can be extracted into codes or arrays based on its content. The splitting process supports to send one file on multiple connections and also is a form of encryption itself.

Destination host needs splitting setting parameters from the web-side manager to assemble data fragments into the original data file. Fragments arrive at a destination port in sequence, so the receiving module deals with them from multiple ports in sequence.

4.2.2 Encrypting and Decrypting

Data fragments are encrypted in the encrypting module. Encryption algorithm is optional. By default, we adopt a simple and effective bit-transformation which keeps data size the same. Encryption strategy depends on users' security level and receiver will be notified the corresponding decryption strategy by web-side manager and then decrypt data fragments.

4.2.3 Information Flow of the Client-Side Subsystem

Figure 3 demonstrates information flow of the client-side subsystem. When a user sends data, splitting module of the sending module on the sender requests splitting setting parameters from splitting setting module on the web-side manager. The manager replies the sender splitting and encrypting configuration and also send it to the receiving module on destination host. Split data following instruction from the manager into fragments and encrypt them and send to the network. Having received data fragments, receiving module decrypts and assembles them in sequence of receiving.

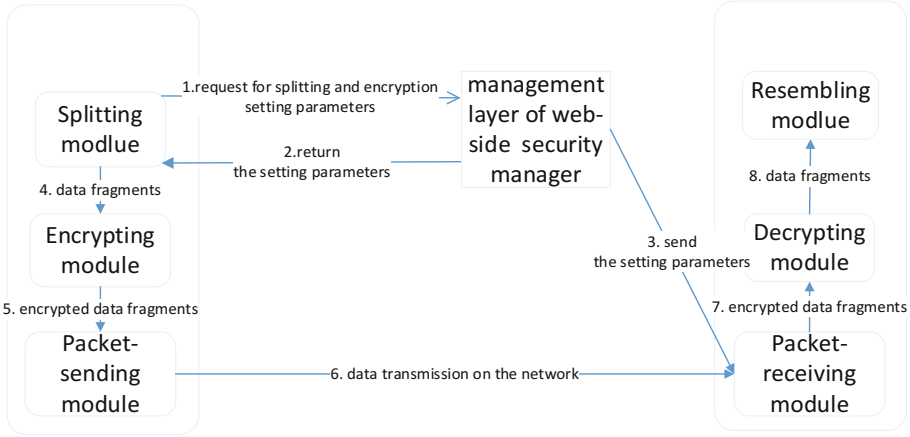


Fig. 3. Information flow chart of the client-side subsystem

4.3 The Web-Side Security Manager

The web-side security manager has a three-layer architecture. As in the right part of Fig. 2, the manager has a management layer, a controlling layer and a forwarding layer. Forwarding strategies and parameters are set in the management layer, with those the controlling layer decides transmission paths and distributes flow tables to switches in the forwarding layer.

4.3.1 Forwarding Layer

The forwarding layer comprises all switches in the network. The switches forward packets according to flow tables which is distributed by SDN controllers, so the packets are transmitted along certain paths specified by the controlling layer.

4.3.2 Management Layer

In the management layer, there are modules set up users' security settings and dealing with requests from the client-side subsystem and modules interacting with controllers in the controlling layer, as the top of Fig. 4.

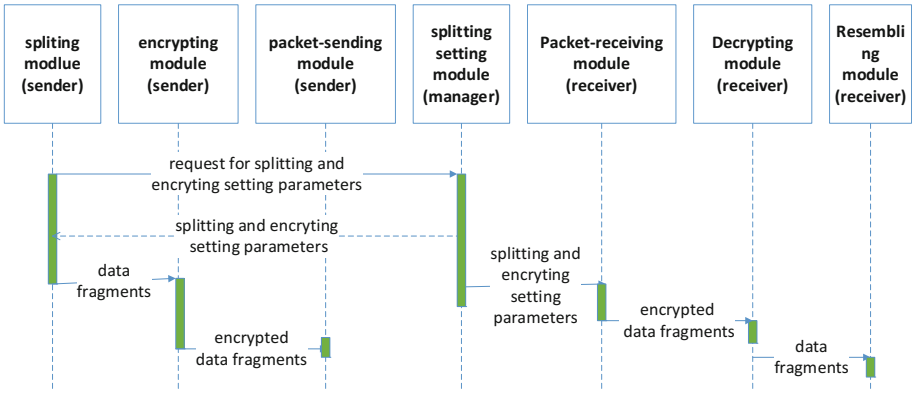


Fig. 4. The interacting sequence diagram of sending process

The management layer includes a connection setting module, a user management module, an encryption setting module, a time-slot setting module, an authority management module, a splitting setting module and a controller interaction module.

The user management module manages administrators who have authority to change the settings in the management layer and users who access the security transmission service. The connection setting module sets up the number of transmission connection. The encryption setting module decides which form of encryption will be used. The time-slot setting module determines the size of time-slot, the duration before a transmission path changes to another. The authority management module ranks users in different security levels corresponding to several security transmission strategies. The splitting setting module decides which form of splitting will be used. The controller interaction module is designed to implement interactions between the controlling layer and the management layer.

4.3.3 Controlling Layer

The controlling layer is actually based on a cluster of SDN controllers. It contains a management-layer interaction module interacting with the management layer, an automatic routing module, a time-slot scheduling module, a multi-connection scheduling module, a topology detecting module and other SDN controller modules

(as in the middle of Fig. 4). The management-layer interaction module gets forwarding strategies from the management layer and these strategies will be realized in the controlling layer. The automatic routing module gets topology of the network from the topology detecting module, and depending on a path algorithm calculates multiple shortest paths as candidates, then according to the decided time-slot strategy chooses several eligible paths as final transmission paths. The time-slot scheduling module uses the parameters set by the time-slot setting module to schedule transmission. The multi-connection scheduling module supports multiple users to use the service respectively at a same time and ensures all strategies not to interfere each other.

4.3.4 Working Process of the Web-Side Security Manager

The working process of the Web-side security manager is like this: link discovery module of SDN controllers detects state of network and send topology information to topology detecting module. Attained topology information, automatic routing module need time-slot scheduling and multi-connection strategies to decide transmission path. Then time-slot scheduling module and multi-connection scheduling module requests related parameters from management-layer interaction module. These parameters are generated in the management layer, so controller interaction layer calls connection setting module and time-slot setting module to generate setting parameters. Time-slot setting module requests user's security level from user management module to decide time-slot scheduling rules and returns the rules to time-slot scheduling module. Automatic routing module decides transmission path and corresponding flow table entries according to time-slot rules provided by time-slot scheduling module, multi-connection rules and topology information from topology detecting module. Finally controllers distribute flow table entries that automatic routing module decides to OpenFlow switches and transmission on specified path in different time-slot is realized.

4.4 Interactions Among Modules

Figure 4 is the interacting sequence diagram of sending process. First, the sending module of the sender's client-side subsystem logins to the web-side security manager. When the validated user need to send data with security transmission service, the sending module sends a request to the manager to access security transmission service. Received a sending request, the management module of the manager sends a request to destination host to get available port numbers, and generates multi-connection and time-slot strategies informing the controllers of. Then the management module returns splitting and encrypting strategies and destination ports list to the source host. The user can send out its data after receiving response from the manager.

5 Multi-connection and Time-Slot Scheduling

The multi-connection and time-slot forwarding strategy is the core of the security transmission service. It divides data flows into different logical paths based on both time and network links in order to make the amount of information that could be acquired on a certain link to minimum.

5.1 Multi-connection

Multi-connection supports to simultaneously transmit split data files through multiple data paths, one of which from a source port of a terminal host to a destination port is considered as a connection. Multiple connections have no affiliation with each other and each of them use an independent operating mechanism. Multi-connection can be realized by an accessory module of the client-side subsystem, and also can be realized on the entrance switch. Multi-connection divides one data flow into multiple data flows, which efficiently decreases the possibility to restore information of original data. And transmitting data through multiple connections simultaneously obviously approves transmission efficiency.

Multi-connection means that every available port of the source host connects to a port of the destination host. A transmission task will be completed by multiple connections. Every connection has multiple actual transmission paths and the transmission is according to the time-slot scheduling strategy, which will be discussed in next section.

5.2 Time-Slot Scheduling

Time-slot scheduling is implemented on every connection. Between two terminal hosts in a SDN network, there are always more than one transmission paths and for one connection at one point of time only one path is used to transmit data. Time-slot scheduling transfers the transmission from one candidate path to another in different time-slots. So, at different time packets of a data flow is forwarded by different switches. This strategy avoids certain switches capturing the whole data flow and accomplish security at the expense of time.

The time-slot scheduling strategy is implemented by time-slot setting module. The time-slot setting module generates rule of time-slot scheduling. A process activated by this module generates a random number n ($n \geq 1$) every unit time, which indicates every unit time is divided into n of time-slots. Most of programming languages provide methods to generate a random number and we adopt the most commonly used one, linear congruence [7]. Then pick out a number m ($m \leq n$), that represents m of time-slots is valid. Only in these m of valid time-slots, switches will be forwarding packets of the data flow, otherwise switches do not forward or forward fake data. The time-slot scheduling rule will be sent down to controllers, assisting controllers in formulating routing strategy.

The time-slot forwarding strategy is forwarding data flows following the time-slot scheduling rule. First, for an eligible time-slot, the time-slot scheduling module chooses a valid path to transmit data. When the time-slot is over, the transmission path switches to another valid candidate path. Switching from one path to another can be realized by a switching strategy based on adding new flow table entries to switches and expiration of old entries. Controllers add flow table entries to switches on the new path from source to destination to instruct new transmission path. Then packets of this data flow will be transmitted on the new path and after a short time (shorter than a time-slot) the old flow table entries on the switches on the old path will expire. If there is overlap between new path and old path, new flow table entry will replace the old one on the

overlapping switches. This way guarantees packets transmitting on the link against loss. Figure 5 shows the process of distributing flow table entries when switching from path 1 to path 2.

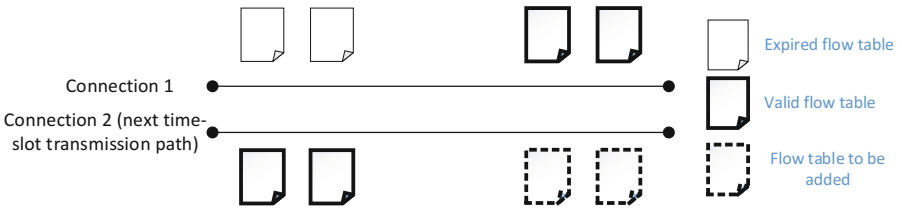


Fig. 5. The process of distributing flow table entries when switching from path 1 to path 2

For example, the process of time-slot scheduling on a connection between host1 to host2 in a SDN network that has topology of Fig. 8 is:

1. The SDN controller chooses a valid path to transmit (Path 1), and distributes flow table entries to switches on this path, in sequence from the one link to the destination host (Host 2) to the one link to the source host (Host 1). These entries have a valid time which is slightly longer than a time-slot and beyond the time they expire and will be deleted from the flow tables of the switches.
2. Host 1 begins to send a certain number of packets to host 2. At this time, only switches on Path 1 have flow table entries of this data flow in the network and lead to host 2, so packets are forwarding along Path 1.
3. After a time-slot, the controller chooses another transmission path (Path 2) and distributes new flow table entries to switches on this path, in sequence from the one link to the destination host (Host 2) to the one link to the source host (Host 1). For the entries expire beyond the valid time, it is not necessary for the controller to deal with the old entries.
4. As the new flow table entry replaces the old one on the entrance switch (the one link to the source host), the packets sent by Host 1 begin to be forwarded along Path 2.
5. After another time-slot, the controller chooses Path 3 and repeats the same distributing process. The transmission path will be switched to Path 3.
6. Repeat the procedures above until all packets arrived Host 2.

The time-slot scheduling combined with an appropriate algorithm of path-selection makes transmission flexible and unpredictable. To the users applying this strategy, transmission efficiency does not reduce and transmission security is guaranteed.

6 Experiment and Analysis

Experiment on simulation platform was conducted to verify the effectiveness of the security transmission service deployed with a SDN controller on a virtual machine. We choose Mininet [8] which is a system for rapidly prototyping large networks on the

constrained resources of a single laptop to simulate a SDN network with OVS switches [9] and hosts with name spaces. And in this experiment, we deploy our service on ONOS [10], a SDN controller. The experiment simulation platform is demonstrated in Fig. 7.

Simulated a SDN network with topology as in Fig. 6, we activate ONOS installed the security transmission service and on the web UI of ONOS all OVS switches on this experimental network can be found and flow tables on them can be viewed.

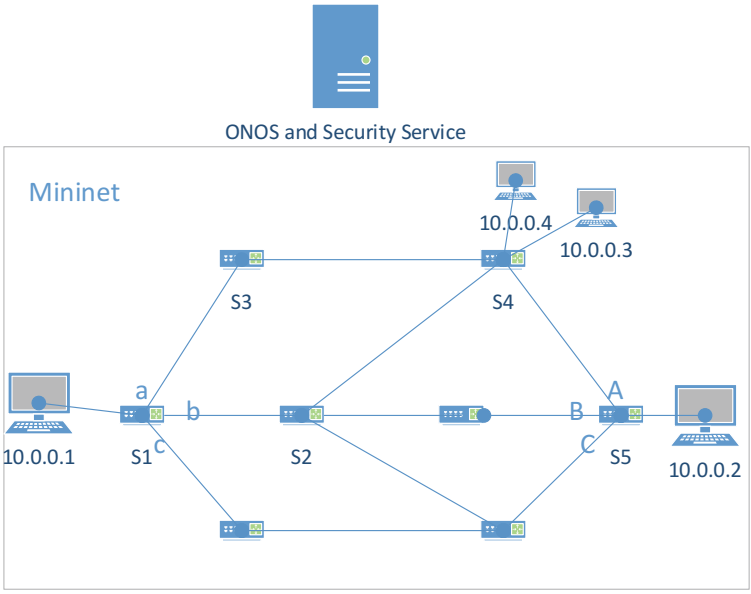


Fig. 6. Experiment network topology on the simulation platform

We send a large amount of packets from host on 10.0.0.1 to host on 10.0.0.2, and the time-slot setting module has been set to choose three paths as candidates. Examine the flow table on the switch S3 that has packets arriving, we find that at this point of time, it has a certain entry in its flow table. When the switch S2 that has packets arriving, we find that at the point of time, the certain entry in flow table on S3 has been deleted. It shows the flow table has been changed from time to time. And examine all the switches we can see packets flow in three transmission paths in different periods of time and that verifies that the time-slot scheduling strategy is working.

There is Table 1 showing parts of result data of the number of packets captured on port a, b, c on S1 and port A, B, C on S5, when we send a certain number of packets from 10.0.0.1 to 10.0.0.2 with different size of time-slot. Capture packets on ports of switch S1 and S5 and the result shows that three ports of each switch received equal amount of packets. That means packets transmitted on three paths and we cannot figure out how the size of time-slot matters to forwarding from the data. It is obvious that corresponding to three transmission paths, almost one third of the packets are transmitted on each path.

Table 1. Experiment with different time-slot and number of packets

Time-slot(s)	Number of packets	Port a	Port b	Port c	Port A	Port B	Port C
5	10,000	3,095	3,044	3,861	3,095	3,044	3,861
	50,000	16,747	17,456	15,797	16,747	17,456	15,797
	100,000	32,493	35,060	32,447	32,493	35,060	32,447
4	100,000	32,853	33,846	33,301	32,853	33,846	33,301
2	100,000	30,934	35,456	33,610	30,934	35,456	33,610
0.5	100,000	32,605	33,588	33,807	32,605	33,588	33,807
0.05	100,000	33,159	33737	33,104	33,159	33737	33,104

If switch S4 is hijacked, irrespective of multi-connection and encryption, one third of packets at most will be revealed. With more paths selected as candidates of time-slot scheduling and appropriate size of the time-slot, the security of transmission is obviously enhanced.

7 Summary

In this paper, we put forward a solution to data leakage on switch level. The security transmission service aims at reducing the risk of revealing information in transmission process and utilizes time division and multiple path strategies to improve security of networks. Experiment proves that this security scheme based on routing by time works and can decrease the percentage of capturing targeted data flow from one or more switches to some degree. Attending other encryption and distribution measures, safety of users' data is guaranteed at a great extent.

SDN architecture enables easy and direct management of networks, which provides basis of the security transmission service. At present, the service operates on a relatively static and small-scaled network. About the appropriate size of time-slot, multi-path algorithm and numbers of chosen paths of time-slot scheduling, further studies and later experiment remain to be continued.

Acknowledgments. This work is supported by Beijing Natural Science Foundation (4174099) and National High Technology Research and Development Program (863 Program) of China (2015AA050203).

References

1. Nadeau, T.D., Gray, K.: Software Defined Networks. O'Reilly Media Inc, Massachusetts (2013)
2. Feamster, N., Rexford, J., Zegura, E.: The Road to SDN. Queue **11**(12), 87–98 (2013)
3. Scott-Hayward, S., O'Callaghan, G., Sezer, S.: SDN security: a survey. Future Netw Serv. **2013**, 1–7 (2013)

4. Mizrahi, T., Moses, Y.: Time-based updates in software defined networks. In: Proceedings of the Second ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking (HotSDN), pp. 163–164 (2013)
5. Mizrahi, T., Moses, Y.: On the necessity of time-based updates in SDN. Open Networking Summit (2014)
6. Ishimoli, A., Farias, F., Cerqueira, E., Abelem, A.: Control of multiple packet schedulers for improving QoS on OpenFlow/SDN networking. In: Second European Workshop on Software Defined Networks, pp. 81–86 (2013)
7. Granger, P.: Statistic Analysis of linear congruence equalities among variables of a program. CAAP Colloquium on Trees in Algebra and Programming, **493**(10), 169–192 (1991)
8. Lantz, B., Heller, B., McKeown, N.: A network in a laptop: rapid prototyping for software-defined networks. In: Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks. ACM (2010)
9. Pan, H.Y., Wang, S.Y.: Optimizing the SDN control-plane performance of the Openvswitch software switch. In: IEEE Symposium on Computers and Communication (ISCC). IEEE, pp. 403–408 (2016)
10. Muqaddas, A.S., Bianco, A., Giaccone, P., et al.: Inter-controller traffic in ONOS clusters for SDN networks. In: IEEE International Conference on Communications (ICC). IEEE, pp. 1–6 (2016)



Construction of Quasi-Cyclic LDPC Codes with Diagonal Structure of Parity-Check Matrices

Hengzhou Xu[1,2(✉)], Huaan Li[2], Min Zhu[2], Bo Zhang[1], and Baoming Bai[2]

[1] School of Network Engineering, Zhoukou Normal University, Zhoukou 466001, China
hzxu@zknu.edu.cn

[2] State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China
bmbai@mail.xidian.edu.cn

Abstract. Quasi-cyclic (QC) LDPC codes whose parity-check matrices have diagonal structure play an important role in channel coding of 5G communications. In this paper, we study an algebraic-based method for constructing QC LDPC codes with diagonal structure of parity-check matrices. We first analyze the cycle structure of this class of QC LDPC codes and then divide the diagonal parity-check matrix into two parts, i.e., the diagonal matrix and the non-diagonal matrix. By employing the masking technique, we design the non-diagonal matrix based on prime field and QC LDPC codes with diagonal structure of parity-check matrices are proposed. Numerical results show that the constructed QC LDPC codes perform much better than the WiMAX-LDPC codes.

Keywords: LDPC code · Diagonal matrix · Prime field · Masking Girth

1 Introduction

Low-density parity-check (LDPC) codes [1] were proposed by Gallager in 1962. Owing to the limitation of the computing, it was not until the late 1990s that LDPC codes with capacity-approaching performance were rediscovered [2,3]. Later, LDPC codes have attracted lots of researcher and engineers' attention in both industry and academia [4,5]. Quasi-cyclic (QC) LDPC codes [6], as a special class of LDPC codes, have advantages of easy hardware implementation of the encoder and decoder [7], fast decoder convergence [8], and low error-floor [9]. Hence, QC LDPC codes have been adopted by many communication standards, such as the DVB-S2 standard [10] and the Consultative Committee for Space Data Systems (CCSDS) standard [11].

As is well known, LDPC codes are a subset of the class of linear block codes. Therefore, the encoding of LDPC codes can be divided into two major classes: (1)

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 460–469, 2018.
https://doi.org/10.1007/978-3-319-78130-3_48

encoding based on the generator matrix; and (2) encoding based on the structural parity-check matrix, such as the parity-check matrix with diagonal structure [12]. The first is a general encoding of linear block codes, and it depends on the generator matrix. Since the generator matrix obtained from the parity-check matrix may not contain structural properties which are suitable for hardware implementation, such encoding is not often used in the practical applications. However, the parity-check matrix with some structures can be directly used to encode. The diagonal parity-check matrices [13] are a kind of such matrices. Recently, LDPC codes have been selected as the coding scheme for Enhanced Mobile Broadband (eMBB) data channel [14] and LDPC codes with diagonal structure of parity-check matrices are considered as the candidate standard codes for 5G communications.

In this paper, we focus on QC LDPC codes with diagonal structure of parity-check matrices, referred to as diagonal QC LDPC codes. This class of LDPC codes had been constructed in the IEEE 802.16e (WiMAX) standard [15]. Although these codes have good performance in the waterfall region, they have high error floor. This may be due to the parity-check matrices of these codes containing weight-2 columns. In 2016, Bocharova *et al.* presented a unified approach to search for and optimize these LDPC codes [16]. Simulation results show that their optimized codes outperform the WiMAX-LDPC codes and have much lower error floor, but need a greedy search by means of computer. Hence, it is interesting to directly construct this class of good codes based on algebraic method.

In this paper, we study the construction of diagonal QC LDPC codes. According to the structure of the diagonal parity-check matrix, we divide it into two sub-matrices, i.e., the diagonal matrix and the non-diagonal matrix. By analyzing the cycle structure of this class of LDPC codes, we find that the existence of cycles is mainly determined by the non-diagonal matrix. It is well known that cycles of length 4 can be effectively avoided in the Tanner graphs of LDPC codes constructed based on algebraic tools, such as finite fields [17], combinatorial designs [18–20], and finite geometries [21,22]. On this basis, we propose an algebraic based method for constructing QC LDPC codes whose parity-check matrices are diagonal. In order to facilitate the introduction and understanding of the proposed method, we employ the masking technique to construct the non-diagonal matrix based on prime fields and a class of diagonal QC LDPC codes is obtained. Numerical results show that the constructed QC LDPC codes perform much better than the WiMAX-LDPC codes under iterative decoding over the additive white Gaussian noise (AWGN) channel.

2 Preliminaries

An (n, k) LDPC code is defined by the null space of an $m \times n$ sparse parity-check matrix $\mathbf{H}$ with rank $(n - k)$ $(\leq m)$. If the column weight and row weight of the matrix $\mathbf{H}$ are J and L, respectively, we say the code defined by the null space of such $\mathbf{H}$ is (J, L)-regular; otherwise, we say the resultant code is irregular. If

$\mathbf{H}$ is an array of circulant permutation matrices (CPMs) and/or zero matrices (ZMs) of the same size, the resultant code is quasi-cyclic (QC). For convenience, consider the following array $\mathbf{H}$ of CPMs and/or ZMs of size $p \times p$.

$$\mathbf{H} = \begin{bmatrix} \mathbf{I}(p_{1,1}) & \mathbf{I}(p_{1,2}) & \cdots & \mathbf{I}(p_{1,L}) \\ \mathbf{I}(p_{2,1}) & \mathbf{I}(p_{2,2}) & \cdots & \mathbf{I}(p_{2,L}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{I}(p_{J,1}) & \mathbf{I}(p_{J,2}) & \cdots & \mathbf{I}(p_{J,L}) \end{bmatrix}, \tag{1}$$

where, for $1 \le i \le J, 1 \le j \le L, p_{i,j} \in \{\{-1\} \cup \mathbb{Z}_p\}$, and for $p_{i,j} \in \mathbb{Z}_p, \mathbf{I}(p_{i,j})$ is a CPM of size $p \times p$ created by simultaneously shifting $p_{i,j}$ positions of each row of the identity matrix $\mathbf{I}$ to the right, and $\mathbf{I}(-1)$ is a $p \times p$ ZM. It can be seen that the positions of nonzero elements in $\mathbf{H}$ are exactly determined by a matrix $\mathbf{P}$, known as *exponent matrix*, where

$$\mathbf{P} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,L} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ p_{J,1} & p_{J,2} & \cdots & p_{J,L} \end{bmatrix}. \tag{2}$$

Notice that, for $1 \le i \le J, 1 \le j \le L, p_{i,j} \ne -1, p_{i,j}$ is called the shift value of CPM $\mathbf{I}(p_{i,j})$. In other words, the matrix $\mathbf{H}$ and the exponent matrix $\mathbf{P}$ are corresponding one by one.

Tanner graph [23] is a graphical form to describe an LDPC code. The Tanner graph is a bipartite graph in which two parts of points correspond to the columns and rows of $\mathbf{H}$, respectively. In coding theory, these two types of points are called variable nodes and check nodes. An edge connecting the variable node i to the check node j exists if and only if the row-i and column-j element in $\mathbf{H}$ is 1. A cycle in the Tanner graph is a family of connected edges which start and end at the same node and contain no other edges no more than once. The length of the shortest cycle in the Tanner graph is called the girth. In general, cycles of length 4 degrade the performance when decoded with the iterative algorithm. Hence, a constraint is usually imposed on $\mathbf{H}$ to avoid the cycles of length 4 in the Tanner graph of an LDPC code. This is known as the row-column constraint (or RC-constraint) [4] on $\mathbf{H}$: no two columns (or two rows) have more than one nonzero element in the same position. Notice that an LDPC code created by $\mathbf{H}$ which satisfies the RC-constraint has girth at least 6.

3 Cycle Analysis of QC LDPC Codes with Diagonal Structure of Parity-Check Matrices

In this section, we analyze the cycle structure of QC LDPC codes whose parity-check matrix is diagonal. There are many types of diagonal matrices, such as single diagonal matrices, bidiagonal matrices, and their mixed forms. For convenience, we only study the bidiagonal matrix $\mathbf{H}$ here. Notice that the following analysis and method are also applicable to the other diagonal matrices.

We first divide the diagonal parity-check matrix $\mathbf{H}$ into two parts: the non-diagonal matrix and the diagonal matrix, i.e.,

$$\mathbf{H} = [\mathbf{H}_1 \,|\, \mathbf{H}_2]\,,$$

where $\mathbf{H}_1$ is the non-diagonal matrix and $\mathbf{H}_2$ is the diagonal matrix. Similarly, if $\mathbf{H}$ consists of CPMs and/or ZMs, the exponent matrix $\mathbf{P}$ of $\mathbf{H}$ can also be divided into two sub-matrices. For an example, when such $\mathbf{H}$ is bidiagonal, its corresponding exponent matrix $\mathbf{P}$ can be expressed as

$$\begin{aligned}\mathbf{P} &= [\mathbf{P}_1 \,|\, \mathbf{P}_2]_{J\times L} \\ &= \left[\begin{array}{cccc|ccccc} p_{1,1} & p_{1,2} & \cdots & p_{1,L-J+1} & 0 & -1 & \cdots & -1 & -1 \\ p_{2,1} & p_{2,2} & \cdots & p_{2,L-J+1} & 0 & 0 & \cdots & -1 & -1 \\ p_{3,1} & p_{3,2} & \cdots & p_{3,L-J+1} & -1 & 0 & \cdots & -1 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{J-1,1} & p_{J-1,2} & \cdots & p_{J-1,L-J+1} & -1 & -1 & \cdots & 0 & 0 \\ p_{J,1} & p_{J,2} & \cdots & p_{J,L-J+1} & -1 & -1 & \cdots & -1 & 0 \end{array}\right], \end{aligned} \tag{3}$$

where $\mathbf{P}_1$ is a $J \times (L - J + 1)$ matrix and $\mathbf{P}_2$ is a $J \times (J - 1)$ matrix. In order to facilitate the encoding based on the parity-check matrix $\mathbf{H}$, $p_{1,L-J+1}$ and $p_{J,L-J+1}$ are set as nonnegative integers and, for $1 < i < J$, $p_{i,L-J+1}$ have one element and only one element being nonnegative integer and the other elements being -1.

Based on Eq. (3) in [24], we can see that the existence of the cycles is related to the elements in the exponent matrix $\mathbf{P}$. Hence, the cycles of an LDPC code whose exponent matrix is $\mathbf{P}$ given by Eq. (3) can be classified into the following three categories:

1. Cycles generated from the matrix $\mathbf{P}_1$;
2. Cycles generated from the matrix $\mathbf{P}_2$;
3. Cycles generated between the matrix $\mathbf{P}_1$ and the matrix $\mathbf{P}_2$;

Consider the second kind of cycles, i.e., the cycles which are generated from the matrix $\mathbf{P}_2$. According to the bidiagonal structure of $\mathbf{P}_2$, it can be easily seen that this kind of cycles does not exist.

Consider the third kind of cycles, i.e., the cycles which are generated between the matrix $\mathbf{P}_1$ and the matrix $\mathbf{P}_2$. If such a cycle of length $2(i_1 - i_0 + 1)$ exists, then the structure of this cycle is given in Fig. 1. Based on Eq. (3) in [24], we can see that the necessary and sufficient conditions for the existence of such cycles is

$$p_{i_0,j_0} = p_{i_1,j_0} \pmod p,$$

where $1 \le i_0, i_1 \le J, 1 \le j_0 \le L$, and $i_0 \ne i_1$. Hence, to avoid this kind of cycles, the **column constraint** should be imposed on $\mathbf{P}_1$: *all nonnegative elements in each column are different from each other.*

Based on the above analysis, we can see that the cycles of a diagonal QC LDPC code are mainly determined by the non-diagonal sub-matrix of the exponent matrix. Hence, the design of the diagonal QC LDPC codes can be simplified as the construction of an exponent matrix which satisfies the column constraint.

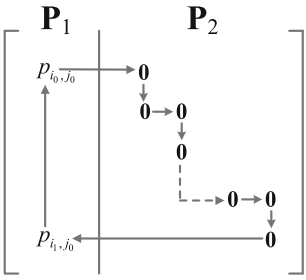


Fig. 1. Cycles generated between the matrix $\mathbf{P}_1$ and the matrix $\mathbf{P}_2$.

4 Construction of QC LDPC Codes with Diagonal Structure of Parity-Check Matrices

In this section, we propose a method to construct QC LDPC codes with diagonal parity-check matrices and numerical simulation results are also given.

4.1 Design of Non-diagonal Exponent Sub-matrix

Based on the cycle analysis in the above section, we can see that construction of diagonal QC LDPC codes is equivalent to the design of the non-diagonal exponent sub-matrices which satisfy the column constraint. In the following, we present a method for constructing non-diagonal exponent sub-matrices based on prime fields and then diagonal QC LDPC codes are obtained.

Consider the prime field $\mathbb{F}_p$ where p is a prime. We construct a bidiagonal QC LDPC code with code rate $(1 - J/L)$ and code length pL whose exponent matrix consists of CPMs and/or ZMs of size $p \times p$. Obviously, the size of the exponent matrix is $J \times L$. Based on Eq. (3) and the above cycle analysis, we only need to construct a $J \times (L - J + 1)$ non-diagonal exponent sub-matrix $\mathbf{P}_1$. Consider the following $J \times p$ matrix

$$\mathbf{w} = [w_{i,j}] = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,p} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ w_{J,1} & w_{J,2} & \cdots & w_{J,p} \end{bmatrix} = \begin{bmatrix} 1 \cdot 1 & 1 \cdot 2 & \cdots & 1 \cdot p \\ 2 \cdot 1 & 2 \cdot 2 & \cdots & 2 \cdot p \\ \vdots & \vdots & \ddots & \vdots \\ J \cdot 1 & J \cdot 2 & \cdots & J \cdot p \end{bmatrix},$$

where $w_{i,j} \in \mathbb{F}_p$ and $w_{i,j} = i \cdot j \pmod{p}$ for $1 \le i \le J, 1 \le j \le p$. By choosing $(L - J + 1)$ columns from the matrix $\mathbf{w}$, we can obtain a $J \times (L - J + 1)$ matrix, denoted by $\mathbf{P}_1^*$. Notice that the size of selection space of $(L - J + 1)$ columns from $\mathbf{w}$ is

$$\binom{p}{L - J + 1} = \frac{p!}{(L - J + 1)! \cdot (p - L + J - 1)!}.$$

How to choose $(L - J + 1)$ columns which can result in good LDPC codes is of interest. Significant work about this had been given in [25] and then we employ

their method to choose good columns here. By replacing some elements of $\mathbf{P}_1^*$ with -1, we can obtain the required $J \times (L - J + 1)$ matrix $\mathbf{P}_1$ with a given degree distribution. Notice that this replacement technique is called masking [26,27]. Here the existing masking matrices are employed or the required masking matrices are designed based on the method in [27]. Notice that the method in [27] can also be used to directly construct the required $J \times (L - J + 1)$ matrix $\mathbf{P}_1$. Based on the resulting $J \times (L - J + 1)$ matrix $\mathbf{P}_1$ and Eq. (3), we can obtain a $J \times L$ exponent matrix and the bidiagonal QC LDPC code with code rate $(1 - J/L)$ and code length pL is constructed.

For all elements of each column of the matrix $\mathbf{w}$, we select any two elements, i.e., p_{i_0,j_0} and p_{i_1,j_0}. Since

$$(i_0 - i_1) \neq 0 \pmod{p},$$

and p is a prime, then

$$p_{i_0,j_0} - p_{i_1,j_0} = i_0 \cdot j_0 - i_1 \cdot j_0 = (i_0 - i_1) \cdot j_0 \neq 0 \pmod{p}.$$

That is, the matrix $\mathbf{w}$ satisfies the column constraint. In other words, cycles generated by the designed $J \times (L - J + 1)$ matrix $\mathbf{P}_1$ and the bidiagonal matrix $\mathbf{P}_2$ do not exist. Hence, the non-diagonal exponent sub-matrix $\mathbf{P}_1$ mainly affects the error performance of the resultant bidiagonal QC LDPC codes. Notice that, by suitably optimizing the elements of each column, the other algebraic-based methods can also be used to construct the matrix $\mathbf{P}_1$ and we omit them here. Moreover, we can also search for non-diagonal exponent matrices which satisfy the column constraint.

4.2 Simulation Results

In order to show the efficiency of our proposed method, we compare the constructed codes with the WiMAX-LDPC codes in [15], and numerical results are given in this subsection. Notice that the simulation parameters are the AWGN channel and the binary phase shift keying (BPSK) modulation. The received words are decoded using the sum-product algorithm (SPA) with 50 iterations.

Here we take the $(576, 288)$ WiMAX-LDPC code in [15] as an example. This code has the following exponent matrix of size 12×24:

$$\begin{bmatrix}
-1 & 94 & 73 & -1 & -1 & -1 & -1 & -1 & 55 & 83 & -1 & -1 & 7 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
-1 & 27 & -1 & -1 & -1 & 22 & 79 & 9 & -1 & -1 & -1 & 12 & -1 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
-1 & -1 & -1 & 24 & 22 & 81 & -1 & 33 & -1 & -1 & -1 & 0 & -1 & -1 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
61 & -1 & 47 & -1 & -1 & -1 & -1 & -1 & 65 & 25 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
-1 & -1 & 39 & -1 & -1 & -1 & 84 & -1 & -1 & 41 & 72 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 \\
-1 & -1 & -1 & -1 & 46 & 40 & -1 & 82 & -1 & -1 & -1 & 79 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & -1 & -1 & -1 & -1 & -1 \\
-1 & -1 & 95 & 53 & -1 & -1 & -1 & -1 & -1 & 14 & 18 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & -1 & -1 & -1 & -1 \\
-1 & 11 & 73 & -1 & -1 & -1 & 2 & -1 & -1 & 47 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & -1 & -1 & -1 \\
12 & -1 & -1 & -1 & 83 & 24 & -1 & 43 & -1 & -1 & -1 & 51 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & -1 & -1 \\
-1 & -1 & -1 & -1 & -1 & 94 & -1 & 59 & -1 & -1 & 70 & 72 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & -1 \\
-1 & -1 & 7 & 65 & -1 & -1 & -1 & -1 & 39 & 49 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 \\
43 & -1 & -1 & -1 & -1 & 66 & -1 & 41 & -1 & -1 & -1 & 26 & 7 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0
\end{bmatrix}.$$

Based on the above method and the masking matrix of the $(576, 288)$ WiMAX-LDPC code, we can design the following exponent matrix of size 12×24 based on the prime field $\mathbb{F}_{23}$:

$$\begin{bmatrix}
-1 & 2 & 3 & -1 & -1 & -1 & -1 & -1 & 9 & 10 & -1 & -1 & 13 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
-1 & 4 & -1 & -1 & -1 & 12 & 14 & 16 & -1 & -1 & -1 & 1 & -1 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
-1 & -1 & -1 & 12 & 15 & 18 & -1 & 1 & -1 & -1 & -1 & 13 & -1 & -1 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
4 & -1 & 12 & -1 & -1 & -1 & -1 & -1 & 13 & 17 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
-1 & -1 & 15 & -1 & -1 & -1 & 12 & -1 & -1 & 4 & 9 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 \\
-1 & -1 & -1 & -1 & 7 & 13 & -1 & 2 & -1 & -1 & -1 & 3 & 9 & -1 & -1 & -1 & -1 & 0 & 0 & -1 & -1 & -1 & -1 & -1 \\
-1 & -1 & 21 & 5 & -1 & -1 & -1 & -1 & -1 & 1 & 8 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & -1 & -1 & -1 & -1 \\
-1 & 16 & 1 & -1 & -1 & -1 & 10 & -1 & -1 & 11 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & -1 & -1 & -1 \\
9 & -1 & -1 & -1 & 22 & 8 & -1 & 3 & -1 & -1 & -1 & 16 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & -1 & -1 \\
-1 & -1 & -1 & -1 & -1 & 14 & -1 & 11 & -1 & -1 & 18 & 5 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & -1 \\
-1 & -1 & 10 & 21 & -1 & -1 & -1 & -1 & 7 & 18 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 \\
12 & -1 & -1 & -1 & -1 & 3 & -1 & 4 & -1 & -1 & -1 & 6 & 18 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0
\end{bmatrix}.$$

Here the selected column indexes of $\mathbf{w}$ are 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13. By replacing the elements of the above exponent matrix with the corresponding CPMs/ZMs of size 23×23, we can construct a $(552, 276)$ bidiagonal QC LDPC code. The bit/word error performance of the proposed $(552, 276)$ QC LDPC code and the comparable $(576, 288)$ WiMAX-LDPC code is shown in Fig. 2. We can see that, at the word error rate (WER) of 10^{-4}, our proposed code achieves about 0.5 dB coding gain over the $(576, 288)$ WiMAX-LDPC code. Moreover, the error floor of the proposed code can not be found at the bit error rate (BER) of 3×10^{-7}. This may be caused by the following reason. The total number of short cycles of our proposed code is less than that of the WiMAX-LDPC code. The cycle distributions of these two bidiagonal QC LDPC codes are given in Table 1. It can be seen from Table 1 that, compared with the WiMAX-LDPC code, our proposed codes have much less short cycles of length no more than 10. That is, cycles of lengths 4, 6, 8, and 10 play an important role in the design of LDPC codes, and this result is consistent with the conclusion in [28].

Table 1. The cycle distribution of the proposed $(552, 276)$ QC LDPC code and the comparable $(576, 288)$ WiMAX-LDPC code

	Length & multiplicity							
	4	6	8	10	12	6 & 8	6 & 8 & 10	6 & 8 & 10 & 12
WiMAX-LDPC code	0	528	7344	75816	920508	7872	83688	1004196
Proposed LDPC code	0	690	6670	73715	958410	7360	81075	1039485

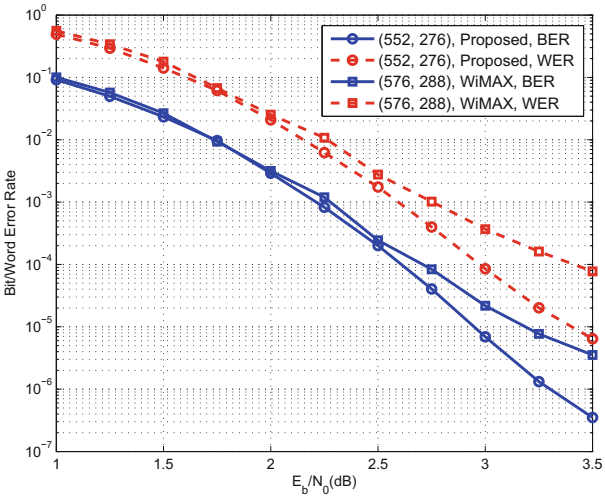


Fig. 2. Bit/Word error performance of the proposed $(552, 276)$ QC LDPC code and the comparable $(576, 288)$ WiMAX-LDPC code in [15].

5 Conclusion

In this paper, we studied on the construction of QC LDPC codes with parity-check matrices owning diagonal structure. By analyzing the cycle structure of this class of QC LDPC codes, we found that the cycles of such codes are mainly determined by the non-diagonal part of its exponent matrix. Hence, by combining with the masking technique, we constructed the non-diagonal matrix based on prime field and QC LDPC codes with diagonal structure of parity-check matrices are obtained. Numerical results show that the proposed QC LDPC codes have good performance over the AWGN channel. Notice that the non-diagonal matrix can also be constructed based on the other algebraic tools, such as finite geometries and combinatorial designs, and this will result in another class of good QC LDPC codes with diagonal structure of parity-check matrices.

Acknowledgment. This work was supported in part by National Natural Science Foundation of China under Grants 61771364, 61701368, and 61103143, the Joint Funds of the National Natural Science Foundation of China under Grants U1504601 and U1404622, Key Scientific and Technological Project of Henan under Grants 162102310589 and 172102310124, Key Scientific Research Projects of Henan Educational Committee under Grant 18B510022, and the School-Based Program of Zhoukou Normal University under Grant ZKNUB2201705.

References

1. Gallager, R.: Low-density parity-check codes. IRE Trans. Inf. Theory **8**, 21–28 (1962)
2. MacKay, D., Neal, R.: Near shannon limit performance of low density parity-check codes. Electron. Lett. **32**, 1645–1646 (1996)
3. MacKay, D.: Good error-correcting codes based on very sparse matrices. IEEE Trans. Inf. Theory **45**, 399–431 (1999)
4. Ryan, W.E., Lin, S.: Channel Codes: Classical and Modern. Cambridge University Press, New York (2009)
5. Zhou, H., Li, Q., Tong, W., Kausar, S., Zhu, H.: P-Aware: a proportional multi-resource scheduling strategy in cloud data center. Clust. Comput. **19**, 1089–1103 (2016)
6. Fossorier, M.P.C.: Quasi-cyclic low-density parity-check codes from circulant permutation matrices. IEEE Trans. Inf. Theory **50**, 1788–1793 (2004)
7. Huang, Q., Diao, Q., Lin, S., Abdel-Ghaffar, K.: Cyclic and quasi-cyclic LDPC codes on constrained parity-check matrices and their trapping sets. IEEE Trans. Inf. Theory **58**, 2648–2671 (2012)
8. Zhang, L., Lin, S., Abdel-Ghaffar, K., Zhou, B.: Circulant arrays: rank analysis and construction of quasi-cyclic LDPC codes. In: 2010 IEEE International Symposium on Information Theory, pp. 814–818. IEEE Press, New York (2010)
9. Li, J., Liu, K., Lin, S., Abdel-Ghaffar, K.: Quasi-cyclic LDPC codes on two arbitrary sets of a finite field. In: 2014 IEEE International Symposium on Information Theory, pp. 2454–2458. IEEE Press, New York (2014)
10. Digital Video Broadcasting (DVB), Europe Telecommunication Standards Institution (2009)
11. Short Block Length LDPC Codes for TC Synchronization and Channel Coding, CCSDS 231.1-O-1 (2015)
12. Zhang, P., Yu, S., Liu, C., Jiang, L.: Efficient encoding of QC-LDPC codes with multiple-diagonal parity-check structure. Electron. Lett. **50**, 320–321 (2014)
13. Li, P., Leung, W.K., Phamdo, N.: Low density parity check codes with semi-random parity check matrix. Electron. Lett. **35**, 38–39 (1999)
14. 3GPP TSG RAN WG1 Meeting ♯87: RAN1 Chairmans Notes (2016)
15. Air Interface for Fixed Broadband Wireless Access Systems, IEEE Standard P802.16e/D12 (2005)
16. Bocharova, I., Kudryashov, B., Johannesson, R.: Searching for binary and nonbinary block and convolutional LDPC codes. IEEE Trans. Inf. Theory **62**, 163–183 (2016)
17. Zhang, L., Lin, S., Abdel-Ghaffar, K., Ding, Z., Zhou, B.: Quasi-cyclic LDPC codes on cyclic subgroups of finite fields. IEEE Trans. Commun. **59**, 2330–2336 (2011)
18. Lan, L., Tai, Y., Lin, S., Memari, B., Honary, B.: New constructions of quasi-cyclic LDPC codes based on special classes of BIBD's for the AWGN and binary erasure channels. IEEE Trans. Commun. **56**, 39–48 (2008)
19. Falsafain, H., Esmaeili, M.: Construction of structured regular LDPC codes: a design-theoretic approach. IEEE Trans. Commun. **61**, 1640–1647 (2013)
20. Xu, H., Feng, D., Sun, C., Bai, B.: Construction of LDPC codes based on resolvable group divisible designs. In: 2015 International Workshop on High Mobility Wireless Communications, pp. 111–115. IEEE Press, New York (2015)
21. Diao, Q., Tai, Y., Lin, S., Abdel-Ghaffar, K.: LDPC codes on partial geometries: construction, trapping set structure, and puncturing. IEEE Trans. Inf. Theory **59**, 7898–7914 (2013)

22. Diao, Q., Li, J., Lin, S., Blake, I.F.: New classes of partial geometries and their associated LDPC codes. IEEE Trans. Inf. Theory **62**, 2947–2965 (2016)
23. Tanner, R.M.: A recursive approach to low complexity codes. IEEE Trans. Inf. Theory **27**, 533–547 (1981)
24. Tasdighi, A., Banihashemi, A.H., Sadeghi, M.: Efficient search of girth-optimal QC-LDPC codes. IEEE Trans. Inf. Theory **62**, 1552–1564 (2016)
25. Zhao, S., Ma, X.: Construction of high-performance array-based non-binary LDPC codes with moderate rates. IEEE Commun. Lett. **20**, 13–16 (2016)
26. Liu, Y., Li, Y.: Design of masking matrix for QC-LDPC codes. In: IEEE Information Theory Workshop, pp. 1–5. IEEE Press, New York (2013)
27. Xu, H., Feng, D., Luo, R., Bai, B.: Construction of quasi-cyclic LDPC codes via masking with successive cycle elimination. IEEE Commun. Lett. **20**, 13–16 (2016)
28. Sun, C., Xu, H., Feng, D., Bai, B.: (3, L) Quasi-cyclic LDPC codes: simplified exhaustive search and designs. In: 2016 9th International Symposium on Turbo Codes and Iterative Information Processing, pp. 271–275. IEEE Press, New York (2016)

Check for updates

Analyzing the Impact of Redundant Paths on Multistage Switch Performance

Gao Ya[1(✉)], Pan Weitao[2], and Zheng Ling[2]

[1] Wuxi Institute of Technology, Wuxi, Jiangsu, China
gaoya@wxit.edu.cn
[2] State Key Laboratory of the Integrated Service Networks (ISN), Xidian University, Xi'an, Shaanxi, China

Abstract. The impact of redundant paths on the performance of onboard multistage switches is investigated systematically for the first time. A calculation model based on Markov chain for throughput performance of Clos-network switch with finite buffers is proposed and proven to have considerable accuracy. Simulations under various parameters are performed to evaluate the impact of redundant paths on the re-sequencing delay, end-to-end delay, as well as the throughput performances. The conclusions drawn in this study can serve as a theoretical guide in the design of onboard multistage switch fabrics.

Keywords: Onboard switch · Clos-network · Re-sequencing
Queueing analysis

1 Introduction

The serious decline in the reliability of onboard switching is induced by the harsh space radiation environment [1–3], such as single-event upset, single-event functional interrupt, and single-event latch-up (SEL). Traditional onboard switching architectures, such as shared memory switch, crossbar switch, and knock-out network [4,5], cannot exchange information between input and output ports if a link fault occurs and thus have low reliability. The continuous improvement in satellite link rate [6,7] makes it difficult for a single chip to achieve the required processing capacity.

Clos-network is an important solution in fabricating high-speed larger-capacity onboard switches because of its modularity, scalability, and reliability improved by multiple paths between input and output ports [8–11]. Central stage buffered Clos-network switches, such as memory-memory-memory (MMM) switch [12,13] and space-memory-memory (SMM) switch [9,14], do not need complex centralized schedulers [15] or costly memory speedup [10,11], thus gaining increasing attention.

The combined-input-and-crosspoint-queued crossbar-based SMM Clos-network (CICQ-SMM) [9] reduces the influence of link faults or crosspoint faults

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 470–478, 2018.
https://doi.org/10.1007/978-3-319-78130-3_49

on switch performance by adding central modules (CMs) to increase the redundant paths between input and output ports. Out-of-sequence cells will be re-sequenced at the output ports of the Clos-network. The dispatching algorithm proposed for CICQ-SMM switch [9] achieves 100% throughput with infinite buffers and complex scheduling. However, employing an infinite buffer is unrealistic. If the buffers of the last two stages are infinite, then no bound exists for re-sequencing buffer.

It's harder to evaluate the performance of a switch with finite buffers than that of a switch with infinite buffers [16]. For a central stage buffered Clos-network switch with finite buffers, in-depth studies are needed to clarify the effect of redundant paths on switch performance and the overhead introduced by increasing redundant paths. As input modules (IMs) are used to balance input traffic, buffers can be removed if IMs adopt a predetermined cyclic shift configuration and no contention occurs at all. Thus, we consider crosspoint-queued crossbar-based SMM (CQ-SMM) Clos-network in this study. The impacts of the number of redundant paths and the buffer size of each module on the throughput and delay performance of CQ-SMM Clos-network are investigated through theoretical analysis and simulation.

2 The CQ-SMM Switch Architecture

Figure 1 shows the CQ-SMM Clos-network, denoted as $C(m, n, k)$, which includes k IMs with size of $n \times m$, m central modules (CM) with size of $k \times k$, and k output modules (OM) with size of $m \times n$. The switch size of the CQ-SMM, denoted as N, equals the product of n and k, i.e., $N = nk$. m $(m \geq n)$ paths exist between input and output ports. Here, we define that the Clos-network switch has s redundant paths, where $s = m - n$, $s \geq 0$.

We assume that packets are segmented into cells and time is slotted. Each module of IM is a bufferless crossbar and adopts a predetermined cyclic shift con-

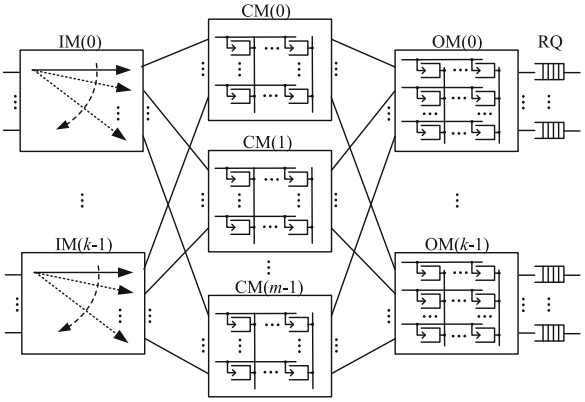
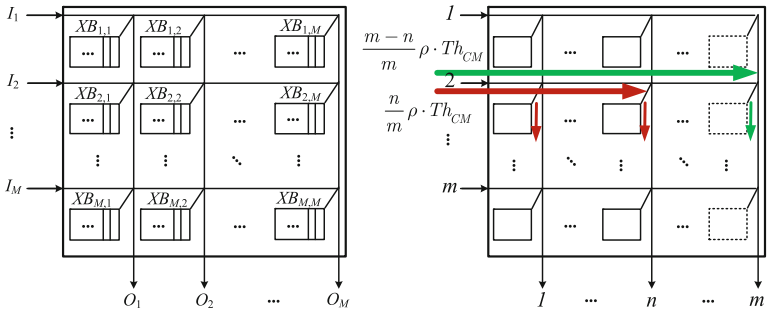


Fig. 1. The CQ-SMM Clos-network switch architecture

figuration [14] that repeats every m time slots. CM and OM employ crosspoint-queued switch [17,18] as the element of each module, and the sizes of each crosspoint buffer are B_C and B_O, respectively. Arrived cell will be dropped if the crosspoint buffer is full. Each output port of CMs and OMs adopts random or round-robin scheduling algorithms. Cells will enter the re-sequencing queues (RQs) after departing from OMs.

3 Performance Analysis

As the last two stages of CQ-SMM switch adopt CQ crossbar architecture, we can calculate the throughput of a CQ switch using an iterative calculation method [18–21] based on the Markov chain model before we deduce the throughput of the CQ-SMM switch.



(a) CQ switch architecture (b) The alternative model of output module

Fig. 2. The CQ architecture and the alternative model of output module

Figure 2(a) shows an $M \times M$ CQ switch, and each crosspoint buffer can accommodate no more than B cells. Arrivals to the CQ switch are Bernoulli i.i.d. uniform traffic with load $\rho \in [0, 1]$. Each time slot is divided into arrival and departure phases. In our case, the arrival phase is followed by the departure phase.

- **During the arrival phase**, if a cell arrives at input g and is destined to output h, it is stored in buffer $XB_{g,h}$, where $1 \leq g, h \leq M$. The cell will be dropped upon arrival if $XB_{g,h}$ is full. At each time slot, a single cell arrives at one crosspoint buffer with a probability of ρ/M.
- **During the departure phase**, the output port randomly selects a nonempty crosspoint buffer. $d(t)$ is defined as the probability that one crosspoint buffer, such as $XB_{g,h}$, is selected and served.

Let $\pi_j(t)$ be the steady-state probability that the observed crosspoint buffer $XB_{g,h}$ contains j cells at the tth time slot. $B+1$ states exist for each crosspoint buffer. Buffers are empty in the initial phase, i.e.,

$$\pi_0(0) = 1 \tag{1}$$

$$\pi_j(0) = 0, 1 \leq j \leq B \tag{2}$$

Let $p_{i,j}$ be the state transition probability, which represents the probability that the buffer state of the observed crosspoint buffer changes from $\pi_i(t-1)$ to $\pi_j(t)$, where $0 \leq p_{i,j}(t) \leq 1$, and $0 \leq i, j \leq B$. For each crosspoint buffer, at most, one cell could arrive and one cell could leave during one time slot. We can determine that $p_{i,j} = 0$, if $j > i+1$ or $j < i-1$.

When $i = 0$,

$$p_{0,0} = \frac{\rho}{M} d(t) + (1 - \frac{\rho}{M}) \tag{3}$$

$$p_{0,1} = \frac{\rho}{M}(1 - d(t)) \tag{4}$$

When $1 \leq i \leq B-1$,

$$p_{i,i-1} = (1 - \frac{\rho}{M}) d(t) \tag{5}$$

$$p_{i,i} = \frac{\rho}{M} d(t) + (1 - \frac{\rho}{M})(1 - d(t)) \tag{6}$$

$$p_{i,i+1} = (1 - \frac{\rho}{M}) d(t) \tag{7}$$

When $i = B$,

$$p_{B,B-1} = d(t) \tag{8}$$

$$p_{B,B} = 1 - d(t) \tag{9}$$

All buffers related to one output port are assumed to have the same state distribution to simplify the model of departure states. The probability that one crosspoint buffer is not empty equals $1 - \pi_0(t-1)(1-\rho/M)$. If i nonempty buffers exist, then one of them will be selected to transmit its head-of-line cell with a probability of $1/i$. The departure probability can be calculated as:

$$d(t) = \sum_{i=1}^{M} \frac{1}{i} C_{M-1}^{i-1} [1 - \pi_0(t-1)(1-\rho/M)]^{i-1} [\pi_0(t-1)(1-\rho/M)]^{M-i} \tag{10}$$

Now, based on above equations, we can calculate $\pi_j(t)$ using iterative method,

$$\pi_j(t) = \sum_{i=0}^{B} p_{i,j} \pi_i(t-1), 0 \leq i, j \leq B \tag{11}$$

Given the uniform distribution of traffic and symmetry of the switch fabric, we can study the performance of one output port instead of the entire switch fabric [16]. If all crosspoint buffers of the same output line are empty, then no cell will be scheduled. Thus, the relative throughput can be calculated as

$$Th(t) = \frac{1 - [\pi_0(t)(1-\rho/M)]^M}{\rho} \tag{12}$$

The system reaches a stable state when t tends to infinity. Consequently, throughput becomes constant, which is a function of the switch size, buffer size, and input load of CQ switch, and can be expressed as $Th = f(M, B, \rho)$.

The last two stages of the CQ-SMM switch adopt CQ switch architectures, and the input traffic of OM is only the output traffic of CM. Each output port of CM randomly selects a nonempty crosspoint buffer, and the destination ports of cells are distributed evenly among all ports. Accordingly, we assume that the traffic arriving OM is independent.

- CM is a $k \times k$ CQ switch with a buffer size of B_C. The input load of CM is balanced with an increase in CM numbers and equals $\frac{n}{m}\rho$.
- OM is an $m \times n$ CQ switch with a buffer size of B_O, and the input load equals $\frac{n}{m}\rho \cdot Th_{CM}$. Some adjustments need to be performed before the above-mentioned formula can be applied to this situation. The switch fabric of OM can be viewed as an $m \times m$ CQ switch with an adjustment input load, which is equal to the actual input load times a scale factor $\frac{m}{n}$, i.e., $\frac{m}{n} \cdot [\frac{n}{m}\rho \cdot Th_{CM}] = \rho \cdot Th_{CM}$, as shown in Fig. 2(b).
- Finally, the throughput of the CQ-SMM switch is given as follows:

$$\begin{aligned} Th_{Clos} &= Th_{CM} \cdot Th_{OM} \\ &= f(k, B_C, \frac{n}{m}\rho) \cdot f(m, B_O, \rho \cdot f(k, B_C, \frac{n}{m}\rho)) \end{aligned} \tag{13}$$

4 Simulation

We simulate the CQ-SMM switches by adopting the random scheduling algorithm and evaluate its performance under different numbers of redundant paths, buffer sizes of CM and OM, input loads, and switch sizes. Arrivals to the switch are the Bernoulli i.i.d. uniform traffic with load $\rho \in [0, 1]$. Admissible traffic, which means that none of the inputs/outputs are overload, is considered. Each simulation is performed 10 times using 10^4 time slots each time.

4.1 Comparison of Theoretical and Simulation Values

Figure 3 plots the throughput values obtained by theoretical analysis and simulation against input load under different numbers of redundant paths and CM/OM buffer sizes, where $n = k = 4$ and $B_C = B_O = B$. Throughput decreases with an increase in input load. The addition of redundant paths or scaling up of the crosspoint buffer size can improve throughput performance. The theoretical value performs well under different numbers of CMs and buffer sizes and has high accuracy. The maximum deviation equals 3.75×10^{-2} when $m = 6$, $B = 1$, and $\rho = 1$.

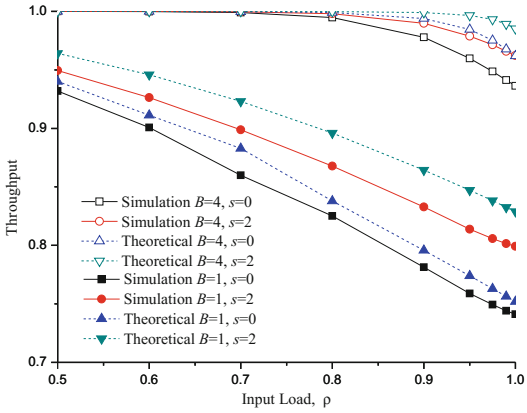


Fig. 3. Throughput of theoretical analysis and simulation as a function of input load

4.2 Different Size of Crosspoint Buffers

Figure 4 plots the throughput values and cell loss rate (CLR) of CM and OM against crosspoint buffer sizes under different numbers of redundant paths, where $n = k = 4$, and $\rho = 1$.

The throughput of the CQ-SMM switch increases with increasing buffer size of CMs and OMs. The CLR of CM falls to a low level when s increases. The theoretical deduction we proposed verifies that the input load of each CM is reduced with the increase in redundant paths, and the CQ switch can achieve high throughput at light load. The CLR of OM increases with an increase in redundant paths because the input load of OM increases and exceeds the processing capacity of OM.

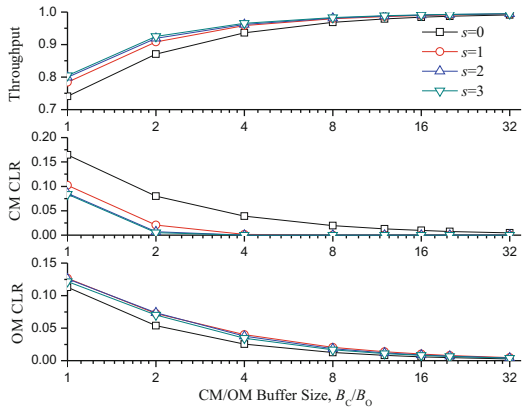


Fig. 4. The throughput and cell loss rate as a function of crosspoint buffer size

The average delays at CM, OM, and RQ are collected separately and constitute end-to-end delay, as shown in Fig. 5, where $n = k = 4$, and $\rho = 1$. When $s = 0$, i.e., $m = n = 4$, no redundant path exists. The delay performance at different modules shows a linearly increasing trend with increase in crosspoint buffer size. The increase in redundant paths leads to an obvious increase in delay in OM and RQ. The average delay in RQ indicates the extent of the out-of-sequence and the required size of the re-sequencing buffer. Compared with the end-to-end delay, the delay in CMs is negligible when redundant paths increase.

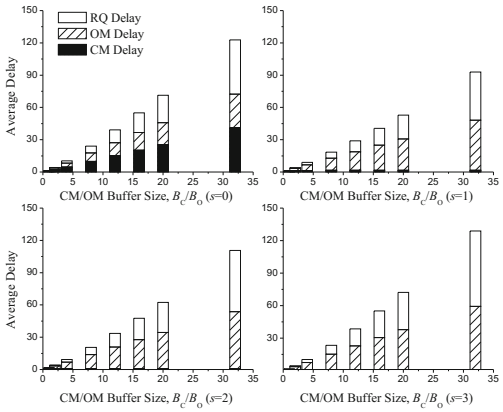


Fig. 5. The delay performance as a function of crosspoint buffer size

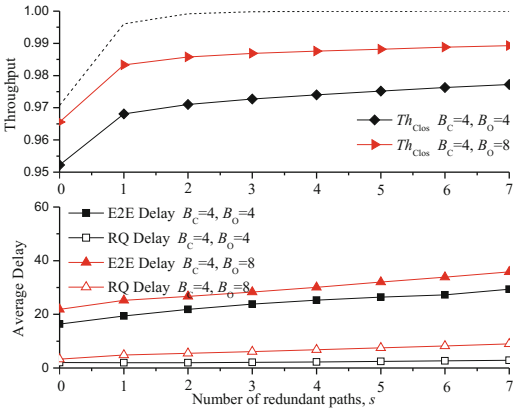


Fig. 6. Throughput and delay performance as a function of the number of redundant paths

4.3 Different Number of Redundant Paths

Figure 6 plots the throughput and average delay against the number of redundant paths under the Bernoulli i.i.d. uniform traffic of load $\rho = 1$ with $n = k = 8$ and $B_C = 4$. When the number of redundant paths increases from 0 to 7, the traffic is balanced among more CMs and the input load of each CMs is reduced. Thus, CM can obtain high throughput performance, as indicated in Fig. 6 by dotted lines. Doubling the buffer size B_O or adding another redundant path can result in nearly the same increase in throughput. Nevertheless, the former nearly doubles the size of required re-sequencing buffers, whereas the latter requires less re-sequencing buffers.

5 Conclusion

Onboard switches cannot be realized with an arbitrary size re-sequencing buffer because of environmental limitations. The extent of out-of-sequence should be controlled below a certain degree by adopting buffers with finite capacity and adding feedback control if necessary. Providing an accurate analytical model for the delay and throughput of the Clos switch is useful for us to meet the overall performance requirement when we design buffers of the on-board switch.

The main conclusions of this paper can be summarized as following:

- The throughput of Clos-network switch is not higher than that of CMs or OMs.
- Increasing redundant paths can reduce the load arrived at each CM, thereby allowing CM to achieve a high throughput with a small buffer size.
- If the throughput approaches the theoretical limit, further increasing the number of CMs or buffer sizes may improve the throughput but with rapid increase in switch overhead, such as re-sequencing buffers.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grant No. 61502204 and No. 61306047, the Fundamental Research Funds for the Central Universities under Grant No. JB140112, and the Qing Lan Project of Jiangsu.

References

1. Wirthlin, M.J.: FPGAs operating in a radiation environment: lessons learned from FPGAs in space. J. Instrum. **8**(2), 1–8 (2013)
2. Maurer, R.H., Fraeman, M.E., Martin, M.N., et al.: Harsh environments space radiation environment effects and mitigation. Johns Hopkins APL Tech. Dig. **28**(1), 17–29 (2008)
3. Siegle, F., Vladimirova, T., Ilstad, J., et al.: Mitigation of radiation effects in SRAM-based FPGAs for space applications. ACM Comput. Surv. **47**(2), 1–34 (2015)
4. Courville, N., Bischi, H., Zeng, J.: Critical issues of onboard switching in DVB-S/RCS broadband satellite networks. IEEE Wirel. Commun. **12**(5), 28–36 (2005)

5. Gilderson, J., Cherkaoui, J.: Onboard switching for ATM via satellite. IEEE Commun. Mag. **35**(7), 66–70 (1997)
6. Koishi, Y., Suzuki, Y., Takahashi, T., et al.: Research and development of 40Gbps optical free space communication from satellite/airplane. In: Proceedings of the 2011 International Conference on Space Optical Systems and Applications (ICSOS), pp. 88–92. IEEE Press, Santa Monica (2011)
7. Botta, A., Pescape, A.: On the performance of new generation satellite broadband internet services. IEEE Commun. Mag. **52**(6), 202–209 (2014)
8. Liu, K., Yan, J., Lu, J.H.: Onboard interconnected parallel Clos-network. J. Tsinghua Univ. **55**(8), 838–843 (2015)
9. Liu, K., Yan, J., Gao, X.L., Lu, J.H.: Fully distributed fault tolerant scheduling for onboard Clos-network switching. J. Electron. Inf. Technol. **38**(6), 1377–1384 (2016)
10. Zhang, M.S., Qiu, Z.L., Gao, Y., Li, J.: Reverse dispatching scheme for satellite Clos-network switches. J. Xidian Univ. **40**(4), 96–101 (2013)
11. Zhang, M.S., Qiu, Z.L., Gao, Y., Li, J.: Divide-and-conquer dispatching scheme for satellite Clos-network switches. J. Electron. Inf. Technol. **34**(11), 2734–2740 (2012)
12. Dong, Z.Q., Rojas-Cessa, R., Oki, E.: Buffered Clos-network packet switch with per-output flow queues. IET Electron. Lett. **47**(1), 32–34 (2011)
13. Gao, Y., Qiu, Z.L., Zhang, Z.: Providing 100% throughput in buffered Clos-network switches with in-sequence service. IET Commun. **8**(1), 133–139 (2014)
14. Zhang, M., Qiu, Z.L., Gao, Y.: Space-memory-memory Clos-network switches with in-sequence service. IET Commun. **8**(16), 2825–2833 (2014)
15. Kao, Y.H., Chao, H.J.: Design of a bufferless photonic Clos network-on-chip architecture. IEEE T. Comput. **63**(3), 764–776 (2014)
16. Li, X., Zhou, Z., Hamdi, M.: Space-memory-memory architecture for Clos-network packet switches. In: Proceedings of the IEEE ICC, Seoul, Korea, pp. 1031–1035 (2005)
17. Kanizo, Y., Hay, D., Keslassy, I.: The crosspoint-queued switch. In: Proccedings of the IEEE INFOCOM, Rio De Janeiro, Brazil, pp. 729–737 (2009)
18. Chen, G., Zhao, Y.J., Pei, D., Sun, Y.Q.: Analyzing the impact of buffer capacity on crosspoint-queued switch performance. J. Commun. Netw. Kor. **18**(3), 523–530 (2016)
19. Radonjic, M., Radusinovic, I., Cvorovic, J.: Iterative throughput calculation for crosspoint queued switch. IEICE Trans. Commun. **93**(12), 3635–3638 (2010)
20. Cvorovic, J., Radusinovic, I., Radonjic, M.: Buffering in crosspoint-queued switch. In: Proceedings of the 17th Telecommunications Forum (TELFOR), Belgrade, Serbia, pp. 24–26 (2009)
21. Gao, Y., Pan, W.T., Zheng, L.: Improved analytical model for performance evaluation of crosspoint-queued switch under uniform traffic. IET Netw. **6**, 81–86 (2017). https://doi.org/10.1049/iet-net.2017.0016

Big Data Network Track



Positive and Unlabeled Learning for Mobile App User and Server Interaction Prediction

Binbin Wang(✉), Ke Yu, Xiaofei Wu, Fang Wei, Wan Jiang, and Di Pan

Beijing University of Posts and Telecommunications, Beijing 100876, China
wangbb1994@gmail.com, {yuke,wuxf,weifang,jiangwan,pandi}@bupt.edu.cn

Abstract. With the fast development of mobile Internet and wireless communication technology, especially 4G communication technology widely used, people increasingly rely on a large number of mobile applications (Apps). In this paper, we propose a machine learning method to accomplish the task of mobile App user and server interaction prediction based on mobile Internet traffic. As in the real traffic flow data, which is collected by network traffic monitors from the public network of ISPs, only the connection information between users and servers is recorded, i.e. we only have positive examples and a large set of unlabeled examples. We design a PU learning (Learning from Positive and Unlabeled examples) based method to predict the interaction between mobile users and servers. We use the real traffic flow data from the most popular mobile App in China, i.e. Tencent QQ, as the experimental dataset. Experimental results illustrate that our proposed method is effective.

Keywords: PU learning · Link prediction
Mobile App user and server interaction prediction
Semi-supervised learning

1 Introduction

Due to the rapid development of mobile Internet and smartphone technology, people can easily access information and services anywhere at anytime. At the same time, mobile users spend more and more time to experience various mobile applications (Apps) on their smart devices such as smartphones and tablets. Understanding mobile user behavior, especially user behavior of mobile App usage, is significant and valuable not only for App developer, but also for Internet Service Providers (ISPs), for better network management and service provisioning.

The existing researches on mobile App usage prediction are mainly based on App usage log from App servers or sensor log from smart devices. Usually these datasets are small and individual-dependent. Actually Internet traffic data from ISPs can be explored to analyze and predict user behavior. The traffic flow data is collected from traffic monitoring systems or network routers. A single traffic

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 481–491, 2018.
https://doi.org/10.1007/978-3-319-78130-3_50

flow record is defined as one or more packets using a particular protocol sent from a specific source port and host to a specific destination port and host, over some time interval [1]. The traffic flow data is de facto big data and provides a whole view of user behavior across different applications.

In this study, we focus on mobile App usage prediction based on Internet traffic flow data. Our main task is to predict whether a mobile user will visit a specific App server based on historical traffic flow data. The detailed information in a traffic flow record include flow duration, mobile phone number or destination and source IP addresses, occurring time, total amount of bytes and packets in the flow, mobile phone brand and model, and application type, etc. How to convert these information into feature vectors to represent user and server interaction pattern effectively is a primary challenge. Moreover, as in the real traffic flow data, only the connection information between users and servers is recorded. We regard it as a positive example if the user has a connection record with the server else it's an unlabeled example, thence we only have a few positive examples and a large amount of unlabeled examples. We innovatively design a PU learning (Learning from Positive and Unlabeled examples) based method to accomplish the prediction task. We choose the traffic flow data from the most popular mobile App in China, i.e. Tencent QQ, as the experimental dataset, and experimental results illustrate that our proposed method is effective.

The rest of the paper is arranged as follows. We summarize related work on App usage prediction and PU learning in Sect. 2. And in Sect. 3, the feature selection and PU learning method are proposed. The experimental results are described in Sect. 4. Section 5 gives the conclusions.

2 Related Work

With the widely used of smartphones, mobile applications (Apps) have been developed quickly to satisfy the needs of users. Many researches have focused on the mobile App user and server interaction prediction task, i.e. App usage prediction problems. In [2], the authors discovered two kinds of features by analyzing the log data of App usage: the Implicit Feature from App usage relations, and the Explicit Feature from sensing readings of built-in sensors. And a personalized feature selection algorithm was proposed to furthermore predict Apps usage. In [3], the authors developed an AppNow widget based on App usage history, which is able to predict users' Apps usage. [4] collected a variety of contextual information, including the user profile, last used App, location, time, and so on, to predict the user's App usage. [5] collected 37 sensor reading as features to finish App usage prediction. As we can see, those researches on mobile App usage prediction are mainly based on App usage log from App servers and sensor log from smart devices, usually these datasets are small and individual-dependent. In this paper, the real traffic flow data will be used to predict App usage.

PU learning has been studied in the past years to settle the situation lack of negative training data. Originally, it was proposed to complete the task of

text classification [6,7]. Subsequently, it was introduced to other classification learning areas. In [8], PU learning was used for graph classification. In [9], the authors proposed a PU learning based algorithm called LELC for data stream classification. In [10], to detect deceptive reviews, a semi-supervised model called MPIPUL based on Latent Dirichlet Allocation (LDA) and SVM was proposed. [11] designed a novel PU learning algorithm PUDI for disease gene identification.

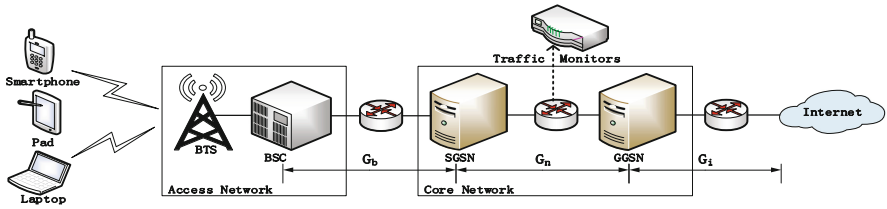


Fig. 1. Traffic flow data collection process

The existing PU learning method can be divided into three categories [12]. The first category of methods adopts a two-step strategy, identifying several reliable negative examples from unlabeled dataset, and then applying classification algorithms such as EM, Naive Bayes and SVM based on the positive examples and reliable negative examples to solve the classification problem [6,13–16]. The second category of methods estimates statistical queries over positive and unlabeled examples [17–19]. The third category of methods simplifies this classification task to the task of learning with high one-sided noise by regarding the unlabeled examples as noisy negative examples [6,7,20]. In this paper, we design a two-step strategy based PU learning method to accomplish the task of mobile App usage prediction.

3 Proposed Prediction Method

3.1 Dataset and Feature Selection

In our study, the Internet traffic flow data was collected by specific high-performance traffic monitors in an ISP's operational mobile network, which provides services for a large number of users in a city of southern China. As shown in Fig. 1, the traffic monitors are deployed between GGSN (Gateway GPRS Support Node) and SGSN (Serving GPRS Support Node), to collect packets at different network interfaces, such as G_i, G_b, and G_n. In addition to supporting line-speed packet parsing and accurate flow composition, the traffic monitors provide real-time traffic classification by combining DFI (deep flow inspection), DPI (deep packet inspection), and port-based application deduction technologies.

The traffic flow data used in this paper was gathered within a 96-h period on December, 2013. There were over 1.5 billion traffic flows recorded each 24-h period (0:00 am 11:59 pm). The detailed entries in each flow record include

mobile phone number or source and destination IP addresses, occurring time, flow duration, application type, total amount of bytes and packets in the flow, and mobile phone brand and model, etc. For application type, there are more than 15 categories and more than 100 sub-categories identified, including: Web, Search, Advertising, E-commerce, Music, Video, Game and so on. For mobile phone, there are more than 20 brands and more than 100 models identified, such as Apple, Samsung, and Huawei, which are popular in Chinese market.

Based on the traffic flow data collected, we can use a feature vector consisting of two parts to represent the connection relationship between the user and server. The first part of the vector is some of the attributes associated with the user, such as the first three digits of the user's phone number (i.e. 139 and 138 belonging to China Mobile, 133 and 189 belonging to China Telecom, 130 and 131 belonging to China Unicom), terminal brand, user active time by hour, connection duration, etc. The latter part is some properties associated with the server, including server active time by hour, average number of connections, service category, etc. The composition of the feature vector is shown in Table 1 in detail, including 8 categories of user-related information and 8 categories of server-related information. The total dimensions of the feature vector is 194.

Table 1. The selected features

	Feature
User related	The first three digits
	User's terminal brand
	Active time by hour
	Number of connection types
	Number of connections
	Connection duration
	Average up packets
	Average down packets
Server related	Number of connection types
	Service category
	Service sub-category
	Active time by hour
	Number of connections
	Connection duration
	Average up packets
	Average down packets

3.2 PU Prediction Method

By using the above feature vectors, we propose a PU learning method based on two-step strategy to predict mobile App user and server interaction.

Step 1: extracting a number of reliable negative examples from the unlabeled set based on spy dataset.
Step 2: building a classifier based on positive and reliable negative examples.

At first, we randomly select a set S of positive examples from all positive data (P) and put them in unlabeled data set (U). Then, we train a binary classifier based on $(P - S)$ and $(U + S)$, and classify each example in $(U + S)$, evaluate a probability threshold t using S, assign every example in U with the prediction probability greater than t as reliable negative examples (RN). The basic idea of identifying reliable negative examples from unlabeled data is shown in Fig. 2. Finally, we build another classifier using P and RN to predict whether a user will connect to the server in the future. The detailed algorithm for extracting those most likely negative examples (RN) and predicting the future connection is given in Algorithm 1. Among them, the classifier has many options, such as Logistic Regression, Naive Bayes, Decision Tree, SVM, Random Forest, etc. Different classifiers have a important influence on the experimental results. In the later experiments, we chose three of them (Logistic Regression, Decision Tree and Random Forest) to do the experiment separately.

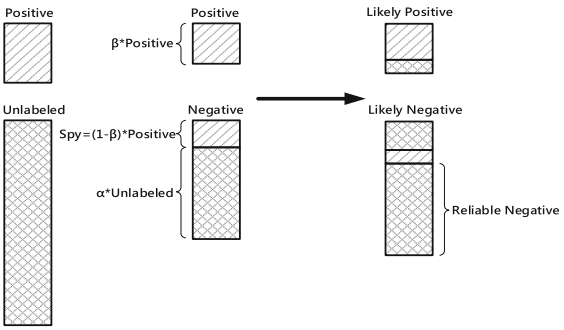


Fig. 2. Extracting reliable negative examples from unlabeled data

We now elaborate how to evaluate the threshold t. Let the set of spies S be $\{s_1, s_2, s_3, \cdots, s_k\}$, and the probabilistic label c_0 assigned to s_i be $P(c_0|s_i)$, where c_0 is the negative class. We can use the maximum probability of S predicted to negative as the threshold value t, i.e.,

$$t = max\{P(c_0|s_1), P(c_0|s_2), P(c_0|s_3), \cdots, P(c_0|s_k)\},$$

which indicates that the unlabeled examples with probability predicting to be negative class greater than t are regarded as reliable negative examples.

Algorithm 1. PU-based App Interaction Prediction Algorithm

Input:

The set of positive examples, P;
The set of unlabeled examples, U;
Test set, T;
α;
β

Output:

Prediction Label Vector FL, $\forall u \in T$;

initializing Reliable Negative Set $RN = \emptyset$;
calculating Unlabeled Subset $US = Sample(U, \alpha)$;
calculating Spy Set $SP = Sample(P, \beta)$;
assigning each example in $P - SP$ the label +1;
assigning each example in $US \cup SP$ the label -1;
producing a classifier C_1 using $P - SP$ and $US \cup SP$;
classifying each $u \in US \cup SP$ using C_1;
deciding a probability threshold t using SP;
for each $u \in US$ **do**
    predicting its probability $p(-1|u)$ using classifier C_1;
    if its probability $p(-1|u) > t$ **then**
        $RN = RN \cup u$
    end if
end for
producing a classifier C_2 using P and RN;
classifying each $u \in T$ using C_2;
return Prediction Label Vector FL, $\forall u \in T$

4 Experiment Results

4.1 Data Preprocessing

Tencent QQ is one of the most popular instant messaging applications in China. QQ also offers users a variety of services, including music, online social games, micro blogging, shopping, movies, and group and voice chat. We collected 4-day traffic flow data in December 2013 from the real network, about billions of records. We constructed Spark-based traffic data processing platform to analyze such a large amount of data. We distilled the flow data related to QQ, and randomly selected 5000 users and extracted all records related to the 5000 users. This forms the experimental dataset which we will analyze later. The dataset includes about 2.2 million records and 1684 distinct servers. Based on the filtered data, we use the feature selection method proposed in Sect. 3. A to calculate the relevant characteristics of every user and server, and generate a feature vector for each user and server pair. To predict QQ user and server connection relationship, we use the data of the first three days as the training data, and the fourth day as the testing data. The experimental dataset is summarized in Table 2.

Table 2. Experimental QQ dataset

	Day1+2+3 for training	Day4 for testing
Number of user nodes	5000	5000
Number of server nodes	1684	1684
Number of edges	8420000	8420000
Positive examples	235586	77835
Unlabeled examples	8268837	8342165

4.2 Evaluation Measures

We use the F score to measure the performance of our proposed method. F score has been widely used to estimate the performance of PU learning algorithms in the literature [6,9,13,21]. It is defined as, $F = 2pr/(p + r)$, where r is the recall, and p is the precision. F score reports an average impact of both recall and precision. The F score will be small if either of them is small. The F score will be large only when both of them are large. As we want to identify positive examples, this is suitable for our purpose, and it is undesirable to have either too small recall or too small precision.

We also report accuracies in our experimental results. The accuracy is defined as, the ratio of the amount of examples that are correctly classified to the total amount of examples. However, it should be pointed out that accuracy does not completely report the performance of our algorithm, as our datasets has a large percentage of negative examples. We believe this reflects realistic situations. Under such cases, the value of accuracy can be large, but a small amount of positive examples may be correctly classified.

4.3 Experimental Results

For all user and server pairs, only a few will have connection records in the real network, i.e. there is only a few labeled positive examples. The proportion of labeled positive and unlabeled examples is very unbalanced, which can have a significant impact on the subsequent training and forecasting process. In this experiment, we performed a random sampling of the unlabeled data, the sampling ratio is α. We studied the effects of α values on experimental results, i.e. $\alpha = 0.05$, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35 and 0.40. The results in terms of F score and accuracy are shown in Figs. 3, 4 and 5 and Table 3.

To hunt for reliable negative examples, we will sample a spy set from training positive set, the sampling ratio is $1 - \beta$, i.e. $1 - \beta$ of the training positive set is added to the unlabeled dataset. We studied the effects of different β values (from 0.2 to 0.9) on experimental results. The result in terms of F score and accuracy is shown in Fig. 6 and Table 4.

We also used a normal method as baseline, in which unlabeled examples are regarded as negative directly, put together with the exist positive examples to

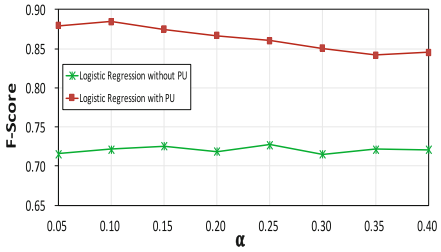


Fig. 3. The effect of parameter α with Logistic Regression classifier ($\beta = 0.8$)

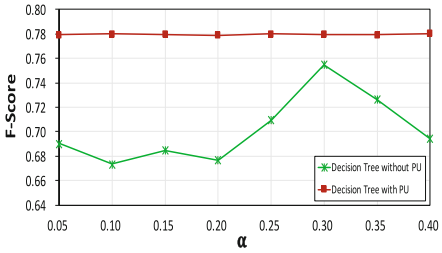


Fig. 4. The effect of parameter α with Decision Tree classifier ($\beta = 0.8$)

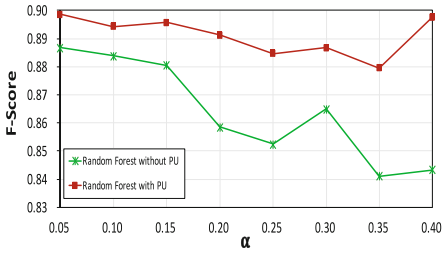


Fig. 5. The effect of parameter α with Random Forest classifier ($\beta = 0.8$)

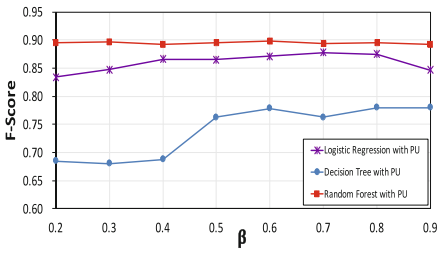


Fig. 6. The effect of parameter β ($\alpha = 0.15$)

train a traditional binary classifier to do the prediction task. We compared the baseline method without PU learning and our proposed PU-based prediction algorithm. The results in Figs. 3, 4 and 5 shown that our proposed algorithm is superior.

In the PU learning algorithm, we need to use some classifiers. The choice of the classifiers is critical and can have a great impact on the final result. We compared the effect of 3 common classifiers (Logistic Regression, Decision Tree and Random Forest) on the algorithm performance in Figs. 3, 4, 5 and 6.

From the results in Figs. 3, 4, 5 and 6 and Tables 3 and 4, we can draw the following conclusions:

1. The PU prediction algorithm produces a preferable result on the task of mobile App user and server interaction prediction. Regardless of which classifier was used in the algorithm, the proposed PU prediction algorithm improved the final results. This reflects the effectiveness of using PU learning method in App user and server interaction prediction.
2. The classifier is very important to our proposed algorithm. Different classifiers have a significant effect on the prediction results. We can see from the experimental results, Random Forest is the best choice in our prediction task.

Table 3. Accuracy of different α and different method

α value	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
Logistic Regression without PU	0.6187	0.6310	0.6438	0.6292	0.6479	0.6248	0.6393	0.6384
Logistic Regression with PU	0.8804	0.8854	0.8734	0.8619	0.8541	0.8404	0.8285	0.8337
Decision Tree without PU	0.5604	0.5250	0.5522	0.5352	0.6009	0.6844	0.6352	0.5704
Decision Tree with PU	0.7222	0.7234	0.7227	0.7218	0.7236	0.7229	0.7221	0.7241
Random Forest without PU	0.8797	0.8798	0.8854	0.8711	0.8671	0.8759	0.8575	0.8587
Random Forest with PU	0.8957	0.8911	0.8910	0.8849	0.8767	0.8787	0.8695	0.8941

Table 4. Accuracy of different β and different method

β value	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Logistic Regression with PU	0.8478	0.8576	0.8718	0.8702	0.8757	0.8806	0.8734	0.8347
Decision Tree with PU	0.5486	0.5397	0.5553	0.6946	0.7205	0.6947	0.7227	0.7227
Random Forest with PU	0.8914	0.8937	0.8875	0.8912	0.8942	0.8893	0.8910	0.8883

3. In our PU prediction algorithm, the parameters of α and β have little effect on the experimental results. This may be due to the fact that the positive and unlabeled examples in the original dataset are unbalanced, and we have taken sampling from the unlabeled dataset to balance the proportion.

5 Conclusion

This paper studied the problem of mobile App user and server interaction prediction. As we only have positive examples, we propose a PU learning based method to solve this problem. Based on real traffic data, we describe how to choose proper features to represent the connection relationship of users and servers in detail. And we design a PU prediction algorithm based on two-step strategy to forecast this connection. Finally, we use real QQ traffic flow data to verify the effectiveness of the algorithm. Experimental results indicate that our proposed PU prediction method is better than using classifier directly judged by F score and accuracy.

Acknowledgment. This work is supported by the National Natural Science Foundation of China under Grant No. 61601046 and No. 61171098, and is partially supported by the 111 Project of China under Grant No. B08004, and EU FP7 IRSES Mobile Cloud Project under Grant No. 612212.

References

1. Quittek, J., Zseby, T., Claise, B., Zander, S.: Requirements for IP flow information export (IPFIX), Technical report (2004)
2. Liao, Z.X., Li, S.C., Peng, W.C., Philip, S.Y., Liu, T.C.: On the feature discovery for app usage prediction in smartphones. In: 2013 IEEE 13th International Conference on Data Mining (ICDM), pp. 1127–1132. IEEE (2013)
3. Liao, Z.X., Lei, P.R., Shen, T.J., Li, S.C., Peng, W.C.: Mining temporal profiles of mobile applications for usage prediction. In: 2012 IEEE 12th International Conference on Data Mining Workshops (ICDMW), pp. 890–893. IEEE (2012)
4. Huang, K., Zhang, C., Ma, X., Chen, G.: Predicting mobile application usage using contextual information. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 1059–1065. ACM (2012)
5. Shin, C., Hong, J.H., Dey, A.K.: Understanding and prediction of mobile application usage for smart phones. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 173–182. ACM (2012)
6. Liu, B., Dai, Y., Li, X., Lee, W.S., Yu, P.S.: Building text classifiers using positive and unlabeled examples. In: Third IEEE International Conference on Data Mining, ICDM 2003, pp. 179–186. IEEE (2003)
7. Lee, W.S., Liu, B.: Learning with positive and unlabeled examples using weighted logistic regression. In: ICML, vol. 3, pp. 448–455 (2003)
8. Zhao, Y., Kong, X., Philip, S.Y.: Positive and unlabeled learning for graph classification. In: 2011 IEEE 11th International Conference on Data Mining (ICDM), pp. 962–971. IEEE (2011)
9. Li, X.L., Yu, P.S., Liu, B., Ng, S.K.: Positive unlabeled learning for data stream classification. In: Proceedings of the 2009 SIAM International Conference on Data Mining, pp. 259–270. SIAM (2009)
10. Ren, Y., Ji, D., Zhang, H.: Positive unlabeled learning for deceptive reviews detection. In: EMNLP, pp. 488–498 (2014)
11. Yang, P., Li, X.L., Mei, J.P., Kwoh, C.K., Ng, S.K.: Positive-unlabeled learning for disease gene identification. Bioinformatics **28**(20), 2640 (2012)
12. Zhang, B., Zuo, W.: Learning from positive and unlabeled examples: a survey. In: 2008 International Symposiums on Information Processing (ISIP), pp. 650–654. IEEE (2008)
13. Liu, B., Lee, W.S., Yu, P.S., Li, X.: Partially supervised classification of text documents. In: ICML, vol. 2, pp. 387–394. Citeseer (2002)
14. Yu, H., Han, J., Chang, K.C.C.: PEBL: positive example based learning for web page classification using SVM. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 239–248. ACM (2002)
15. Li, X., Liu, B.: Learning to classify texts using positive and unlabeled data. In: IJCAI, vol. 3, pp. 587–592 (2003)
16. Fung, G.P.C., Yu, J.X., Lu, H., Yu, P.S.: Text classification without negative examples revisit. IEEE Trans. Knowl. Data Eng. **18**(1), 6–20 (2006)
17. Denis, F., Gilleron, R., Tommasi, M.: Text classification from positive and unlabeled examples. In: Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2002, pp. 1927–1934 (2002)
18. Denis, F., Laurent, A., Gilleron, R., Tommasi, M.: Text classification and co-training from positive and unlabeled examples. In: Proceedings of the ICML 2003 Workshop: The Continuum from Labeled to Unlabeled Data, pp. 80–87 (2003)

19. Blum, A. Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory, pp. 92–100. ACM (1998)
20. Zhang, D., Lee, W.S.: A simple probabilistic approach to learning from positive and unlabeled examples. In: Proceedings of the 5th Annual UK Workshop on Computational Intelligence (UKCI), pp. 83–87 (2005)
21. He, J., Zhang, Y., Li, X., Wang, Y.: Naive Bayes classifier for positive unlabeled learning with uncertainty. In: Proceedings of the 2010 SIAM International Conference on Data Mining, pp. 361–372. SIAM (2010)



Game-Based Mobile Traffic Sharing with WiFi

Rong Wang, Zhifei Wang, Jibing Wu, Hongbin Huang(✉), Su Deng, and Yahui Wu

Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Deya Road 109, Changsha 410073, China
wangzhifeicn@163.com,
{wangzhifei,hbhuang,wuyahui}@nudt.edu.cn,
wujibing_2007@126.com, sudeng@sohu.com

Abstract. Mobile data traffic is important for the remote device control in Internet of Things (IoT). Due to heterogeneous usage and subscribed data caps, mobile traffic sharing is attractive for the users. This paper mainly studies how the users share the leftover traffic efficiently. In particular, it presents a framework that allows the users to share their traffic based on WiFi. The framework can collect all the users' bids, and then the sellers allocate their traffic according to the buyers' bids. In such scenario, the competition between the buyers exists, so how to bid efficiently is critical for the buyers. This paper models the problem as the non-cooperative game. According to the number of sellers, it considers two different cases. The existence of the Nash Equilibrium in both cases is mathematically proved.

Keywords: Traffic sharing · WiFi · Non-cooperative game

1 Introduction

The mobile communication is one important channel to transmit the information, which is used to remotely control various devices in IoT. Combing with the consumption in online content and mobile commerce, the users' mobile data traffic increase rapidly. To decrease the expenses, the users often subscribe the monthly data cap from the ISPs, such as the China Mobile, AT&T, etc. [1]. However, the users' data usage may be heterogeneous, and may not fit their data cap very well. In this case, some users may need extra data, and have to pay overage fees, which are expensive. Correspondingly, the other users may have leftover traffic, which will be cancelled in the next month. Therefore, if these users can share the leftover traffic with others by charging certain fees, it may be good for all of them. At present, some ISPs have presented the platforms for the users to buy/sell their traffic. From example, the China Mobile proposes the 2CM project that allows the users to share their traffic [2]. The AT&T proposes similar platform in USA [1]. In such platforms, the sellers/buyers will show the price of the traffic they want to sell/buy, and the ISPSs will act as a middle man to match the buyers and sellers based on their price. However, the ISPs do not introduce any guidance for the users to submit their bids, so the users do not know how to bid efficiently. Based on

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 492–501, 2018.
https://doi.org/10.1007/978-3-319-78130-3_51

the 2CM platform, the work [3] proposes a bidding policy for the users. In [3], if there is a seller whose price is below the buyer's bids, the buyer can buy the traffic successfully. However, if such seller does not exist, the buyer cannot get the additional traffic at all. Sen et al. introduces a framework to understand the trade-offs in choosing between individual data caps and shared data plans [4]. The work [5] mathematically explores whether the users have the incentive to share their traffic.

Though above works solve many problems in the traffic sharing process, there are still some challenges. First, the platforms established by the ISPs do not exist in most areas. For example, the consumers in the Chinese mainland cannot use the 2CM project at all, but the demand for traffic sharing is more common. Second, the buyers' optimal biding policy closely depends on the sellers' allocation policy. The work [3] proposes a *threshold*-based allocation policy, which is similar to the Amazon's resource allocation policy [6]. As shown in next section, such policy may not be proper in some scenarios. Third, state of the art works do not consider the competition between the buyers efficiently. To overcome the first challenge, the work [7] presents a platform that allows users (buyers or primary users, PUs) to buy the data from others (sellers or assistant users, AUs) indirectly. The platform is based on WiFi, where the AUs download data from the Internet and forward it to PUs though WiFi. The platform can be used in small areas very well, such as the official building, the students' dormitory, etc. However, it just considers one buyer, so it does not consider the competition between the buyers.

Based on the existing works, we present a new platform. Different from [7], in our platform, the traffic is shared directly. Note that in [7], the PUs download the data and then forward the data to AUs, so the traffic is shared indirectly. This strategy will bring two problems. First, the downloading process will occupy certain resources of PUs and may have certain impact on other applications of the PUs. Second, the privacy of the AUs may be obtained by PUs. For this reason, in our platform, the seller set themselves as the WiFi hotspot. Therefore, if the buyers connect the hotspot built by the sellers, the buyers can share the sellers' traffic directly. In the platform, the sellers can control the volume of the traffic that one specific buyer can get through controlling the buyer's connecting time. Based on such platform, we propose the proportional allocation policy, where the sellers allocate their leftover traffic according to the buyers' bids, so the competition between the buyers exists. We model such problem as a non-cooperative game. In particular, we first consider the case with only one seller. Then, we explore the case with multiple sellers. In this case, the buyers not only decide their total bids, but also need to consider how to use the bids efficiently, so the problem is more complex. Through mathematical analysis, we prove that the Nash Equilibrium exists in both cases. The theoretical results in this paper provide a guideline for the buyers to bid for the traffic efficiently. Moreover, our work can be used to extend the work [7] to the case where multiple buyers exist.

We would like to draw your attention to the fact that it is not possible to modify a paper in any way, once it has been published. This applies to both the printed book and the online version of the publication. Every detail, including the order of the names of the authors, should be checked before the paper is sent to the Volume Editors.

2 Related Works

The works [1–5, 7] have explored the traffic sharing problem. Our work is also similar to the resource sharing in the cloud, which is very hot recently. At present, many works explore such problem, and these works can be divided into two categories. The first category mainly considers the providers' charging policy to maximize their final revenue [8]. The last category tries to study the optimal biding policy of the users [6], which is much closer to our work. However, these works are still different from ours. First, the resources in the cloud are different from the traffic. For example, the leftover traffic will be cancelled in the next month, and this phenomenon may encourage the sellers to sell more traffic. Second, the users' biding policy depends on the resources' allocation policies, which are heterogeneous. For example, the Amazon provides three different policies: reserved, on-demand and spot. In this paper, we propose a new allocation policy, where the traffic the users can obtain is proportional to their bids. The new policy has certain advantage. For example, there are two buyers (with demand a, b) and one seller (with traffic c, $a > c$, $b > c$). In the Amazon's allocation policies, at most one buyer can get the traffic from the seller, so the seller just gets one buyers' bid. In the new policy, the seller can receive both buyers' bid and allocate the traffic proportionally. In addition, both buyers can get certain traffic with the price lower than the ISP's. Therefore, the new policy may be good for all the users.

Note that our work is different from the popular mobile traffic offloading problem, which tries to let the users access the internet with WiFi instead of the traditional 3G, 4G, etc. [9, 10]. In the traffic sharing problem, the users still use the traditional communication method through the sellers.

3 Sharing Framework

We first present the sharing framework. We suppose that there are N buyers and M sellers. The volume of the traffic that the j-th seller wants to sell is denoted as C_j. The basic sharing framework is shown in Fig. 1.

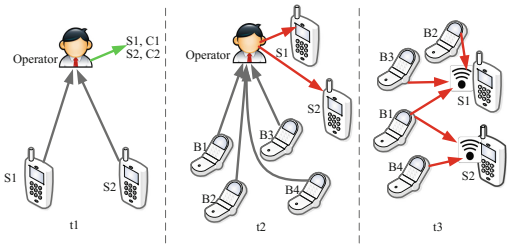


Fig. 1. The traffic sharing framework

In the framework, there is an operator, who can collect the users' bids. For example, in the time *t1*, the sellers *S1* and *S2* submit their bids *C1* and *C2* to the operator. Then,

the operator publishes the bids. Based on the sellers' bids, the buyers submit their bids in time *t2*. The bids of the buyers mainly include the price they are willing to pay for each seller. In time *t3*, the operator will tell the buyers' bids to the sellers *S1* and *S2*. Next, the sellers will set themselves as the hotspots to serve the buyers according to the received bids. Once the buyers connect the hotspot, they can share the sellers' traffic. In addition, the sellers can control the volume of the traffic the buyers obtain by controlling the connecting time. In such framework, many challenges arise. In this paper, we just consider how to bid efficiently for the buyers.

The buyers gain certain utility after buying additional traffic. Generally speaking, the utility increases with the traffic. However, the increasing gradient may decrease when the traffic increases. Such phenomenon can be denoted by the famous α-fair utility function, which has been used in [3]. Then, if a buyer gain c volume of traffic, its utility can be denoted as

$$V(c) = \theta c^{1-\alpha}/(1-\alpha) \tag{1}$$

In (1), θ is a positive constant which represents the scale of the utility. α is also a constant, whose value belongs to [0, 1). For different buyers, the values of θ and α may be different. For simplicity, this paper assumes that they are the same for different buyers. However, the work reported in this paper can be easily extended to the case without such assumption. Then, based on (1), we can get the payoffs of the j-th buyer as follows,

$$U_j = \varepsilon_j V(c_j) - b_j = \varepsilon_j \theta c_j^{1-\alpha}/(1-\alpha) - b_j \tag{2}$$

b_j denotes the bids of the j-th buyer. ε_j is a positive constant and is used to regulate the weight of the utility and cost.

When $M = 1$, there is only one seller, so the buyers does not need to allocate their bids for different sellers. However, when there are multiple sellers, the buyers must consider above problem, and the problem is more complex. Due to the limitation of the space, this paper only considers the case where $M = 2$, which can be easily extended to the case with $M > 2$.

4 Mathematical Analysis

4.1 Game Modeling when $M = 1$

The game can be described as follows.

1. *Players*: the N buyers.
2. *Actions*: the actions of the players are their bids. As shown above, the j-th buyer's bids are b_j. Since the players will not bid arbitrarily high values, we can define $b_{min} \leq b_j \leq b_{max}$. b_{min} and b_{max} as the lower and upper bounds of the buyers' bids.
3. *Allocation Policy*: The leftover traffic is allocated according to the bids of the buyers. In particular, we define

$$c_j = Cb_j / \sum_{i=1}^{N} b_i \tag{3}$$

The c_j denotes the volume of traffic the j-th buyer gets. Because there is only one seller, we use C to denote the volume of traffic that it wants to sell. Note that if the sum of all the buyers' bids equals to 0, none of the buyers can get additional traffic. Then, we assume $b_{min} > 0$, and above problem will not appear. Our work can be used without such assumption.

4. Payoffs: the payoffs of the buyers are shown in (2).

4.2 Nash Equilibrium when $M = 1$

The bids of the players belong to $[b_{min}, b_{max}]$, and $[b_1, b_2, \ldots, b_N]$ belong to $[b_{min}, b_{max}]^N$ forms the strategy space. Therefore, the players form a continuous pure strategy game [11]. For such game, we can get Theorem 1.

Theorem 1: The Nash Equilibrium of the game exists.

Proof: First, we can see that the set of the pure strategies is compact and convex. Then, according to Theorem 2.2 in [11], the sufficient conditions for the existence of Nash Equilibrium in the continuous pure strategy game are: for any player j, if we fix the bids of other players, the payoffs of j should be quasi-concave with respect to b_j. The conditions are shown and proved in Lemma 1. Therefore, the Nash Equilibrium exists. ■

Lemma 1: When the bids of other players are fixed, U_j is quasi-concave with respect to b_j, $1 \leq j \leq N$.

Proof: The function U_j is quasi-concave when it satisfies one of the following conditions: (i) non-decreasing, (ii) non-increasing, (iii) there exists b_{j*}, where U_j is non-decreasing for $b_j < b_{j*}$ and non-increasing for $b_j > b_{j*}$.

Based on (2), we have

$$\dot{U}_j = (U_j + b_j)(1/b_j - 1/\sum_{i=1}^{N} b_i) - 1 \tag{4}$$

$$\ddot{U}_j = -2(U_j + b_j)(1/b_j - 1/\sum_{i=1}^{N} b_i)(1/\sum_{i=1}^{N} b_i) < 0 \tag{5}$$

Therefore, the function (4) decreases with b_j. Then, we can easily find that the following formula holds.

$$\begin{aligned} &\dot{U}_j(b_{\max}) \geq 0, U_j \textit{ satisfies condition } i) \\ &\dot{U}_j(b_{\min}) \leq 0, U_j \textit{ satisfies condition } ii) \\ &\textit{others}, U_j \textit{ satisfies condition } iii) \end{aligned} \tag{6}$$

This shows that above conditions hold. Therefore, U_j is quasi-concave, and Lemma 1 follows. ■

When the condition (i) holds, b_{max} is the best policies. If condition (ii) holds, U_j non-decreases, so the player will bid the lowest price. In case (iii), the best policy can be obtained by letting the function (4) be zero.

4.3 Game Modeling when $M = 2$

When there are two sellers, a specific buyer can buy the traffic from both sellers. Then, its bids will include two parts. We let b_{j1} and b_{j2} denote the bids of the j-th buyer for the first and second sellers, respectively. Then, we have $b_j = b_{j1} + b_{j2}$. The volume of traffic that this buyer can get satisfies,

$$c_j = \sum_{k=1}^{M} (b_{jk} C_k / \sum_{i=1}^{N} b_{ik}), M = 2 \tag{7}$$

4.4 Nash Equilibrium when $M = 2$

Because the strategy space remains unchanged, so the game is still a continuous pure strategy game. Therefore, if the payoffs are quasi-concave, the Nash Equilibrium exists.

The problem can be divided into two different cases. In *case 1*, the bids b_{j1} and b_{j2} are independent. This case can be used in the scenario where b_{max} is very big (or very small), and the sum of b_{j1} and b_{j2} at the best policy is not more than b_{max} (or smaller than b_{min}). In *case 2*, the sum of b_{j1} and b_{j2} at the best policy is bigger than b_{max} (or smaller than b_{min}). Therefore, b_{j1} and b_{j2} in this case are dependent. Then, the best policy should satisfy $b_{j1} + b_{j2} = b_{max}$ (or b_{min}). Generally speaking, the upper bound b_{max} is very big, and should not constraint the buyers to get more utilities. This means that if the optimal biding policy satisfies $b_{j1} + b_{j2} > b_{max}$, the buyer j will increase the upper bound and still adopt the optimal policy. For this reason, we can just consider the *case 1*. In such case, we have

$$\frac{\partial c_j}{\partial b_{j1}} = (c_j - b_{j2} C_2 / \sum_{i=1}^{N} b_{i2})(1/b_{j1} - 1/\sum_{i=1}^{N} b_{i1}) \tag{8}$$

Where j belongs to [1, 2]. The derivative about b_{j2} can be got by the same way. Then, Eq. (9) follows.

$$\frac{\partial U_j}{\partial b_{j1}} = \varepsilon_j \frac{\theta}{1-\alpha} c_j^{-\alpha} (c_j - \frac{b_{j2} C_2}{\sum_{i=1}^{N} b_{i2}})(1/b_{j1} - \frac{1}{\sum_{i=1}^{N} b_{i1}}) - 1 \tag{9}$$

Furthermore, we can get the following formula.

$$\frac{\partial (U_j / b_{j1})}{\partial b_{j1}} \leq 0 \tag{10}$$

Then, we can get similar result as (6). Therefore, U_j is quasi-concave with respect to b_{j1}. By the same method, we can prove that the function is also quasi-concave with

respect to b_{j2}. This demonstrates that the Nash Equilibrium exists and can be got easily based on (9) and (10).

When $M > 2$, we can get similar equations to (8) and (9). Furthermore, we can prove that the partial differential of U_j is quasi-concave. Then, the Nash Equilibrium still exists.

5 Performance Evaluation

We develop an app of the framework based on Android OS, and deploy it on 20 classmates in our laboratory. Based on the app, each user can submit their bids at any time. After receiving a bid, we act as the operator and update the publishing information. Both the buyers and sellers may adjust the next biding policy based on such information. However, this paper just considers the buyers' policy. Note that the buyers that buy the traffic from different sellers are independent. Therefore, we can consider two specific buyers and suppose that both are willing to buy traffic from a seller j with the bid $C = 1$. Without loss of generality, we set $\theta = 1$, $\varepsilon_1 = 10$, $\varepsilon_2 = 20$ and $\alpha = 0.4$. The lower and upper bound of the bids satisfies: $b_{min} = 1$, $b_{max} = 10$. Then, we explore the final payoffs of the buyers. Due to the limitation of space, we just use the function U_1 as an example, and we set $b_2 = b_{min}$ and $b_2 = b_{max}$. Next, Fig. 2 can be obtained.

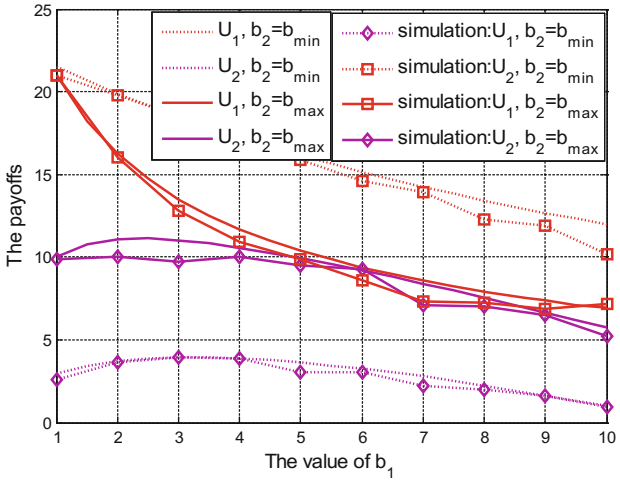


Fig. 2. The changing rule of the payoffs with different bids.

The result in Fig. 2 shows that both the payoffs of the two players are quasi-concave, and this is consistent with our theoretical results. In addition, we can find that the payoffs of the second buyer also change with b_1, and this means that one buyer's payoffs depend on others' bids. On the other hand, the deviations between the simulation and theoretical results are very small (not bigger than 5.688%), so our model is accurate. In fact, the users may move, so the WiFi connection may be disrupted. For this reason, a buyer may not use all the bought traffic, and then the deviations arise.

Now, we want to explore how the Nash Equilibrium changes with the system parameters. Here, we just consider ε_1 as an example, and let ε_1 increase from 1 to 20. Other parameters remain unchanged. Then, we get Fig. 3.

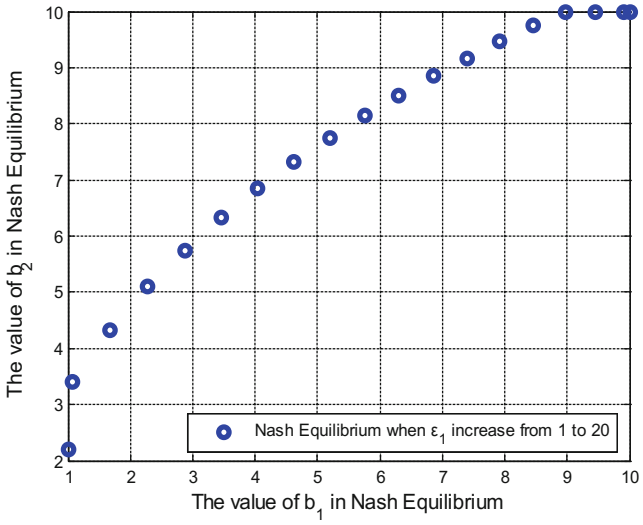


Fig. 3. The bids in Nash Equilibrium ($M = 1$)

The result shows that both the bids at the Nash Equilibrium increase with ε_1. This is because when ε_1 is bigger, the first buyer can get more rewards by buying additional traffic. Then, this buyer may increase its bids with the hope of getting more traffic. However, the leftover traffic in the market is limited, so the second buyer has to increase its bids, too. Otherwise, the traffic it can get will decrease, and this further decreases its final payoffs. Note that in the Nash Equilibrium, both the bids are positive, so the seller can get positive rewards by selling its leftover traffic. This means that the traffic sharing is good for both users.

Then, we begin to consider the scenario where $M = 2$. When ε_1 increases from 1 to 20, we can obtain Fig. 4.

The result shows that both the bids change with ε_1. Therefore, if the weight of the utilities by buying additional traffic is different, the bidding strategy will be different, too. We also can find that the total bids of the first buyer are smaller than the second buyer's. In fact, we set $\varepsilon_1 > \varepsilon_2$, this phenomenon demonstrates that if the weight is bigger, the buyers may have the incentive to give more prices. In addition, we can see that b_{22} first increases and then decreases. This is because that b_{12} increases with ε_1. In this case, the second buyer has to increases b_{22}, or the traffic it can get from 1-th seller will decreases. However, due to the increasing of b_{22}, the total bids of the second buyer (b_2) may reach to the upper bound (b_{max}). Moreover, due to the increasing of b_{11}, b_{21} has to be increased, too. For these reasons, b_{22} decreases when ε_1 is big enough. In such case, the buyer may increase the upper bound and use more money to get the more traffic.

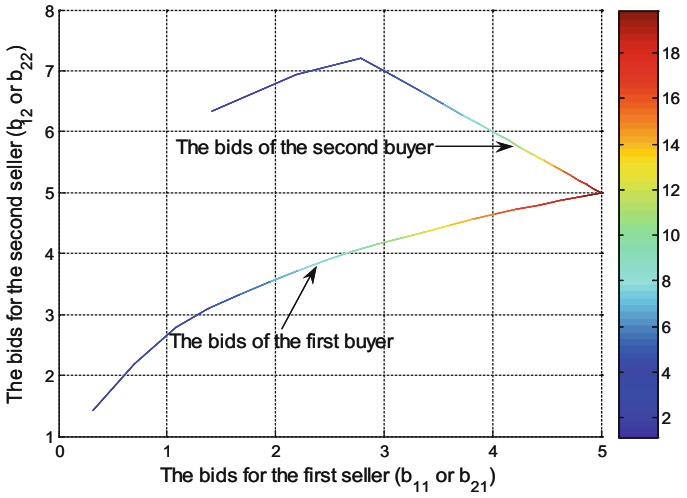


Fig. 4. The bids in Nash Equilibrium when ε_1 increases from 1 to 20 ($M = 2$)

6 Conclusions

This paper first proposes a framework to share the traffic without the help of ISPs. Then, it considers the competition between the buyers, and mathematically proves that the Nash Equilibrium exists. The result gives guidance for the users to buy the traffic efficiently.

Acknowledgments. The work reported in this paper is supported by NSFC No. 61401482.

References

1. AT&T.: AT&T sponsored data (2014). http://www.att.com/att/sponsoreddata/en/index.html#fbid=fNjHshoHkg
2. https://www.hk.chinamobile.com/tc/2cm_intro
3. Zheng, L., Joe-Wong, C., Tan, C.W., Ha, S., Chiang, M.: Secondary markets for mobile data: feasibility and benefits of traded data plans. In: Proceedings of IEEE INFOCOM (2015)
4. Sen, S., Joe-Wong, C., Ha, S.: The economics of shared data plans. In: Proceedings of WITS (2016)
5. Jin, Y., Pang, Z.: Smart data pricing: to share or not to share? In: Proceedings of IEEE INFOCOM Workshops on Smart Data Pricing (2014)
6. Zeng, L., Joe-Wong, C., Tan, C., Chiang, M., Wang, X.: How to bid the cloud. In: Proceedings of ACM SIGCOMM, pp. 71–84 (2015)
7. Yu, T., Zhou, Z., Zhang, D., Wang, X., Liu, Y., Lu, S.: Indapson: an incentive data plan sharing system based on self-organizing network. In: Proceedings of IEEE INFOCOM, pp. 1545–1553 (2014)

8. Feng, G., Garg, S., Buyya, R., et al.: Revenue maximization using adaptive resource provisioning in cloud computing environments. In: Proceedings of IEEE Grid, pp. 192–200 (2012)
9. Wang, X., Chen, M., Kwon, T., Jin, L., Leung, V.: Mobile traffic offloading by exploiting social network services and leveraging opportunistic device-to-device sharing. IEEE Wirel. Commun. **21**(3), 28–36 (2014)
10. Lee, K., Lee, J., Yi, Y., Rhee, I., Chong, S.: Mobile data offloading: how much can WiFi deliver? IEEE/ACM Trans. Netw. **21**(2), 536–550 (2013)
11. Reny, P.J.: Non cooperative games: equilibrium existence. In: The New Palgrave Dictionary of Economics, 2nd edn. (2005)



Continuous Cell Zooming Algorithm Towards Energy Efficient in Random Heterogeneous Cellular Networks

Jie Sheng, Yang You(✉), Dong Ma, Canyan Zhu, and Fan Xu

School of Rail Transportation, Soochow University,
Suzhou 215021, Jiangsu, China
yyou@stu.suda.edu.cn

Abstract. This paper presents an energy efficient aware continuous cell zooming algorithm based on green random cellular heterogeneous networks (HetNet). Firstly, we derive a two-tier Poisson Voronoi Tessellation (PVT) random heterogeneous network model in which the macro base stations (MBS), small-cells base stations (SBS), macro users (MU) and small-cell users (SU) locations are drawn randomly from independent homogeneous Poisson point processes (HPPPs). Secondly, by updating the SUs' locations in SBSs periodically, the radius of the cells can be adjusted adaptively to achieve cell zooming. The performance of this work is examined under different user densities and ratios compared with reference algorithm. Numerical evaluations show that our proposed algorithm can significantly decrease system energy consumption and improve energy efficient.

Keywords: Cell zooming · Energy efficiency · PVT · HetNets

1 Introduction

During recent years, the strong construction of 4G networks and the popularity of intelligent terminal equipment brought about a huge upgrade in bandwidth and the rate of communication networks for user equipment (UE). However, the contradiction between the energy consumption and the operating cost of the network is more and more prominent. According to research [1, 2], information and communication technology industry (ICT) has become the world's fifth largest energy-consuming industry with 5% of the world's carbon emissions, in which over 60% is base station energy consumption. The large amount of energy consumption has stimulated the 5G standardization agencies and researchers to build an environmental protection 5G cellular network [3].

Cell zooming (CZ) is a technique [4] that ultimately reduces the energy consumption in a cellular network by adjusting transmit power of base station and changing the size of the cell, including both macro base stations (MBS) and small-cell base stations (SBS). The term "small-cell" contains femtocell, microcell, picocell and metrocell. A study [5] has confirmed that cell zooming can roughly reduce the total power consumption by about 20%. It is considered to be one of the potential strategies

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 502–511, 2018.
https://doi.org/10.1007/978-3-319-78130-3_52

for developing green communication. Meanwhile, another research [6] proposed that it is a change to realize how dynamic cell zooming optimize and effect the cell zooming on the quality of service (QoS) and how to further enhance the energy efficiency (EE) when the QoS is constrained.

In order to complete the challenge, Tun [7] and Balasubramaniam [8] proposed and evaluated the currently more advanced cell zooming algorithms: the continuous zooming algorithm, the fuzzy zooming algorithm, and the discrete zooming algorithm. According to the results of these two studies, we can find that the continuous zooming algorithm is more efficient than the other two algorithms. After that, Tun and Kunavut [9] proposed two kinds of cell zooming algorithms based on the adaptive address update scheme: TAPU (Time-Adaptive Periodic Update) and LAPU (Location-Adaptive Periodic Update). Compared with TAPU, LAPU can not only reduce the number of updates, but also maintain high level of service quality.

In addition, the actual base stations are not fully deployed as a regular grid model, but rather exhibit random characteristics due to differences in business areas (such as downtown, neighborhoods, parks, outskirts, suburbs) and capacity (networks and links) [10]. As a result, the traditional hexagonal grid model is too ideal to be used in modeling single-tier and random diversified multi-tier heterogeneous cellular networks [11]. Recently, a new modeling method based on stochastic geometric theory is widely used in modeling cellular networks [12]. It can not only maintain a high level of QoS, but also improve EE.

Based on the above analysis, we construct a two-tier heterogeneous network model in the Poisson Voronoi tessellations (PVT) random cellular networks in this paper, and research a dynamic continuous cell zooming algorithm according to cell user location and QoS requirements. By adaptively updating the cell coverage radius to achieve energy savings, ensuring user throughput requirements, while improving network spectrum utilization and energy efficiency.

The rest of the paper is organized as follows. The next section introduces a two-tier heterogeneous random cellular network system model. Section 3 describes throughput, EE and ASE. Meanwhile, a new proposed continuous cell zooming algorithm is provided in. The simulation parameters and results are mentioned in Sect. 4. Finally, conclusion and future work are pointed out in Sect. 5.

2 System Model

We study the use of orthogonal frequency division multiple access (OFDMA) two-tier heterogeneous cellular network, as shown in Fig. 1, that includes two types of base stations (BS). Assuming that each base station's distribution is based on a separate two-dimensional homogeneous Poisson point process (HPPP), the distribution of MBS is represented by Φ_m, and λ_m is the density of Φ_m. Also, the distribution of SBS is represented by Φ_s, and λ_s is the density of Φ_s.

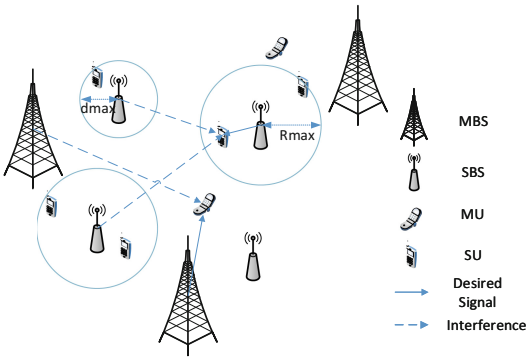


Fig. 1. An illustration of two-tier random heterogeneous cellular network model

At the same time, we assume that the access criterion of the UE is to access the nearest base station. Because the access distance is used as the standard, this will form the Poisson Voronoi tessellations (PVT) as shown in Fig. 2.

As can be seen from Fig. 2, the deployment of macro base station divides entire flat into multiple Voronoi graphics and each Voronoi diagram represents the service range of corresponding macro base station. Each small-cell base station covers a circle as a hot spot area, among which the maximum coverage radius is R_{max}. It can be seen that this Voronoi topology is clearer than the classical rule lattice structure to reflect the highly irregularity of the actual network node deployment.

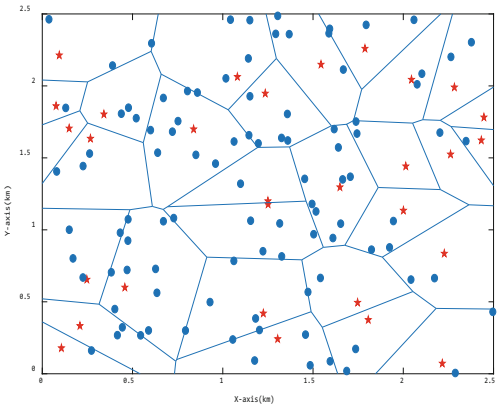


Fig. 2. An example of the PVT random cellular network: red five-pointed stars are MBSs and blue points are SBSs (Color figure online)

We divide the UE into two categories: macro users (MU) and small-cell users (SU) according to the different access points. The small-cell users are distributed in each circular hot spot area as independent HPPP Φ_{su} whose density is λ_{su} and are served by SBS. Meanwhile the macro users are distributed over the entire plane as independent HPPP Φ_{mu} at the density λ_{mu} and are served by the MBS. The average

density of all UE is λ_{UE}. The resources available on the network are divided into N subchannels averagely. Assuming that the base station is assigned a spectrum of resources to communicate with UE, cross-layer interference is generated, and each UE occupies one subchannel to transmit data. Also, we assume that MU can only access the macro cells, while SU will access the small-cells at first. If there are no suitable small-cells to access, the SU will access the nearest macro cell instead.

Define a mode matrix A, when it comes to MBS, we have $A = [1 \quad 0]$, and for SBS the matrix is written by $A = [1 \quad 0]$.

The maximum transmitted power of the base station is given by $P_{max} = \begin{bmatrix} P_{Mmax} \\ P_{Smax} \end{bmatrix}$. where, P_{Mmax} is the maximum transmitted power of MBS and P_{Smax} is that of SBS. Similarly, maximum covering radius of the base station is given by $R_{max} = \begin{bmatrix} R_{Mmax} \\ R_{Smax} \end{bmatrix}$.

Thus, the transmitting power of BS_j for the UE_i can be written as shown in (1):

$$P_{ij} = A * P_{max} * \left[d_{ij}/(A * R_{max})\right]^{\alpha} \tag{1}$$

where, d_{ij} is the distance between UE_i and BS_j, α is the path loss exponent.

The linear power model [13] is referenced in (2) to present the linear relationship between energy consumption of BS and the transmitted power.

$$P_c = A * a * P_{ij} + A * b \tag{2}$$

where, P_c is the average consumed power, coefficient a is amplification factor for compensating the difference between the average transmission power and power consumption caused by the factors of power amplifier, loss of feeder and air-conditioning. Variable b represents power consumption bias that has nothing to do with the transmission power, including the consumed power of signal processing, batteries and cold air conditioning. Without doubt, corresponding to different types of base station, a and b have different values, we have $a = \begin{bmatrix} a_M \\ a_S \end{bmatrix}$ and $b = \begin{bmatrix} b_M \\ b_S \end{bmatrix}$, where M represents MBS and S represents SBS.

3 Proposed Cell Zooming Algorithm

3.1 Methodology

Without loss of generality, by exploiting the HPPP characteristics of the BS and UE locations, we can just focus and analyze the behavior of a typical UE located at the origin and its typical BS. Then the obtained results will be generalized to the whole network.

The signal to interference and noise ratio (SINR) at UE_i in the coverage of BS_j can be written as:

$$SINR = A * P_{ij} * a/(I_{ij} + \sigma^2) \tag{3}$$

where, σ^2 is an additive white Gaussian noise (AWGN) power, and I_{ij} denotes interference from other cells. Formally, I_{ij} is defined below:

$$I_{ij} = \sum\nolimits_{h\in\{\Phi_M,\Phi_S\},h\neq j} A * P_{ih} * a \tag{4}$$

where, h is the other base station of the same type, BS_h generate the same layer and cross-layer interference to UE_i.

According to the Shannon formula, the throughput of UE_i in BS_j is:

$$C_{ij} = W * log_2(1 + SINR) \tag{5}$$

where, W is the spectral bandwidth of active UE in a cell.

The energy efficiency (EE) of an active base station can be expressed as:

$$EE = A * W * C_{ij} * (\lambda_{MU} + \lambda_{SU}) / \sum P_c \tag{6}$$

The area spectral efficiency (ASE) can be expressed as:

$$ASE = C_{ij} * (\lambda_M + \lambda_S) \tag{7}$$

3.2 Continuous Cell Zooming Algorithm

On the basis of the above-mentioned network model and performance parameters, we will propose a new continuous cell zooming algorithm (CCZA), by updating the SUs' locations in SBSs periodically, the radius of the cells can be adjusted adaptively to achieve cell zooming. The specific algorithm implementation steps are as follows.

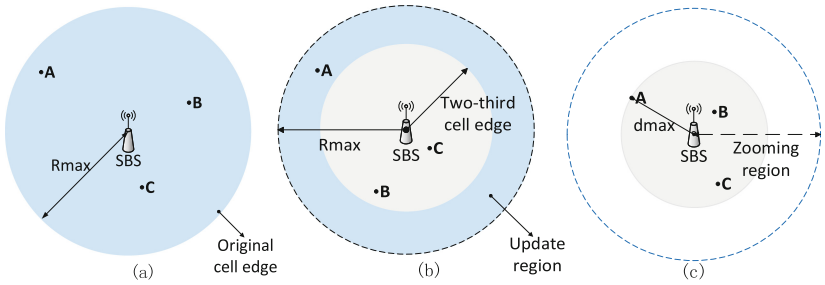


Fig. 3. Conceptual demonstration of cell zooming algorithm (a) the initial state (b) location-adaptive periodic update (LAPU) (c) continuous cell zooming algorithm (CCZA)

Step 1:

Initialize the entire cellular network. As shown in Fig. 3(a), the coverage radius of the base station is R_{max}, also, the traffic load of each MBS and SBS is zero. An update interval T is set for all UEs in the cell and we assumed that the UE model is simplified in T.

Step 2:
2-1: In each period, according to the pilot signal strength, UE measures the distance d_{ij} between the connected base station. When the update time arrives, all users which in the cell perform updates and report its location to the base station. MBSs and SBSs receive and update the message, after that, they find the distance d_{max} from its most distant user in its served UEs.
2-2: SBS coverage radius zoom to the farthest user's position d_{max}, which is shown in Fig. 3(c).
2-3: Calculate P_{ij}, C_{ij}, EE, ASE at this moment.
Step 3:
3-1: If SBS receive 0 message, which indicates that there is no user in the cell, this SBS goes into sleep mode.
3-2: If MBS receive 0 message, this MBS goes into sleep mode. And the PVT model will have changed, so return to step 2.
3-3: Adjust λ_m, λ_s, λ_{su}, λ_{mu} separately, finally, return to step 2.

4 Performance Evaluation and Numerical Results

In this section, we evaluate the performance of our proposed continuous cell zooming algorithm by using MATLAB simulation tool. The parameter values we use in our analysis are reported in Table 1. The simulation environment is the square area with the length of 2.5 km, and there are 30 MBSs and 100 SBSs distributed in the area. The coverage area of a macro cell is generated by PVT algorithm, and the coverage area of a small-cell is the circular determined by cell radius.

Table 1. System parameters.

Parameter	Value
P_{Mmax}	43 dBm
P_{Smax}	33 dBm
R_{Mmax}	500 m
R_{Smax}	100 m
α	3
a_M	22.6 W
a_S	5.5 W
b_M	412.4 W
b_S	32.0 W
σ^2	−80 dBm
W	0.1 MHz

4.1 Performance Analysis of CCZA

First, the energy consumption (EC) of the network for different user ration is depicted in Fig. 4. As shown, with the increase of λ_{UE} in the network coverage area, the total EC

also increases. Meanwhile the network EC increases along with the growth of the proportion of MU. The reason is that, MU can only access the macro cells, while SU will access the small-cells at first. If there are no suitable small-cells to access, the SU will access the nearest macro cell instead. According to the transmitted power parameters of two cells, when the distance is fixed, the required transmission power from the macro cell is higher to meet the user's demand. Therefore, with the increasing number of SU, the algorithm proposed in this paper will effectively reduce the network energy consumption.

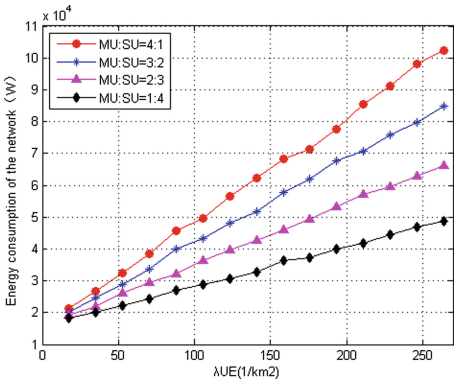


Fig. 4. Energy consumption of the network for different user ratio

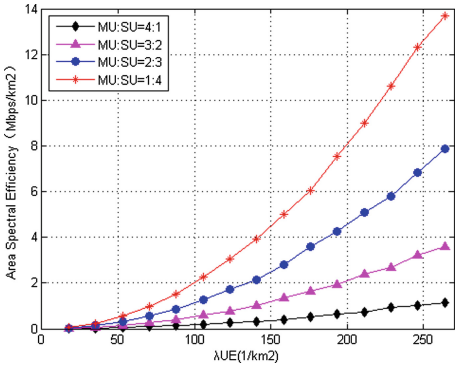


Fig. 5. Area spectral efficiency of the network for different user ratio

Furthermore, Fig. 5 shows the ASE grows monotonically with λ_{UE} and the ratio of SU. With the growth of SU proportion in the network, more and more users connect to small-cells for service. By dynamically scaling the radius of small-cell, the CCZA algorithm can achieve the highest throughput per unit area, so as to make an impressive promotion of ASE of the system.

Figure 6 illustrates the energy efficiency for the PVT random heterogeneous cellular network. It is obvious that, with the increase of λ_{UE}, the network energy efficiency (EE) increases at first and leveled off afterwards, and gradually be attracted to the maximum efficiency of the network. It is noteworthy that the EE of the network grows monotonically with SU proportion in the network, which shows the advantage in improving energy efficiency of CCZA algorithm.

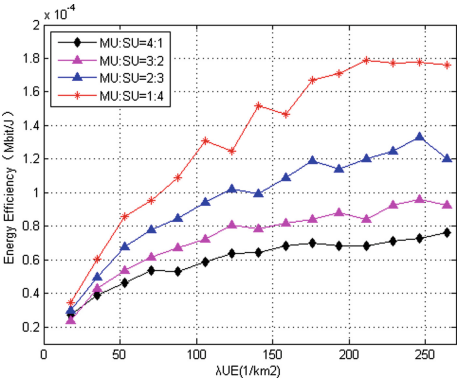


Fig. 6. Energy efficiency of the network for different user ratio

4.2 Performance Comparison Between CCZA and LAPU

In order to further evaluate the effectiveness of the cell zooming algorithm in this paper, we compare the performance of proposed CCZA with LAPU algorithm [7]. Seeing the $2/3R_{Smax}$ as the boundary, the coverage area of small-cell will be classified into two parts by using the LAPU algorithm. As is shown in Fig. 3(b), the external area is defined as an update region, and the internal area is an non-update region. The UEs in the update region will renew its location in each replacement cycle, otherwise the UEs will skip the update. In the comparison of the two algorithms, we considered two scenarios with different user ratios.

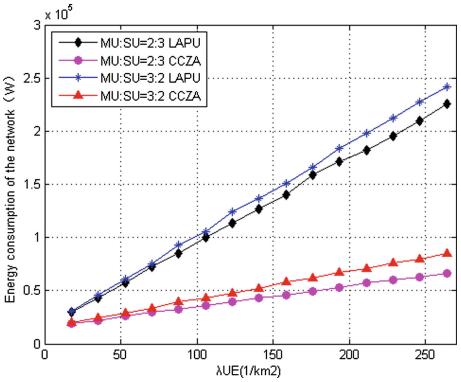


Fig. 7. A comparison of energy consumption for the network

Figure 7 compares the variation of the system total energy consumption by using two algorithms in ratios of the number of MU and SU at 2:3 and 3:2 respectively. As can be seen in the figure, in general the energy consumptions of two algorithms both improve by increasing λ_{UE}. The EC of CCZA is always lower than that of LAPU, and the advantage is more and more significant with the increasing of λ_{UE}. When the user density reached 265/km^2, the EC of LAPU were 225.56 kW and 241.54 kW, and the EC of CCZA were just 66.26 kW and 84.72 kW which were reduced by 70.6% and 64.9% respectively compared with LAPU.

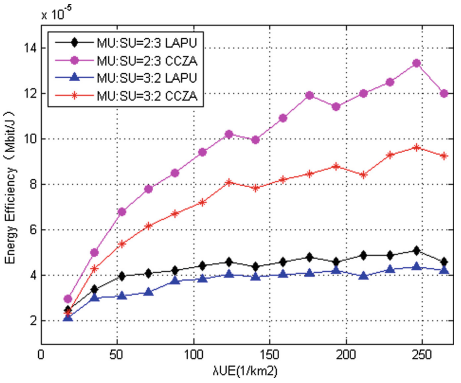


Fig. 8. A comparison of energy efficiency for the network

Figure 8 compares the system EEs with two algorithms in ratios of the number of MU and SU at 2:3 and 3:2 respectively. Obviously, the EEs of two algorithms were all increase at first and leveled off afterwards, which gradually were attracted to the maximum values by increasing of λ_{UE}. The EE of CCZA is always higher than that of LAPU. When λ_{UE} equals to 265/km^2, the EE of LAPU were $4.61 * 10^{-5}$ and $4.22 * 10^{-5}$, and the EE of CCZA reached up to $1.2 * 10^{-4}$ and $9.62 * 10^{-5}$ which were 2.6 times and 2.27 times respectively of the results of LAPU. The reason is that, only the SUs in the external area will update the location and the radius of small-cells can only zooming to $2/3R_{Smax}$ by LAPU, while the CCZA expanded the scope of dynamic area by making further efforts in shrinking the cell radius and comes result of saving a lot of unnecessary energy consumption as well as improving the EE of the system.

5 Conclusion

In this work, we investigated a continuous cell zooming algorithm towards energy efficient in random HetNets. The algorithm provides different services for different kinds of users, and by updating the SUs' locations in small-cells periodically, the radius of small-cells can be adjusted adaptively to achieve cell zooming. The performance evaluation of the algorithm was also discussed by means of modeling and simulation results analyzing.

Numerical evaluations showed that, with the increase of SU proportion in HetNets, the total energy consumption of the system was effectively reduced, and the ASE as well as the EE were improved substantially while meeting the requirement of QoS by adopting the algorithm of this paper. Compared with the performance of LAPU algorithm, the user density growth will lead to a gradually convergence of the EE, and CCZA has a significant improvement in saving energy and improving EE.

Acknowledgments. The work was supported by grants from the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (15KJB510027), the 19th Batch of College Students Extracurricular Academic Research Fund Project of Soochow University (KY2017734B, KY2017727B), and the Foundation Research Project of Jiangsu Province (BK20140359, BK20170346).

References

1. Chin, W.H., Fan, Z., Haines, R.: Emerging technologies and research challenges for 5G wireless networks. IEEE Wirel. Commun. **21**(2), 106–112 (2014)
2. Auer, G., et al.: How much energy is needed to run a wireless network? IEEE Wirel. Commun. **18**(5), 40–49 (2011)
3. Andrews, J.G., et al.: What will 5G be? IEEE J. Sel. Areas Commun. **32**(6), 1065–1082 (2014)
4. Niu, Z., Wu, Y., Gong, J., Yang, Z.: Cell zooming for cost-efficient green cellular networks. IEEE Commun. Mag. **48**(11), 74–79 (2010)
5. Prithiviraj, V., Venkatraman, S.B., Vijayasarathi, R.: Cell zooming for energy efficient wireless cellular network. J. Green Eng. **3**, 421–434 (2013)
6. Tun, K.C., Kunavut, K.: An overview of cell zooming algorithms and power saving capabilities in wireless networks. KMUTNB Int. J. Appl. Sci. Technol. **7**(3), 1–13 (2014)
7. Tun, K.C., Kunavut, K.: Performance evaluation of dynamic cell zooming algorithms in omni-directional and sector-based cells. In: 2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE), Chon Buri, pp. 158–163 (2014)
8. Balasubramaniam, R., Nagaraj, S., Sarkar, M., Paolini, C., Khaitan, P.: Cell zooming for power efficient base station operation. In: 2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC), Sardinia, pp. 556–560 (2013)
9. Tun, K.C., Kunavut, K.: Adaptive location update schemes for continuous cell zooming algorithm in wireless networks. In: 2015 7th International Conference on Knowledge and Smart Technology (KST), Chonburi, pp. 155–160 (2015)
10. ElSawy, H., Hossain, E., Haenggi, M.: Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: a survey. IEEE Commun. Surv. Tutor. **15**(3), 996–1019 (2013)
11. Andrews, J.G., Baccelli, F., Ganti, R.K.: A tractable approach to coverage and rate in cellular networks. IEEE Trans. Commun. **59**(11), 3122–3134 (2011)
12. Ge, X., Yang, B., Ye, J., Mao, G., Wang, C.X., Han, T.: Spatial spectrum and energy efficiency of random cellular networks. IEEE Trans. Commun. **63**(3), 1019–1030 (2015)
13. Fehske, A.J., et al.: Energy efficiency improvements through micro sites in cellular mobile radio networks. In: 2009 IEEE Globecom Workshops, Honolulu, HI, pp. 1–5 (2009)



D2D Assisted MU-MIMO Precoding to Enhance LTE Network

Yanqiu Zhang[1], Shu Fang[2(✉)], Yuanchao Han[1,2], and Yu Zeng[1,2]

[1] Department of National key Laboratory on Communications, University of Electronic Science and Technology of China, Chengdu, China
[2] Qingshuihe Campus, No. 2006, Xiyuan Avenue, West Hi-Tech Zone, Chengdu 611731, China
fangshu@uestc.edu.cn

Abstract. Multi-user multiple-input multiple-output (MU-MIMO) has been adopted as a key enabling technique to achieve the capacity demand in 4G wireless networks. However, there are still some problems existing in the frequency division duplex (FDD) MU-MIMO system, such as the channel quality indicator (CQI) mismatch problem, large uplink feedback, low pairing probability of MU communication. In this paper, we exploit a MU-MIMO scheme in FDD assisted by device-to-device (D2D) technique to enhance the conventional LTE network. The proposed scheme could not only solve the above problems in conventional MU-MIMO systems, but also achieve significant performance improvement than conventional MU-MIMO schemes. Since the user pairing process is done at the user side, the feedback of the proposed scheme is reduced dramatically compared with the conventional scheme. Simulation results show that the proposed D2D assisted MU-MIMO scheme has clear superiority in terms of throughput over conventional MU-MIMO scheme in LTE networks. Because with D2D assisted MU-MIMO, the proposed scheme tackles the CQI mismatch problem and increases the multi-user pairing probability to enhance the overall throughput.

Keywords: Device-to-device · Multiple-input multiple-output
Feedback · Channel quality indicator · Precoding

1 Introduction

Multi-user multiple-input multiple-output (MU-MIMO) technique has been adopted as one of the core technologies in LTE-A system. The basic feature of MU-MIMO is to empoly the space division muliple access (SDMA) technology to transmit multiple data streams from multiple users simultaneously [1]. Therefore, the inter-channel interference is existed in MU-MIMO system and precoding technique is adopted for MU-MIMO system so as to combat the interference.

In the time division duplex (TDD) mode, the evolved Node B (eNB) can easily obtain the channel state information (CSI) via channel reciprocity of uplink

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 512–523, 2018.
https://doi.org/10.1007/978-3-319-78130-3_53

channel. However, in frequency division duplex (FDD) mode, where channel reciprocity is not existed, it is necessary to feed CSI back to the eNB for each user in MU-MIMO system [2]. The conventional MU-MIMO scheme in FDD mode will feed back the Rank-1 channel quality indicator (CQI) and precoding matrix information (PMI) without multi-user interference to eNB. After eNB collects all of the feedback information, it will do multi-user scheduling to pair some of the users at the same time-frequency resource block to estimate the CQI with multi-user interference. Therefore, the modulation and coding scheme (MCS) is finally confirmed at eNB according to the estimated CQI at eNB. Therefore, the CQI mismatch problem is always existed in FDD MU-MIMO system. At the same time, the pairing probability of FDD MU-MIMO is low.

Recently, device-to-device (D2D) communication in cellular networks has received much attention thanks to its potential to provide higher date rate [3]. D2D communication can be defined as a direct communication between two UEs. Specifically, unlike the conventional cellular communication where all traffic is routed via eNB, the adjacent UEs can establish direct links in the control of eNB to transmit data to each other under cellular network resources. It has the advantages of improving spectrum utilization, shorting packet delay, and reducing energy consumption, while enabling new peer-to-peer and location-based applications and services [4].

There are some recent research work and literature concerned with tackling the above issues existing in MU-MIMO system. In [5,6], the authors refers to the idea of joint spatial division and multiplexing (JSDM) to realize savings in CSI for FDD mode. It lies in clustering the UEs into groups with approximately similar channel covariance eigenvectors, while users across groups have near orthogonal covariances. To mitigate or avoid mutual interference between cellular and D2D transmissions, a new strategy of setting a restricted area was proposed in [7], in which the cellular users can not coexist with D2D users, thus the interference caused by the uplink cellular users can be controlled within a certain range. In [8], the authors found that interference can be completely overcome if the number of canceled D2D interfering signals is scaled with the number of base station (BS) antennas at an arbitrarily slow rate. In [9], the authors proposed to suppress co-channel interference by maximizing the signal-to-leakage-noise ratio (SLNR) for all users simultaneously.

For the resource allocation of D2D communication in the MU-MIMO system, an enhanced D2D resource allocation scheme under the TDD mode is proposed [10], where eNB acts as a relay. It receives data from the D2D users and forward the data stream when the D2D link is interrupted, thus guaranteeing the quality of D2D communication. Nash bargaining solution and nash competitive games frameworks are proposed to address the power allocation problem for D2D underlay cellular network in [11]. In [12], the author used game theory to study the resource allocation in the networks, and an efficient resource allocation algorithm based on better response dynamic is developed for a feasible suboptimal solution. In [13], the author adopts the subset and iterative concept to complete the precoding selection, which can mitigate the co-channel interference between

D2D and cellular transmissions in the meantime. Several practical methods are presented in [14] for D2D and cellular nodes to cooperate effectively and efficiently in determining their precoder selections from LTE compatible MIMO codebooks. In order to increase the overall sum rate of the system, an algorithm for determining the connection state of all D2D pairs as 0 or 1 is proposed in [15], as for the designing of precoding vectors, the maximum sum rate is regarded as the optimal condition.

To cope with the CQI mismatch problem, reduce the uplink feedback, and increase the pairing probability of conventional FDD MU-MIMO communication, we proposed a D2D assisted MU-MIMO precoding scheme as an alternative to the conventional MU-MIMO system. The proposed new scheme exploits a novel strategy to bring D2D technology into multi-user MIMO network, where the user pairing and the CQI selection process is done at the user side and only the paired D2D users feeds back to the eNB. Therefore, the uplink feedback and the computational complexity at eNB are significantly reduced. Furthermore, the information interaction between UEs is relatively small since only the channel capacity and PMI are needed.

To evaluate the system performance, we compare the D2D asisted MU-MIMO precoding scheme with rank-1 FDD MU-MIMO based on zero forcing (ZF) precoding [16]. Numerical results demonstrate that with the help of D2D technology, conventional FDD MU-MIMO system can achieve higher throughput, smaller uplink feedback, and solve the CQI mismatch problem.

This paper is organized as follows. Section 2 presents the conventional FDD rank-1 MU-MIMO system. In Sect. 3, the detailed process of the D2D assisted FDD rank-1 MU-MIMO scheme is presented. Numerical results and performance evaluations are presented in Sect. 4. Conclusions are drawn in Sect. 5.

Notation: We use $(\cdot)^H$ and $\|\cdot\|_F$ to denote the transpose and Frobenius norm of a vector/matrix respectively. We use $(\cdot)^{-1}$ indicates the inverse of the matrix. We use $(\cdot)^2$ to denote the square of a variable. $(\cdot)_{i,j}$ denotes the i-th row and j-th column of the matrix. $[\cdot]_k$ represents the k-th column of the matrix. We denote by $\mathcal{CN} \sim (m, \sigma^2)$ the complex Gaussian distribution of a random variable having independent Gaussian distributed real and imaginary part, denoted by $\mathcal{CN} \sim (m, \frac{\sigma^2}{2})$, with mean m and variance $\frac{\sigma^2}{2}$.

2 The System Model

The conventional MU-MIMO system in FDD mode is shown in Fig. 1. Consider a MU-MIMO system where N_T antennas are equiped at eNB and N_R antennas are equiped at each UE. Householder codebooks are generated off line both at the eNB and each UE. Each UE feeds back the vector index PMI from the codebook and its corresponding CQI to the eNB. The eNB implements user scheduling and select precoding matrix from the codebook. There are totally X users to be scheduled and $K(K \leq N_T)$ users to be transmitted according to the feedback channel information and the scheduling criterion, then the scheduled UEs will

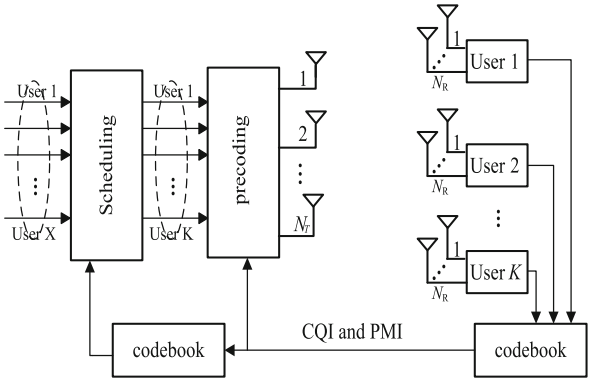


Fig. 1. The system model of conventional MU-MIMO

be precoded by the precoding matrix before transmission. The channel matrix $\mathbf{H_k}$ for the k-th UE is presented as

$$\mathbf{H}_k = \begin{bmatrix} h_{1,1} & \cdots & h_{1,M} \\ \vdots & \ddots & \vdots \\ h_{N,1} & \cdots & h_{N,M} \end{bmatrix}, \tag{1}$$

Where $h_{i,j}$ denotes the channel impulse response coupling the j-th transmitter to the i-th receiver element. The amplitude of $h_{i,j}$ obeys independent and identically Rayleigh-distribution.

For rank-1 FDD MU-MIMO scheme, the precoding algorithm is represented as below

Step 1 Each UE feeds back a maximum rank-1 CQI to eNB through uplink, which can be defined as

$$\mathrm{CQI}_k = \max_{\mathbf{e}_m \in \mathcal{C}} \frac{\|\mathbf{H}_k \mathbf{e}_m\|}{\sigma^2}, \tag{2}$$

Where $\mathbf{e}_m$ denotes the m-th codebook vector of Householder rank-1 codebook, and $\mathcal{C}$ is the rank-1 Householder codebook.

Step 2 eNB predicts the CQI of paired UE on the basis of rank-1 CQI fed back by UEs. Consider user l and user k are paired successfully, then the predicted $\mathrm{CQI}_{k,predict}$ for the k-th user under MU communication is shown as

$$\mathrm{CQI}_{k,predict} = \frac{\mathrm{CQI}_k / (2n) \times \beta_{11}^2 / \|\mathbf{F}_k\|_F^2}{1 + \mathrm{CQI}_k / (2n) \times \beta_{12}^2 / \|\mathbf{F}_l\|_F^2}, \tag{3}$$

$\mathrm{CQI}_{l,predict}$ has similar formulation as $\mathrm{CQI}_{k,predict}$, where user k is regarded as an interference. $\beta_{i,j} = \left(\left(\mathbf{V}' \mathbf{V} \right) \left(\mathbf{V}' \mathbf{V} + \rho \mathbf{I} \right)^{-1} \right)_{i,j}$, and

$\mathbf{V^H} = [\mathbf{e}_k, \mathbf{e}_l]$, $\mathbf{e}_k$ and $\mathbf{e}_l$ are the precoding vector fed back by k-th and l-th user, ρ is taken as the value of SNR. The calculation of $\mathrm{CQI}_{k,predict}$ and $\mathrm{CQI}_{l,predict}$ are implemented when $\mathbf{e}_k \neq \mathbf{e}_l$. If $\mathbf{e}_k = \mathbf{e}_l$, the pairing process fails, user l and user k can not perform the MU communication. Where $\mathbf{F} = [\mathbf{F}_k, \mathbf{F}_l] = \mathbf{V}^H\left(\mathbf{V}\mathbf{V}^H + \rho\mathbf{I}\right)^{-1}$, $\mathbf{F}_k$ and $\mathbf{F}_l$ are the precoding vector for user k and user l respectively when they are paired. The parameter n is an empirical value.

Step 3 eNB operates the user scheduling and communication mode selection based on the rank-1 CQI and $\mathrm{CQI}_{k,predict}$, $\mathrm{CQI}_{l,predict}$. The eNB will select the user pair with the maximum $\mathrm{CQI}_{MU,predict}$, where

$$\begin{aligned} \mathrm{CQI}_{MU,predict} = &\log_2\left(1 + \mathrm{CQI}_{k,predict}\right) \\ &+ \log_2\left(1 + \mathrm{CQI}_{l,predict}\right), \\ k, l = 1, 2, \cdots, K \quad &k \neq l \end{aligned} \tag{4}$$

If the $\mathrm{CQI}_{MU,predict} > \log_2\left(1 + \mathrm{CQI}_{k,predict}\right)$, $k = 1, 2, \cdots, K$, the eNB will launch MU-MIMO precoding. Otherwise, the SU-MIMO precoding begins.

If user k and user l is the optimal user pair, the received signal at the k-th UE can be represented as

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{W}\mathbf{s} + \mathbf{n}_k, k = 1, \ldots, K, \tag{5}$$

$$\mathbf{s} = [p_k s_k \quad p_l s_l], \tag{6}$$

$$\mathbf{W} = [\mathbf{F}_k \quad \mathbf{F}_l]. \tag{7}$$

Where $\mathbf{y}_k$ denotes the received signal of user k, p_k is the transmitted signal power of the k-th data stream, and the addictive noise $\mathbf{n}_k$ obeys distribution $CN(0, N_0)$. The channel matrix at UE side for the k-th user after precoding is

$$\widetilde{\mathbf{H}}_k = \mathbf{H}_k\mathbf{W}. \tag{8}$$

The linear MMSE decoding vector for the k-th user $\widetilde{\mathbf{G}}_k$ correspondingly is shown as

$$\widetilde{\mathbf{G}}_k = \widetilde{\mathbf{h}}_k^H(\widetilde{\mathbf{H}}_k\widetilde{\mathbf{H}}_k^H + \frac{2N_0}{P_0}\mathbf{I}_M)^{-1}, \tag{9}$$

Where $\mathbf{I}_M$ denotes the $M \times M$ identity matrix and P_0 is the transmitted power of all data streams. In order to analyze the system simply and efficiently, we consider equal power allocation in this paper. The equivalent channel matrix of user k after precoding is

$$\widetilde{\mathbf{h}}_k = \mathbf{H}_k\mathbf{F}_k. \tag{10}$$

The SINR for the k-th user with the linear detection is

$$\mathrm{SINR}_k = \frac{p_k\left|\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{F}_k\right|^2}{p_l\left|\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{F}_l\right|^2 + \|\mathbf{G}_k\|_2^2 N_0}. \tag{11}$$

The capacity of the MU-MIMO system can be represented as

$$\mathrm{C}_{MU} = \log_2\left(1 + \mathrm{SINR}_k\right) + \log_2\left(1 + \mathrm{SINR}_l\right). \tag{12}$$

3 MU-MIMO Precoding Scheme with D2D

In the conventional rank-1 FDD MU-MIMO system, the user is usually unware of the multi-user interference. Thus, the UEs feeds back an ideal non-interfering rank-1 CQI/PMI through the uplink. Then the eNB predicts the rank-2 CQI for MU communication based on some precoding algorithm such as zero forcing. Although we can reduce the gap between an practical value and the predicted value of CQI by some advanced algorithms, the predicted value is still inaccurate, that would cause CQI mismatch problem. To cope with the above problems, the D2D technique is introduced in FDD MU-MIMO scenario.

The details of the D2D assisted rank-1 FDD MU-MIMO system is as follows

Step 1 Each user estimates the channel matrix $\mathbf{H}_k$ according to the downlink CSI-RS respectively.

Step 2 Each UE calculates the channel capacity value C_k and the optimal precoding vector and its corresponding index PMI based on $\mathbf{H}_k$ as

$$\mathrm{C}_k = \log_2\left|\mathbf{I} + \frac{1}{N_T N_0}\mathbf{H}_k^H\mathbf{H}_k\right|, \tag{13}$$

Where $\mathbf{I}$ denotes the identity matrix. N_T is the number of transmit antennas, and N_0 is the noise power that obeys $\mathcal{CN} \sim (0, 1)$.

Step 3 Each UE calculates the optimal precoding vector according to rank-1 codebook. The k-th user's CQI is calculated as

$$\mathrm{CQI}_k = \max_{\mathbf{e}_m \in \mathcal{C}} \frac{\|\mathbf{H}_k\mathbf{e}_m\|}{N_0}, \tag{14}$$

PMI_k is the corresponding index of the best precoding vector for user k. $\mathcal{C}$ has been defined in Eq. (2).

Step 4 Each UE broadcasts the channel capacity C_k, $k \in \{1, 2, \cdots, K\}$ and the optimal codebook index PMI_k, $k \in \{1, 2, \cdots, K\}$ respectively via the D2D link.

Step 5 Based on the information from other UEs and each UE's local channel information, user k will calculate the achievable throughput when combining with user p to form a MU-MIMO pair. The achievable MU-MIMO throughput is expressed as

$$\mathrm{C}_{k,p} = \left(\mathrm{C}_k + \mathrm{C}_p\right) \times \left(1 - \left|\mathbf{e}_k\mathbf{e}_p^H\right|\right), \tag{15}$$

Where $\mathbf{e}_k$ is the precoding vector corresponding to the PMI fed back by the k-th UE, $\mathbf{e}_p$ is the precoding vector corresponding to the PMI fed back by the p-th UE. Each user traverses all possible combinations $\mathrm{C}_{k,p}$

to get the set of the approximate sum capacity value $\{C_G\}$. If the set of the approximate sum capacity value $\{C_G\}$ is empty, which indicates that all the UE's codebook index are conflicting and the user pairing fails. Therefore, only the SU communication is performed with the criterion of maximum capacity. If not, turn to step 6.

Step 6 Take a user pair which has the maximum sum capacity value as a optimal pair, denoted as (i, j). We have

$$(i, j) = \underset{k,p\in\{1,2,\cdots,K\}}{\arg\max}\ C_{k,p}. \tag{16}$$

Thus, the i-th UE and the j-th UE are optimal paired UEs. If the UE k is not in the optimal pair (i, j), this UE enters into a silent state. The UE i in the optimal pair feeds back information SINR_i to eNB as (user j is regarded as an interference)

$$\mathrm{SINR}_i = \frac{p_i|\mathbf{G}_i\mathbf{H}_i\mathbf{F}_i|^2}{p_j|\mathbf{G}_i\mathbf{H}_i\mathbf{F}_j|^2 + \|\mathbf{G}_i\|_2^2 N_0}, \tag{17}$$

Where $\mathbf{G}_i$ denotes the detection matrix, p_i is the transmit power of the i-th user, $\mathbf{F}_i$ is the precoding matrix of user i, which is obtained by zero forcing the codebook vector of user i and user j, it can be presented as

$$\mathbf{W} = [\mathbf{F}_i \ \ \mathbf{F}_j] = \mathbf{V}\left(\mathbf{V}\mathbf{V}^H + \rho\mathbf{I}\right)^{-1}, \tag{18}$$

Where $\mathbf{V}^H = [\mathbf{e}_i \ \ \mathbf{e}_j]$, and ρ is a positive number that is generally perceived as the value of SNR, $\mathbf{e}_i$,$\mathbf{e}_j$ are the vectors reconstructed from rank-1 Householder codebook by PMI_i and PMI_j. Where $p_i|\mathbf{G}_i\mathbf{H}_i\mathbf{F}_i|^2$ represents the power of useful data, and $p_j|\mathbf{G}_i\mathbf{H}_i\mathbf{F}_j|^2$ represents the power of interfering data, $\|\mathbf{G}_i\|_2^2 N_0$ represents the noise power after the equalization in receiver end. P_0 denotes the transmit power of all data stream, which meets the following total power constraint

$$p_i + p_j = P_0. \tag{19}$$

To simplify the analysis, we assume that the system allocates power equally. When using linear minimum mean square error (MMSE) detection in the receiver end, we have

$$\mathbf{G}_i = \tilde{\mathbf{h}}_i^H(\tilde{\mathbf{H}}_i\tilde{\mathbf{H}}_i^{\mathrm{H}} + \frac{2N_0}{P_0}\mathbf{I}_N)^{-1}, \tag{20}$$

$$\tilde{\mathbf{h}}_i = \mathbf{H}_i\mathbf{F}_i, \tag{21}$$

$$\tilde{\mathbf{H}}_i = \mathbf{H}_i\left[\mathbf{F}_i\ \mathbf{F}_j\right], \tag{22}$$

Where $\tilde{\mathbf{h}}_i$ is the equivalent channel matrix of the i-th user. User j in the optimal pair can also calculate SINR_j in a similar way (user i is regarded as an interference).

Step 7 User i and user j in the optimal pair select the CQI_i and CQI_j according to SINR_i and SINR_j in the UE side, then feed back CQI_i, CQI_j, SINR_i, SINR_j respectively via uplink to eNB.

Step 8 eNB do MU-MIMO precoding according to feedback PMI indexes for user i and user j. The precoding matrix $\mathbf{W}$ is confirmed according to the PMI fed back from the optimal pair. Where $\mathbf{W} = [\mathbf{F}_i \quad \mathbf{F}_j] = \mathbf{V}\left(\mathbf{V}\mathbf{V}^H + \rho\mathbf{I}\right)^{-1}$, and $\mathbf{V}^H = [\mathbf{e}_i \quad \mathbf{e}_j]$. Data is transmitted to each UE from eNB since then.

4 Numerical Results

In this section, we will evaluate the simulation results of FDD MU-MIMO precoding scheme without or with D2D under the LTE physical layer standard. The proposed FDD MU-MIMO scheme is based on Rank-1 ZF beamforming.

We consider a simplified cellular network model for the performance evaluation. There are $N_T = 4$ antennas at the eNB and $N_R = 2$ antennas at the user terminal (UT). First of all, the achievable rate of rank-1 method and the proposed scheme are compared with the practical achievable rate at the UE side. Then, we evaluate the throughput in LTE network with adaptive modulation and coding. The details of the simulation parameters are given in Table 1.

Figures 2 and 3 give the comparison of practical SINR at UE side and estimated SINR at eNB for rank-1 ZF scheme and proposed D2D assisted MU-MIMO scheme. It is observed that there is a large gap between the curve of rank-1 scheme at eNB and the practical achievable rate at UE side, this is because that eNB predicts the CQI of paired UEs according to the CQI fed

Table 1. Parameters and settings

Parameters	Settings
Channel model	Flat Rayleigh channel/1.4 MHZ
UE feedback cycle	1 ms
CP	Conventional
eNB antenna configuration	4
UE antenna configuration	2
Codebook/codebook number	4 launch antenna of the Householder codebook with 16 vectors
Modulation coding	Adaptive CQI
FFT points	128
UE-side equalization algorithm	MMSE
Channel estimation	Ideal channel estimation
D2D link	Ideal

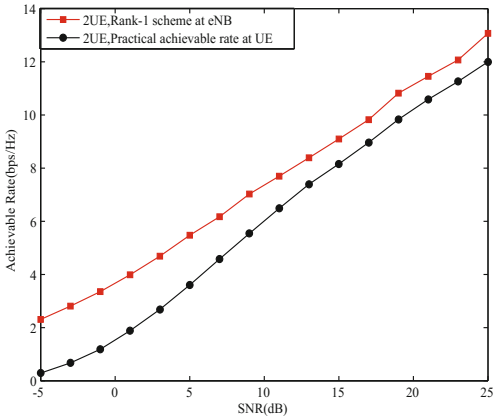


Fig. 2. The achievable rate of rank-1 method at eNB versus the practical achievable rate at UE side

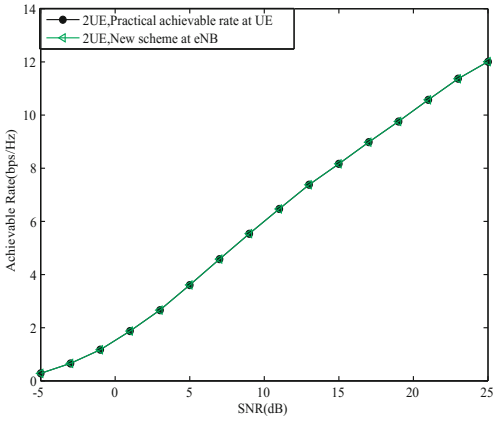


Fig. 3. The achievable rate of new scheme at eNB versus the practical achievable rate at UE side

back by each UE, thus the CQI mismatch problem arise, and the SINR estimated by eNB is not accurate. However, the curve of proposed scheme at eNB and the practical achievable rate at UE side are overlapped, which demonstrates that the CQI mismatch problem is solved in the proposed scheme, and the UE side can acquire accurate SINR from eNB.

Figure 4 compares the throughput of the proposed D2D assisted FDD rank-1 MU-MIMO precoding scheme versus rank-1 ZF precoding of 2UEs with adaptive MCS in LTE network. It can be observed that the proposed precoding method obtain superior throughput gain than the rank-1 ZF precoding method. Since the proposed scheme has tackled the CQI mismatch problem, the accuracy of

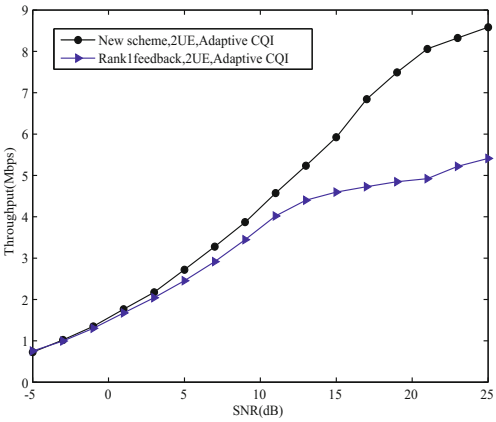


Fig. 4. Throughput of our scheme versus rank-1 ZF precoding method with adaptive MCS for 2UEs

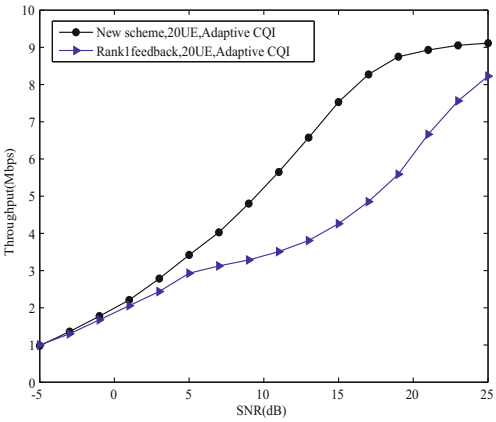


Fig. 5. Throughput of our scheme versus rank-1 ZF precoding method with adaptive MCS for 20UEs

CQI for each user is the reason for the performance improvement when the SNR is under high region.

Figure 5 gives the comparison between Rank-1 ZF precoding scheme and the proposed D2D assisted MU-MIMO precoding scheme with adaptive MCS for 20 UEs. As shown in Fig. 3, when the SNR is higher than 5 dB, the proposed scheme achieved significant throughput gain, which outperforms Rank-1 method. This is because with the accurate SINR, the UE side can calculate the best CQI for each paired UEs. Furthermore, this also lies in the appropriate communication mode of UEs, for UEs choose SU communication in the low SNR region, and UE pairs select MU communication in the high SNR region.

5 Conclusion

In this paper, we propose a novel D2D assisted MU-MIMO precoding under FDD mode to fully exploit the MU-MIMO system. In our scheme, UEs can accomplish pairing at the UE side, and only the successful matched UE pair feeds information back to the eNB, which can substantially reduce the uplink feedback and improve the pairing probability of MU-MIMO. Furthermore, the CQI mismatch problem is solved at the same time to enhance the throughput of MU-MIMO system in FDD mode. In order to estimate the performance of our scheme, we make a comparison with the conventional rank-1 scheme which does not employ the D2D technology, the result demonstrates the superiority of the proposed D2D asisted MU-MIMO precoding scheme.

Acknowledgment. This work is partially supported by the National Science Foundation of China (No. 61401074, No. 61771106), the Fundamental Research Funds for the Central University of China (No. ZYGX2016J019) and the Huawei Foundation (No. YB2015020023).

References

1. Yu, Y., Gu, D.: Enhanced MU-MIMO downlink transmission in the FDD-based distributed antennas system. IEEE Commun. Lett. **16**(1), 37–39 (2012)
2. Yin, H., Cottatellucci, L., Gesbert, D.: Enabling massive MIMO systems in the FDD mode thanks to D2D communications. In: 2014 Asilomar Conference on Signals, Systems and Computers, pp. 656–660 (2014)
3. Feng, D., Lu, L., Yi, Y.W., Li, G.: Device-to-device communications in cellular networks. IEEE Commun. Mag. **52**(4), 49–55 (2016)
4. Lin, X., Andrews, J., Ghosh, A., Ratasuk, A.: An overview of 3GPP device-to-device proximity services. IEEE Commun. Mag. **52**(4), 40–48 (2013)
5. Nam, J., Adhikary, A., Ahn, J.Y., Caire, G.: Joint spatial division and multiplexing: opportunistic beamforming, user grouping and simplified downlink scheduling. IEEE J. Sel. Top. Sig. Process. **8**(5), 876–890 (2014)
6. Adhikary, A., Nam, J., Ahn, J.Y., Caire, G.: Joint spatial division and multiplexing—the large-scale array regime. IEEE Trans. Inf. Theory **59**(10), 6441–6463 (2013)
7. Min, H., Lee, J., Park, S., Hong, D.: Capacity enhancement using an interference limited area for device-to-device uplink underlaying cellular networks. IEEE Trans. Wirel. Commun. **10**(12), 3995–4000 (2011)
8. Lin, X., Heath, R.W., Andrews, J.G.: The interplay between massive MIMO and underlaid D2D networking. IEEE Trans. Wirel. Commun. **14**(6), 3337–3351 (2015)
9. Sadek, M., Tarighat, A., Sayed, A.H.: A leakage-based precoding scheme for downlink multi-user MIMO channels. IEEE Trans. Wirel. Commun. **6**(5), 1711–1721 (2007)
10. Li, J.C.F., Lei, M., Gao, F.: Device-to-device (D2D) communication in MU-MIMO cellular networks. In: Global Communications Conference, pp. 3583–3587 (2012)
11. Baniasadi, M., Maham, B., Kebriaei, H.: Power control for D2D underlay cellular communication: game theory approach. In: 2016 8th International Symposium on Telecommunications (IST), pp. 314–319 (2016)

12. Zhong, W., Fang, B., Qian, Z.: Game theoretic resource allocation for D2D MIMO heterogenous networks. Wirel. Pers. Commun. **84**(1), 273–285 (2015)
13. Li, C.M., Hung, C.F.: Cooperative MIMO precoding for the LTE-A D2D communications. Wirel. Netw. **23**, 2545–2554 (2016)
14. Tang, H., Zhu, C., Ding, Z.: Cooperative MIMO precoding for D2D underlay in cellular networks. In: IEEE International Conference on Communications, pp. 5517–5521 (2013)
15. Wei, L., Hu, R.Q., He, T., Qian, Y.: Device-to-device (D2D) communications underlaying MU-MIMO cellular networks. In: GLOBECOM Workshops, pp. 4902–4907 (2013)
16. R1-093997, Implicit feedback in support of downlink MU-MIMO. Texas Instruments, RAN1#58bis, Miyazaki, Japan, October 2009



High Concurrent Elastic Resource Allocation in Hadoop YARN

Yang Peng, Danyan Luo(✉), Jian Dong, and Zhibo Wu

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
pengyang@ftcl.hit.edu.cn,
{luodanyan,dan,wzb}@hit.edu.cn

Abstract. Efficient resource management to improve the throughput in large-scale cluster has become a research focus with the rapid development of applications of Big Data. YARN (Yet Another Resource Negotiator), as the new generation of resource management system in Hadoop, is more efficient in resource utilization and capable of handling more kinds of workload than previous systems. Due to the fact that a task usually occupies more resources than it actually uses during some stage of its life cycle, a relevant amount of resource is idle and can not be allocated to satisfy the requirements of pending tasks. In order to address the deficiencies of resource allocation in YARN, this paper presents a high concurrent elastic resource allocation strategy named Ballon, which can dynamically adjust the configured resource of a node depending on the actual resource utilization of the node. Moreover, Ballon classifies resource requests of applications into different types. Consequently the elastic resources can be allocated to proper request. Our experiments demonstrate that Ballon cluster can reduce the average execution time of application by at least 10% in most MapReduce application and can increase the resource utilization of cluster.

Keywords: YARN · High concurrency · Elastic resource allocation

1 Introduction

With the rapid development of the Internet and related information industry, the amount of data is growing exponentially, and more and more fields are using big data processing technology as the development of its business [1]. In Hadoop 1.0, MRv1 [2], is regarded as the unitary Hadoop computing framework to deal with large-scale unstructured off-line data. However, with the increase of data scale and workload in distributed system cluster, the limitation of MRv1 is becoming more and more serious [3]. Therefore, Hadoop 2.0 proposed a new resource management system named YARN [4]. YARN abstracts the functions of resource management into an independent general system. It becomes a resource scheduler on which all of the computing framework can be run [5]. When the cluster starts up, each node declares its resource capacity and sends it to the ResouceManager, and applications from different frameworks will send the resource request to the ResourceManager. The ResourceManager checks the available resources and encapsulates each resource request to a fixed

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 524–534, 2018.
https://doi.org/10.1007/978-3-319-78130-3_54

container, and ensures that the resources occupied by containers do not exceed its capacity [6]. As a result of the fixed allocation method, a task usually occupies more resources than it actually uses during some stage of its life cycle. Therefore, YARN has a problem that the resources in container can not be fully used [7]. The elastic scheduling strategy of container is recognized as an effective method to solve above issue [8]. To improve the efficiency of YARN, Genkin and his colleagues [9] put forward a scheduler called KERMIT, which used a window to adjust the parameters of the running containers. Ding [10] presented a scheduler named JellyFish, which continually collected real-time statistics to optimize configuration and resource allocation dynamically during execution of a job. Shao [11] used the task prediction model named AR model to predict the workload of the next period. Yao [12] proposed a novel opportunistic and efficient resource allocation approach, named OpERA, which broke the barriers among the encapsulated resource utilization to reassign opportunistic available resources to pending tasks. Zhao [13] presented an elastic resource management approach, which can dynamically expand or shrink container's size to meet the actual resource needs of tasks. Lin [14] proposed an resource scheduler based on an ant-colony method, which can search more suitable parameters for resource allocation than YARN. Shah [15] proposed a new scheduling strategy based on cloud platform which can virtualize the cluster resource and reallocate it to the task. However, the above approaches are based on regulating each single container, and to adjust the container's size to an appropriate range so that ResourceManager can reallocate the optimized resources directly to the new pending tasks. In the actual environment, the task usually runs in the container as a process and the process has a great fluctuation in the use of resources at run time. The resource usage of a task may exceed the size defined by container at some point due to fluctuation factors while using above approaches, so that the running containers may be killed by ResoureManager and the scheduler's efficiency was reduced. What's more, the management of each container increases the workload of ResourceManager, resulting in an increased response time. Moreover, some of the above approaches did not consider the key problem that the dynamic change of resources in the node may lead to the overload of server node resources, and incur interference and degrade the performance of all active tasks. To address above issues, this paper presents a high concurrent elastic resource allocation strategy named Ballon, which can dynamically adjust the configured resource of a node depending on the actual resource utilization of the node, and Ballon classifies resource requests of applications into different types. Consequently the elastic resources can be allocated to proper request. Moreover, Ballon presents a series of solutions to solve the key problem of incurring interference in elastic scheduling strategy: (1) monitors the actual resource utilization at run time. (2) schedules the resource of node flexibly on the basis of real resource utilization. (3) divides the resource request to different types and determine whether this type of request has the opportunity of being scheduled. (4) arranges the task reasonably by using idle resource to mitigate the severe resource contentions due to resource over-provisioning.

The reminder of this paper is organized as follows. Section 2 describes the structure of our new high concurrent elastic resource allocation. Section 3 presents the different performance comparison between our new approach and existing resource allocation approach in various testing sets. We finally conclude the paper in Sect. 4.

2 The Design of Ballon

In YARN, a job is a program instance that is submitted to a Hadoop system. We assume that every job is finite, which means that every job will be finished in a certain time period with enough resources. A task is a subprocess in a job. We define that the volume of a task is based on the amount of resources it occupies and its execution time, and the total volume of a job is the sum of all the sub-tasks. This paper focuses on the calculation framework of MapReduce, and tasks in a MapReduce job can be categorized into two subsets: map tasks and reduce tasks. Thus the volume of a MapReduce job can be calculated as Eq. 1:

$$S = AUMiMap * SEm + AUMi\ \mathrm{Re} * SEr \tag{1}$$

where S denotes the required volume of completing a MapReduce job. *AUMiMap* denotes the actual resource utilization of all map tasks. *AUMiRe* denotes the actual resource utilization of all reduce tasks. *SEm* denotes the total time of the map tasks cost. *SEr* denotes the total time of the reduce tasks cost. In general, the volume of completing a job can be regarded as a constant in the same experimental environment. It can be concluded that the actual resource utilization ratio is inversely proportional to the execution time of tasks from Eq. 1. Thus, Ballon adds three models on the basis of YARN to improve the utilization of resource.

The overall sketch of Ballon is shown in Fig. 1. Ballon adopts the same double-level scheduling mechanism as YARN. The ResourceManager allocates requested resource to the ApplicationMaster in first layer, and the ApplicationMaster assigns the allocated resources to the internal container in the NodeManager in second layer. The added module in NodeManager named NodeSigar Monitor, can combine total resource of the node with the actual resource utilization in the node's operating system to generate the elastic resource. The other added module in ResourceManager named ResourceRequest Classifier, can classify resource requests of applications into different types. Consequently the elastic resources can be allocated to proper request in the added module named Ballon Scheduler in ResourceManager.

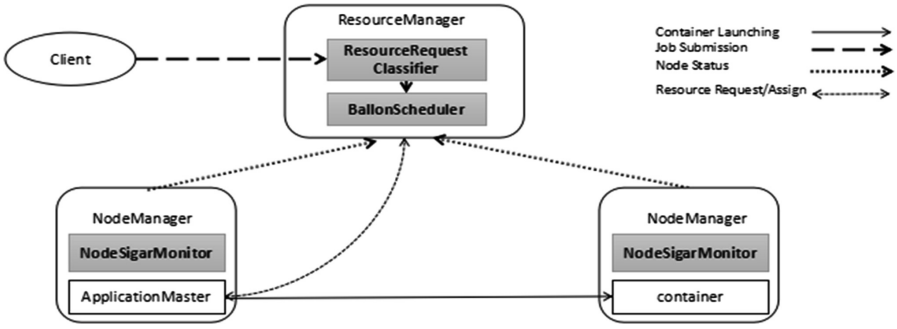


Fig. 1. The architecture of the Ballon-based YARN framework.

2.1 NodeSigar Monitor

NodeSigar Monitor can monitor the utilization of resources of the all containers and the actual utilization ratio of resources on operating system of the node, and then schedule the resources of the node dynamically according to the actual resource utilization ratio it monitors. In order to combine the resource scheduling management of Ballon with the actual resource utilization ratio of each node at run time, we propose the concept of virtual memory, virtual CPU cores, real memory and real CPU cores in Ballon.

Virtual memory/CPU cores: YARN reads it from the configuration file in each node, and reports it to ResourceManager when NodeManager starts. It is a flexible scheduling value.

Real memory/CPU cores: the value of the actual resource in the node's operating system. It is a fixed value.

We add a byte to the field in the heartbeat from NodeManager to ResourceManager, this byte includes the virtual resource of the NodeManager and the actual average utilization ratio of resource within a period of time, and actual average utilization ratio of memory is calculated as Eq. 2:

$$AUMi = \frac{\sum_{lastNb}^{Nb} (total - free - cached - buffers)}{total * (Nb - lastNb)} \tag{2}$$

where *AUMi* denotes the actual average memory resource utilization ratio of the *i*-th node in the two heartbeat intervals. *Nb* denotes the heartbeat time nearly. *LastNb* denotes the last heartbeat time. *Total* denotes the actual amount of memory of the node. *Free* denotes the actual total amount of free memory at a time. *Cached* denotes the amount of memory in file buffer at a time. *Buffers* denotes the amount of memory in disk buffer at a time. Meanwhile, NodeSigarMonitor also monitors the usage of CPU resources, the monitored indicator is the average load of the CPU, namely the average number of processes which CPU runs in the ready state during last time period. Thereby NodeSigar Monitor can obtain the average CPU load of the node over a period of time through the indicator.

After the NodeSigar Monitor obtains the average value of the actual resource (i.e., memory, CPU) usage of the system during the last period of time, and the value will determine whether the resource utilization rate reaches the critical value of flexible scheduling. If not, there is no need for flexible scheduling in the node. Otherwise, virtual resource of the node will be rescheduled by reading the parameters inside the relevant configuration file (i.e., critical value of utilization, expanding/decreasing value of virtual resource). After the completion of scheduling, NodeSigar Monitor will determine whether there exists resource inclination according to the virtual CPU usage and virtual memory usage. The resource inclination denotes that the usage of a resource far exceeds the usage of another resource in a node, resulting in an imbalanced usage of resources at that node. Then NodeSigar Monitor needs to transmit this imbalance to ResourceManager, so that ResourceManager of Ballon can assign reasonable tasks to mitigate this imbalance for the node.

The specific parameters (i.e., expanding/decreasing value of virtual resource, critical value) in the configuration file can be designed by the user according to their own needs, then NodeSigarMonitor introduces these parameters into the resource elastic allocation algorithm to schedule the virtual resource of the node.

```
Resource Elastic Allocation Algorithm(NodeStatus)
  {List<ContainerStatus> containerStatus};
  conStatus = getContainerStauts(containerStatus);
  for(i=0;i<conconStatus.length;i++ )
      conResVores = get virtual cpu vores;
      conResMem = get virtual memory;
  for(N heartbeat interval)
      AUMi = getAUMi();
      cpuLoad[] = Nodesigar.getLoadAverage;
  if(conResMem>expanding*totalMem&&conResVores<
totalCores)
     if(AUMi<realMem/totalMem)
        totalMem = totalMem * expand;
  if(conResVcores>expanding*totalCores&&conResMem<
totalMem)
     cpuLoad[0] = get nearest CPU load
     if(cpuLoad[0]/realVores < MinLoad)
        totalVcroes = totalVcores * expand;
  VcoresPerpent = get virtual utilization of cpu;
  MemPerpent = get virtual utilization of memory;
  if(VcoresPerpent >> MemPerpent )
    node.setFlag(1);//exist io-resource inclinaiton
  else if(MemPerpent >> VcoresPerpent )
    node.setFlag(2);//exist cpu-resource inclinaiton
  else node.setFlag(0);
  return nodeStatus;
```

After NodeSigar Monitor updates NodeStatus according to the algorithm, the updated information will be transmitted to the ResourceManager through the NodeHearteBeat in the way of remote invocation. Then the ResourceManager will reassign reasonable type of container to the node according to the transmitted information.

2.2 ResourceRequest Classifier

When user submits application to YARN cluster, ResourceManager will start a container in a random NodeManager as a task submitted tool called ApplicationMaster. ApplicationMaster submits resource requests to ResourceManager to get the allocation of resources continuously. ResourceRequest Classifier is expected to classify all the pending resource requests into four types (i.e., small ioDemand, large ioDemand, small cpuDemand and large cpuDemand), which will decide whether the resource request can be allocated to the idle resource produced by elastic resource allocation scheduling in NodeSigarMonitor. Therefore, the design of the ResourceRequest Classifier includes

three aspects: (1) confirms the resource-demand type of a resource request. (2) predicts the volume of a resource request. (3) compares the volume with a threshold to determine the volume type it belongs to. Therefore, the accuracy of classifier is the key to ResourceRequest Classifier.

Determining resource-demand type is relatively easy according to the characteristics of the MapReduce computing framework. The computing framework of MapReduce is divided into two stages, namely Map stage and Reduce stage. For each resource request of Map or Reduce, the demanded resources of the request is constant. Therefore, after application is submitted to YARN cluster, ResourceRequest Classifier will identify all resource requests in the application and determine the resource-demand type of each request according to the ratio of the resources in each request. However, the volume of resource request is difficult to be defined, because the task execution time is difficult to be predicted, and the specific execution time depends on the allocation of resources and a series of operating parameters. Through the analysis for the characteristics of MapReduce computing framework, we find that different input data sizes could produce only a different number of tasks. Considering a cluster processing two sort applications, one with 20 GB input data and the other with 100 GB input data, all resource request of these two applications will process similar size of input files, and the only difference between them is the total numbers of MAP and REDUCE task. Thus, we can make a conclusion that the volume of the request in one smaller application is same with the other in larger application. Thus, ResourceRequest Classifier could use the historical data of the same type of request in the same application to predict the volume of the resource request to be executed. Classifier should have the characteristics of high accuracy and low overhead in storage and computing. So we adopt the naive Bayes classifier to predict the volume of resource request. The following six properties of each resource request are considered by classifier to identify the volume type of the request.

- Fr: the framework of resource request.
- Ar: the application of resource request.
- Sr: the stage of application that the resource request belongs to.
- Mr: the memory resource of resource request.
- Cr: the CPU resource of resource request.
- Pr: the progress of the resource request.

According to the algorithm of naive Bayes, once the resource request's features (i.e., P(Sr|C)) are determined, we can calculate the posterior probability (i.e., P(C|Sr)) by using Eqs. 3 and 4.

$$P(\mathrm{Cr}|Fr) \propto P(Cr) * P(Fr|Cr) \tag{3}$$

$$P(\mathrm{Fr}|Cr) = \prod_{i} P(F_{\mathrm{i}}^{r}|Cr) \tag{4}$$

According to the classification algorithm of naive Bayes, the resource request submitted by ApplicationMaster is divided into two types, small and large. Then ResourceManager will encapsulate the resource request into a container, and we can

monitor the various parameters of the container when it runs on the NodeManager. At last, we can calculate the correctness of our classification according to the value.

2.3 Ballon Scheduler

Through above two sections of this paper, during the time of running application, NodeSiagarMonitor adds a flexible scheduling of resources to node to generate idle resources. Moreover, ResourceRequest Classifer divides resource request submitted by application into four types that can be allocated by resource allocation strategy. Therefore, there are two kinds of resources in Ballon scheduler, that is, the normal resources owned by the normal cluster and the elastic resources generated by the NodeSiagarMonitor, resulting that there should be two different strategies of allocating resources in Ballon scheduler. In existing ResourceManager of YARN, there are three kinds of resource scheduler according to different requirements, namely FIFO Scheduler, Capacity Scheduler and Fair Scheduler. All three schedulers use a two-level scheduling model, in which the Capacity and Fair scheduler is a multi-queue scheduler.

Considering the drawback that the cluster resources are not fully used, Ballon Scheduler modifies the original normal resource allocation strategy and increases the allocation strategy of elastic resources on the basis of Capacity Scheduler. For all resources in cluster, if there exists resource inclination, the resource request of the weaker part is given priority to be allocated, otherwise, different type of resources have different assignments. The normal resource allocated for each node is calculated as Eq. 5, and the elastic resource allocated for each node is calculated as Eq. 6:

$$AR_i = C_i - \sum_{\mathrm{t}\in Rn} U_t \tag{5}$$

where *ARi* denotes the available capacity of normal resource in node *i*. *Ci* denotes the total capacity of normal resource in node *i*. *Ut* denotes the capacity of denotes the capacity of resource occupied by running container. *Rn* denotes the running container list.

$$ER\mathrm{i} = ACi - Ci - \sum_{t\in St} U_t \tag{6}$$

where *ERi* denotes the remaining available elastic resources for node *i*. *Aci* denotes the total capacity of resource in node *i*. *St* denotes the list of specific category of resource requests that could be allocated to use elastic resources.

3 Evaluation

3.1 Experimental Setup

We use five physical servers to build an experimental cluster. The cluster includes five nodes. Every node is a virtual machine that is virtualized by XenServer technology in a physical server. Each node contains 8G memory and 12 CPU cores, and the node is

equipped with operating system of CentOS 6.5. A node is set as the master node of our cluster, the other four nodes is set as the slave nodes. We use HiBench 4.0 to synthesize the workload used in our experiments. We use some of the benchmarks shown in Table 1 to compare the performance and resource utilization in Capacity Scheduler of YARN and Ballon.

Table 1. Benchmarks of HiBench 4.0

Application	Resource type	Input data size
WordCount	CPU intensive	50 GB
TeraSort	Map: CPU intensive	50 GB
	Reduce: IO intensive	
Sort	CPU intensive	40 GB
TestDFSIO	IO intensive	80 GB

3.2 Single Job Performance Evaluation

In this section, we submit all of the application in Table 1 to YARN and Ballon individually, and record specific information produced at the running process. Table 2 shows the prediction accuracy of ResourceRequest Classifier after running all the above applications.

Table 2. ResourceRequest classification

	Actual	Classification		
	Request numbers	Small	Large	Accuracy
Small	2328	2028	300	0.8711
Large	364	16	348	0.9540

We observe that our ResourceRequest Classifier is able to accurately categorize most of small and large request with high prediction accuracy ratios, i.e., 87.11% and 95.4%, respectively. Though there is a probability of almost 13% of the small type of request to be mistaken for the large attribute, the error ratio is still in a low range and could prevent us from assigning elastic available resources to this kind of request, which in general will not incur severe resource contention.

The experimental time of each benchmark in the YARN scheduler and Ballon scheduler are shown in Fig. 2. As can be seen from Fig. 2, the testing sets of HiBench, all have been improved in varying degrees in Ballon, and the running time is shortened. Among them, the improvement of performance of WordCount and Sort are obvious, and Ballon scheduler saved 17% of running time at most. There are two reasons for this phenomenon: firstly, these two kinds of tasks have low memory requirement and low actual memory utilization, so that the virtual resources of nodes can be flexibly scheduled by Ballon Scheduler, and the elastic resource can be allocated for extra tasks. Secondly, for these two tasks, most of the request's volume are classified as Small in

ResourceRequest Classifier, so these requests can use the elastic resources with higher priority. But for the application of TestDFSIO, most volume its resource requested are classified as Large, and the range of elastic resource is flat due to the tasks with high memory requirement and large actual memory utilization, thus the running time Ballon optimized offsets the time Ballon spending on Scheduling. The comparison of the concurrency of containers allocation between YARN and Ballon scheduler in the cluster for most application is shown in Fig. 3. We can clearly find that the Ballon Scheduler uses the elastic resource to achieve higher concurrency and shorten the running time.

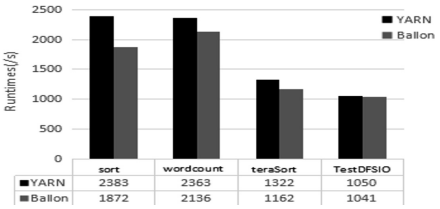


Fig. 2. The performance in YARN and Ballon

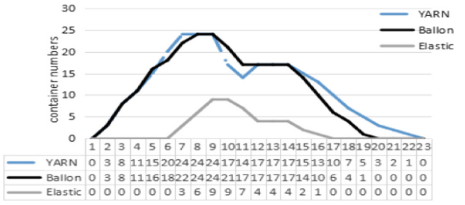


Fig. 3. The containers in YARN and Ballon

3.3 Multiple Jobs Performance

Based on above single job test, the results show that the Ballon scheduler can improve the efficiency of IO-intensive tasks and CPU-intensive tasks. However, in the actual application process, the cluster is often used by multiple users and each user will run different computing framework of application to cluster. Therefore, the design of multiple jobs' test can better demonstrate the practical application of Ballon scheduler's performance. In the single job experiment, the optimized performance of above two types of jobs in Ballon scheduler is different. So we take different combinations of different types of applications to test.

Figure 4 presents that the Ballon scheduler has different degrees of improvement compared with the YARN scheduler in multiple job using case. The difference between single job testing and multiple jobs testing is that the Ballon scheduler greatly improves the efficiency of the combinational applications of CPU intensive and IO intensive. There are three key reasons for this phenomenon: firstly, in the case of multiple applications concurrency, CPU-intensive application requires high demand of CPU cores, but the demand for memory is relatively low. IO-intensive application requires high demand of memory, but the demand for CPU cores is relatively low. Thus in the process of node's resource scheduling, the resource demanded by the two different types of applications are complementary, resulting in a high resource utilization ratio and concurrency of cluster. Secondly, there are multiple types of resource requests, and some specific types of requests could be allocated to idle resource produced by elastic resource scheduling in Ballon scheduler. Moreover, the Ballon scheduling strategy can eliminate the resource inclination of cluster, resulting in increasing the resource utilization of cluster in another way.

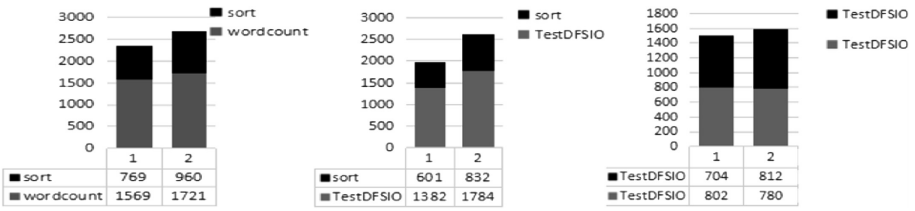


Fig. 4. The time of combinatorial applications in YARN and Ballon

4 Conclusion

In this paper, we presents a high concurrent elastic resource allocation named Ballon, which can dynamically adjust the configured resource of a node depending on the actual resource utilization of the node, and can use this elastic resource to manage the task reasonably. The main features of Ballon are as follows: (1) Ballon takes a comprehensive account of various factors on the virtual resources of YARN. (2) Ballon combines the total virtual resource with the actual resource utilization (i.e., *AUMi* and CPUload) in the node's operating system. (3) Ballon has classified the application's resource request to grant some kinds of resource request the right to use the elastic resource. (4) Ballon can mitigate the severe resource contentions due to resource over-provisioning. With the result of experiments, we prove that the Ballon scheduler can improve cluster utilization and reduce the average running time of the submitted applications.

References

1. Shvachko, K., Kuang, H., Radia, S., et al.: The Hadoop distributed file system, pp. 1–10 (2010)
2. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. In: Conference on Symposium on Operating Systems Design & Implementation. DBLP (2004)
3. Frampton, M.: Storing and configuring data with Hadoop, YARN, and ZooKeeper. In: Frampton, M. (ed.) Big Data Made Easy, pp. 11–56. Apress, Berkeley (2014). https://doi.org/10.1007/978-1-4842-0094-0_2
4. Huang, W., Meng, L., Zhang, D., et al.: In-memory parallel processing of massive remotely sensed data using an Apache Spark on Hadoop YARN model. IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens. **10**, 1–17 (2017)
5. Vavilapalli, V.K., Murthy, A.C., Douglas, C., et al.: Apache Hadoop YARN: yet another resource negotiator. In: Symposium on Cloud Computing, p. 5 (2013)
6. Eadline, D.: Preface to Apache Hadoop YARN: Moving beyond MapReduce and Batch Processing with Apache Hadoop 2. Pearson Schweiz Ag, Zug (2014)
7. Liu, Y.: High availability of network service on docker container. In: International Conference on Measurement, Instrumentation and Automation (2016)

8. Lee, S., Bae, M.-H., Eum, J.-H., Oh, S.: Efficient vCore based container deployment algorithm for improving heterogeneous Hadoop YARN performance. In: Kim, K., Joukov, N. (eds.) ICISA 2017. LNEE, vol. 424, pp. 191–201. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-4154-9_23
9. Genkin, M., Dehne, F., Pospelova, M., Chen, Y., Navarro, P.: Automatic, on-line tuning of YARN container memory and CPU parameters. In: IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Sydney, NSW, pp. 317–324 (2016)
10. Ding, X., Liu, Y., Qian, D.: JellyFish: online performance tuning with adaptive configuration and elastic container in Hadoop YARN. In: 2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS), Melbourne, VIC, p. 831 (2015)
11. Shao, Y., Li, C., Dong, W., Liu, Y.: Energy-aware dynamic resource allocation on Hadoop YARN cluster. In: IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Sydney, NSW, pp. 364–371 (2016)
12. Yao, Y., Gao, H., Wang, J., Mi, N., Sheng, B.: OpERA: opportunistic and efficient resource allocation in Hadoop YARN by harnessing idle resources. In: 2016 25th International Conference on Computer Communication and Networks (ICCCN), Waikoloa, HI, pp. 1–9 (2016)
13. Zhao, Y., Wu, G.: Yadoop: an elastic resource management solution of YARN. In: 2015 IEEE Symposium on Service-Oriented System Engineering, San Francisco Bay, CA (2015)
14. Lin, J., Lee, M.: Performance evaluation of job schedulers on Hadoop YARN. Concur. Comput. Pract. Exp. **28**(9), 2711–2728 (2016)
15. Shah, P.: Adaptive application master for elastic web server farms for cloud based on Hadoop YARN. In: International Conference on Cloud Computing and Big Data, pp. 461–446 (2013)
16. Kakantousis, T.: Scaling YARN: a distributed resource manager for Hadoop (2015)



An Autoencoder-Based Learning Method for Wireless Communication Protocol Identification

Jie Ren[1], Zulin Wang[1,2], and Mai Xu[1(✉)]

[1] School of Electrical and Information Engineering, Beihang University, Beijing, China
{rj278007,wzulin,maixu}@buaa.edu.cn
[2] Collaborative Innovation Center of Geospatial Technology, 129 Luoyou Road, Wuhan, China

Abstract. As protocols play respective roles to fulfill different communication services, it is important to identify protocols before analyzing and managing the system. In the past decade, there have been a lot of researches on protocol identification using machine learning methods, which achieve promising results. However, the features of protocol used for identification mainly rely on engineering skill and domain expertise, which may not be available for the complicated wireless communication systems, such as encryption-based systems. In this paper, we propose an unsupervised-based learning method to make the feature extraction more intelligently and automatically. We first review the limitation of the traditional identification methods, especially the part of feature extraction. After that, an unsupervised deep learning based method, autoencoder, is proposed for automatically extracting the features of the original protocol data. Then, we construct the identification model based on the extracted features and a Support Vector Machine based classifier. Finally, experimental results show the effectiveness of the proposed method.

Keywords: Protocol identification · Feature extraction
Unsupervised learning · Autoencoder · Support vector machine

1 Introduction

In a wireless communication system, MAC protocols play important roles in controlling the channel states and offering different services. Since the heterogeneous network and big data become a prominent trend in modern wireless communication system, the analysis, control and management of the system need to identify the protocols in real time. In this case, the protocol identification technology is widely used for the management of networks. The traditional methods, such as port-based method [1], statical feature mapping method [2], and deep packet inspection(DPI) method [3], have been proposed in identifying some particular types of protocols. However, these methods can be only applied in some certain

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 535–545, 2018.
https://doi.org/10.1007/978-3-319-78130-3_55

conditions, which is impractical in the modern wireless system [4]. The reasons can be summarized as follows.

- Such applications use the dynamic ports for communication, some of which may not be registered in the Internet Assigned Numbers Authority (IANA).
- The encryption technologies have been widely used in modern wireless communication systems.
- The features extracted in some systems are not available in other systems.

Machine learning is one of research fields of artificial intelligence, and the most useful application of machine learning is classification. Meeting the above challenges, machine learning methods [5–7] have been introduced in protocol identification, seen as a kind of classification. In these methods, the signal-level and bit-level characteristics taken from the traffic flows can be regarded as the input to the classification model, and map the traffic flows into different protocol categories [8,9]. In the past decade, a large number of machine learning methods have been applied for protocol identification, such as neural network (NN) [8], hidden markov model (HMM) [10] and support vector machine (SVM) [11]. The experimental results show the effectiveness when using the machine learning based (ML-based) identification methods.

However, the existing ML-based protocol identification methods usually have to extract the features by some hand design feature extractors. In order to make a high performance feature extractor, the prior information of the systems and the expert experience are indispensable for the designers. This leads to inefficiency in analyzing the signal-level and bit-level features. Besides, the existing ML-based methods can be only used in some certain systems, not suitable for other systems or unknown systems. Thus, the traditional feature extractor is always restricted for its resource consumption and feasibility limitation [12].

In this paper, we propose a new feature extractor based on an unsupervised deep learning method, i.e., autoencoder [13]. Its architecture consists of two parts, i.e., encoder and decoder. The former is used to learn the representation of the original data, while the latter is used to reconstruct the input data. Here, we propose the stack autoencoders to construct the feature extractor. After that, we apply the SVM-based classifier to map the protocols into different types. In the simulation, we use real-world data of the most common 3G wireless system, WCDMA, and select three types of the protocols to evaluate the proposed method. The results suggest that the proposed method can automatically extract features, feasible for protocol identification.

The remainder of this paper is organized as follows. Section 2 describes the scheme of our method, which mainly consists of feature extraction and protocol identification; Sect. 3 shows the experimental results; Finally, Sect. 4 concludes the work.

2 Proposed Method

The process of our method can be divided into feature extraction and classification. In this section, an unsupervised deep learning method is developed for

feature extraction, and a supervised machine learning method is applied for classification. The framework of our method for protocol identification is shown in Fig. 1.

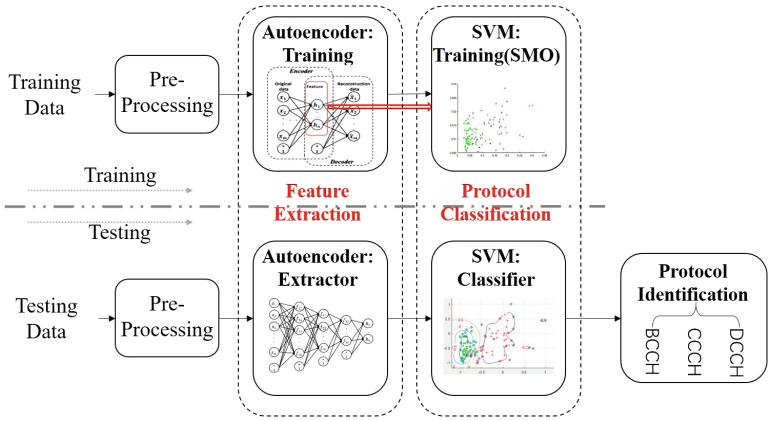


Fig. 1. The framework of the proposed method, which consists of the feature extractor of autoencoder and the classifier of SVM.

2.1 Feature Extractor

Now, we make use of the unsupervised deep learning method, autoencoder, as our feature extractor. In order to demonstrate our method, some notations and their explanations are shown in Table 1.

Table 1. Notation and explanation

Notation	Explanation
$\boldsymbol{S}$	The data set of all protocols
m	The dimension of the input layer
n	The dimension of the hidden layer
N	The number of protocol samples
L	The largest length of protocol sample
$\boldsymbol{x}^{(i)}$	The original data of i-th protocol sample
$y^{(i)}$	The type of i-th protocol sample
$\boldsymbol{h}^{(i)}$	Output of the feature extractor

Autoencoder is a multi-layer neural network that attempts to copy its input to its output [14]. In this model, it consists of an input layer (original data),

an output layer (reconstruction data) and a hidden layer (representation). For reconstructing the input data, the dimension of output layer is the same as the input layer. For extracting the discriminative features from the original data, the dimension of hidden layer is smaller than the input layer ($n < m$). Assume that there is a protocol sample $(\boldsymbol{x}^{(i)}, y^{(i)}) \in \boldsymbol{S}$, where $\boldsymbol{x}^{(i)}$ is the data and $y^{(i)}$ is the type of protocol. The network can be divided into two parts: an encoder and a decoder. The features after encoding $\boldsymbol{x}^{(i)}$ in the encoder is denoted as

$$\boldsymbol{h}^{(i)} = f(\boldsymbol{x}^{(i)}) = \sigma_1(\boldsymbol{W}\boldsymbol{x}^{(i)} + \boldsymbol{b}), \tag{1}$$

for mapping the sample from original data space to feature space $\boldsymbol{h}^{(i)}$. The decoder that maps the data from feature space to original data space is denoted as

$$\hat{\boldsymbol{x}}^{(i)} = g(\boldsymbol{h}^{(i)}) = \sigma_2(\boldsymbol{W'}\boldsymbol{h}^{(i)} + \boldsymbol{b'}). \tag{2}$$

In (1) and (2), $\boldsymbol{W}$ and $\boldsymbol{W'}$ are the weight matrixes of the encoder and decoder network, respectively. $\boldsymbol{b}$ and $\boldsymbol{b'}$ are the biases of each network. σ_1 and σ_2 are activation functions. In order to obtain the optimal solution, the learning process is described as minimizing a reconstruction error function, written as

$$Loss = \frac{1}{N}\sum_{i=1}^{N} \parallel \boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)} \parallel^2 = \frac{1}{N}\sum_{i=1}^{N} \parallel \boldsymbol{x}^{(i)} - \sigma_2(\boldsymbol{W'}\sigma_1(\boldsymbol{W}\boldsymbol{x}^{(i)} + \boldsymbol{b}) + \boldsymbol{b'}) \parallel^2 . \tag{3}$$

The procedure of training the autoencoder is summarized in Algorithm 1 [15].

Algorithm 1. The training of autoencoder.

Input: N: the number of total protocol samples. M: batch size. K: number of batch($K * M \leq N$).
Output: weights: $\boldsymbol{W}$ and $\boldsymbol{W'}$, biases: $\boldsymbol{b}$ and $\boldsymbol{b'}$, extracted feature: $\boldsymbol{h}$.
Initialize: weights initialization, draw randomly from a Gaussian or uniform distribution.
For: $k = 1, 2, ..., K$.

- Feed-forward: Select a batch of protocol samples $\boldsymbol{x}^{(ki)}, i = 1, ..., M$, compute activations at hidden layer $\boldsymbol{h}^{(ki)}$, and then obtain the output $\hat{\boldsymbol{x}}^{(ki)}$ at output layer;
- Loss function calculation: Calculate the deviation of output $\hat{\boldsymbol{x}}^{(ki)}$ from the input $\boldsymbol{x}^{(ki)}$, and then the loss function is $Loss = \frac{1}{M}\sum_{i=1}^{M} \parallel \boldsymbol{x}^{(ki)} - \hat{\boldsymbol{x}}^{(ki)} \parallel^2$;
- Back-propagate: Update the weights($\boldsymbol{W}$ and $\boldsymbol{W'}$) and biases($\boldsymbol{b}$ and $\boldsymbol{b'}$) of each layer.

End

In order to obtain the high level representation, we propose a deep architecture by stacking 4-layer autoencoder, which stack one on top of another. Assume that $\boldsymbol{f}^{(1i)}$ is the representation of the first-layer autoencoder. Then, it is used as the input to the second-layer autoencoder. Here, we denote the representation of

the forth-autoencoder as $\boldsymbol{f}^{(4i)}$, which can be used as the input to a classification model, such as SVM.

Therefore, we construct a feature extractor of 4-layer autoencoder to obtain the high level representation of original protocol data. Back-propagation algorithm is utilized to train the model, and the output of this feature extractor is denoted as $\boldsymbol{h}^{(i)} = \boldsymbol{f}^{(4i)}$. Given the extracted features of protocols, we need to classify these protocols into different types.

2.2 Classifier

SVM is a widely used supervised classification method to separate different kinds of data by searching a hyperplane or set of hyperplanes in a high- or infinite-dimensional space [16]. Usually, data in two-dimensional feature space can be separated by line, and three-dimensional can be separated by plane. Similarly, data in N-dimensional feature space can be separated by $(N-1)$-dimensional hyperplane. Assume that each sample is denoted as $(\boldsymbol{h}^{(i)}, y^{(i)})$, where $\boldsymbol{h}^{(i)}$ represents the protocol features obtaining from the feature extractor, and $y^{(i)}$ represents the type of protocol. Thus, the hyperplane can be expressed as

$$\boldsymbol{w} \cdot \boldsymbol{h}^{(i)} + d = 0, \tag{4}$$

where $\boldsymbol{w}$ is the weights and d is the bias.

In order to find the optimal solution, the SVM model is processed by maximizing the margin of the classifier, as well as solving the dual problem in

$$\begin{gathered} \min \tfrac{1}{2} \parallel \boldsymbol{w} \parallel^2 + C \textstyle\sum_{i=1}^{m} \xi^{(i)}, i = 1, ..., N \\ s.t. \;\; y^{(i)}(\boldsymbol{w}^T h^{(i)} + d) \geq 1 - \xi^{(i)}, i = 1, ..., N \\ \xi^{(i)} \geq 0, i = 1, ..., N, \end{gathered} \tag{5}$$

where $\xi^{(i)}$ is the slack and parameter C determines the tradeoff between increasing margin-size and ensuring samples lie on the correct side. We use the Radial Basis Function (RBF) to map the features from a low-dimensional space to a high-dimensional space, in which the data can be separated. The RBF kernel function is as follows,

$$K(\boldsymbol{h}^{(i)}, \boldsymbol{h}^{(j)}) = e^{-\gamma \| \boldsymbol{h}^{(i)} - \boldsymbol{h}^{(j)} \|^2}, \gamma \geq 0, \tag{6}$$

where $\boldsymbol{h}^{(i)}$ and $\boldsymbol{h}^{(j)}$ are features of two arbitrary samples in the input set $\boldsymbol{S}$. γ is the parameter of kernel function.

Therefore, the optimization of this classifier can be solved by lagrange multiplier and Karush-Kuhn-Tucker conditions as the following formulation,

$$L(\boldsymbol{w}, d, \xi, \alpha, \gamma) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi^{(i)} - \sum_{i=1}^{N}\alpha_i[y^{(i)}(\boldsymbol{w}^T\boldsymbol{h}^{(i)} + d) - 1 + \xi^{(i)}] - \sum_{i=1}^{N}\gamma_i\xi^{(i)}, \tag{7}$$

where α_i and γ_i are both lagrange multipliers. Then, we use the sequential minimal optimization (SMO) algorithm to train the model.

3 Experimental Results

For the task of protocol identification, we developed a 4-layer autoencoder as our feature extractor, and the SVM as our classifier. To evaluate the performance of our method, we considered the real-world data of the most common 3G wireless communication system, WCDMA, and selected three most important logical channel protocols, BCCH (Broadcast Control Channel), CCCH (Common Control Channel) and DCCH (Dedicated Control Channel). Also, we compared the results with the traditional protocol identification methods.

3.1 Data Collection

Training data is the foundation for machine learning, especially for deep learning. Here, we used a wireless communication terminal, typed Sony Errison K800i, and the corresponding software, TEMS, to collect the communication data. The terminal is able to support GSM and WCDMA communication system, while the software is an advanced cellular network diagnostics tool to collect and analyze the communication information. In our experiment, we only acquired the data of WCDMA to verify our method. Totally, we spent about 3 weeks, more than 504 hours to collect the WCDMA communication data. During this time, there were more than 400 times of the calls, either inbound or outbound, in total about 375 min for all of these calls. The collected protocol samples used in our experiment are shown in Table 2.

Table 2. Data set of our experiment

Protocol type	Length of protocol	Number of protocols
BCCH	1,029	2,015
CCCH	1,015	3,113
DCCH	960	4,815

Since the longest protocol sample was 1029 (bits), we set the dimension of the input layer to 1029 ($m = 1029$), while the short ones were filled with zero at the tail. In order to balance the number of different types of protocols, we randomly selected 2000 samples from each type of protocol.

3.2 Identification Model

After collecting data, we constructed the protocol identification model based on the proposed method of Sect. 2. The model consists of two parts, the feature extractor and the classifier. First, we constructed a 4-layer autoencoder to extract features from the original data. Afterwards, we used these features as the input to the SVM-based classifier for identifying different types of protocols. The architecture of this model is shown in Fig. 2.

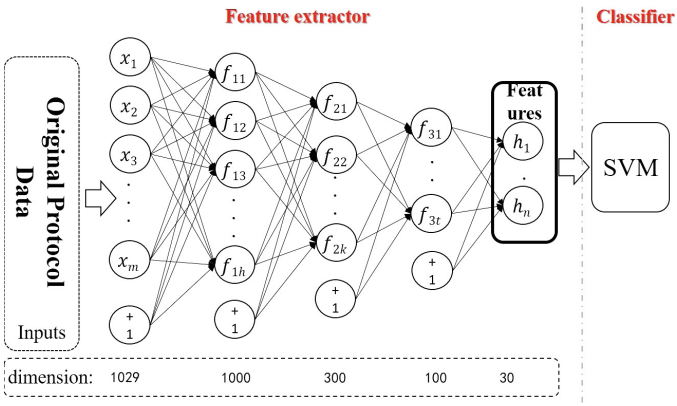


Fig. 2. The structure of our protocol identification model, which consists of two parts: feature extractor and classifier.

The first part is a feature extractor. It was trained by a 4-layer autoencoder, and we used the part of encoder as our feature extractor. In Fig. 2, $\boldsymbol{f}_{i*}$ is the representation of the i-th layer autoencoder, and $\boldsymbol{h} = \boldsymbol{f}_{4*}$ is the output of the feature extractor. The dimension of the input layer was 1029, while the dimension of the four hidden layers were 1000, 300, 100 and 30. The 30-dimension output feature vector $\boldsymbol{h}$ was then used as the input to the SVM classifier. In our experiment, we trained the model by minimizing the reconstruction error, with back-propagation algorithm.

In the second part, we used SVM as the classifier to identify three types of communication protocols. The input to this part was the 30-dimension feature vector from the 4-layer autoencoder. The output was a 1-vs-3 vector, which represents one of the three types. In our experiment, we trained the model by SMO algorithm, and applied the RBF kernel function.

3.3 Results and Analysis

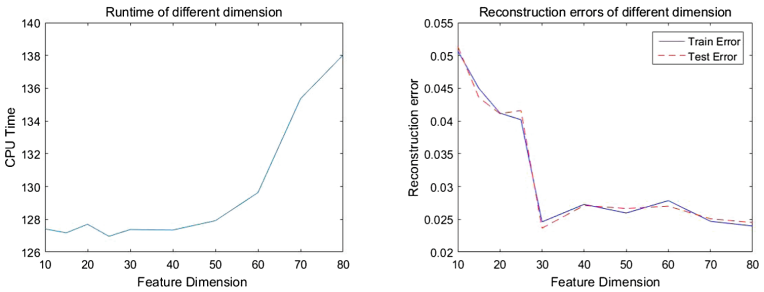
In our experiment, we used 60% of the protocol data samples to train the identification model, and 20% of the samples were used to tune the hyper-parameters. The last 20% were testing examples.

Feature Extraction. First, we trained the feature extractor by minimizing the reconstruction error, and the average error on each dimension is denoted as

$$E(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{N}\sum_{i=1}^{N}|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}| = \frac{1}{N*L}\sum_{i=1}^{N}(\sum_{j=1}^{L}|x_j^{(i)} - \hat{x}_j^{(i)}|). \tag{8}$$

Recall that $\boldsymbol{x}^{(i)}$ represents the original data of each protocol, and $\hat{\boldsymbol{x}}^{(i)}$ represents the reconstruction data of each protocol. The length of $\hat{\boldsymbol{x}}^{(i)}$ is equal to that of

$\boldsymbol{x}^{(i)}$, denoted as L. The runtime and reconstruction error of different feature dimension are shown in Fig. 3.



(a) Runtime (b) Reconstruction error

Fig. 3. The runtime and reconstruction errors of different feature dimension.

Figure 3(a) shows the runtime of different feature dimension. The number of parameters arises along with the increased feature dimension, thus leading to more runtime. From Fig. 3(b), we can see that the lower reconstruction error can be achieved, given more extracted features.

Through the training and validation, we can obtain the minimal reconstruction error when the feature dimension is 30. Meanwhile, the runtime of obtaining 30-dimension features is less than others. By using this extractor, we can automatically extract features from the original data samples, and each sample is represented by a 30-dimension feature vector, used as the input to the classifier.

Identification Results. After the above feature extraction, we used its output feature vector as the input to the classifier. In this classifier, we applied the one-against-one SVM to identify three types of protocols. Here, we make a comparison with other feature extraction methods and classification methods.

(1) In the first experiment, we compare our method with the traditional identification method, which depends on hand-engineered features with some statistical technologies. Note that Yuan et al. [17] provided 19 optional features and applied the SVM for traffic classification. So for this comparison, we selected some of the provided features to construct a four-dimension feature vector, including the data receiving time, data receiving direction, length of the flow and the byte ratio. Then, SVM was used to classify the types of protocols, using the features of the conventional method, and our proposed method in this paper respectively. The identification results are shown in Fig. 4.

From Fig. 4, we can see that the identification results of each type of protocol are 81–85% for the traditional method, while the results of our method are

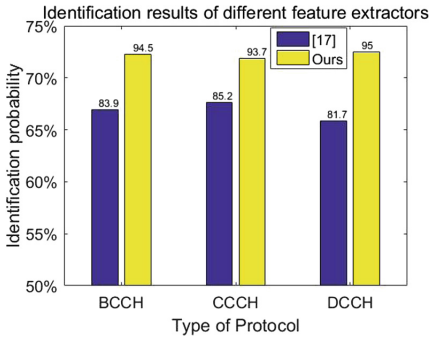


Fig. 4. Identification results of different feature extraction methods.

92%–95%. The total identification result of our method is improved about 12%, compared to the traditional method [17]. By this comparison, we can see that the autoencoder-based feature extractor is more efficient than the traditional method. At the same time, the autoencoder-based method makes the process of feature extraction more intelligently and automatically.

(2) In the second experiment, we compare our method with different classifiers. Here, we compared three different types of classifiers, including HMM, NN and SVM. For this comparison, the input to these classifiers were all extracted by using the same autoencoder-based method, and the training set, the validation set and the testing set were all the same. The HMM and NN classifiers were the same as the model mentioned in [8,10]. The results of these three classifiers are shown in Fig. 5.

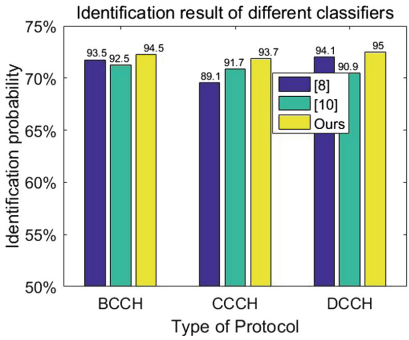


Fig. 5. Identification results of different classifiers, including HMM, NN and SVM.

From Fig. 5, we can see that nearly all of the identification results of the three classifiers are above 90%, and these results show the effectiveness of the proposed

deep learning based feature extractor. Thus, the autoencoder is competent for the task of feature extraction and may improve the identification precision. Besides, we can see that the SVM based classifier performs best for protocol identification among all three approaches. Therefore, the effectiveness of our proposed method can be validated.

4 Conclusion

The performance of protocol identification methods for wireless communication mainly depends on the features extracted from original data. The traditional feature extraction methods for protocol identification mainly rely on engineering skill and domain expertise, which consume much computational time and resource. In this paper, we introduced a deep learning method, autoencoder, to extract features from original data, meanwhile, achieving automation in the procedure of feature extraction. In addition, SVM was applied to classify three types of protocols, based on the extracted features. Some experimental results are presented in this paper, and the results show the effectiveness of the proposed method in protocol identification of WCDMA system.

References

1. Moore, A.W., Papagiannaki, K.: Toward the accurate identification of network applications. In: Dovrolis, C. (ed.) PAM 2005. LNCS, vol. 3431, pp. 41–54. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-31966-5_4
2. Kang, H.-J., Kim, M.-S., Hong, J.W.-K.: A method on multimedia service traffic monitoring and analysis. In: Brunner, M., Keller, A. (eds.) DSOM 2003. LNCS, vol. 2867, pp. 93–105. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-39671-0_9
3. Sen, S., Spatscheck, O., Wang, D.: Accurate, scalable in-network identification of P2P traffic using application signatures. In: WWW, pp. 512–521. ACM (2004)
4. Wright, C.V., Monrose, F., Masson, G.M.: On inferring application protocol behaviors in encrypted network traffic. J. Mach. Learn. Res. **6**, 2745–2769 (2006)
5. Williams, N., Zander, S., Armitage, G.J.: A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. Comput. Commun. Rev. **36**(5), 5–16 (2006)
6. Nguyen, T.T.T., Armitage, G.J.: A survey of techniques for internet traffic classification using machine learning. IEEE Commun. Surv. Tutor. **10**(1–4), 56–76 (2008)
7. Arndt, D.J., Zincir-Heywood, A.N.: A comparison of three machine learning techniques for encrypted network traffic analysis. In: CISDA, pp. 107–114. IEEE (2011)
8. Auld, T., Moore, A.W., Gull, S.F.: Bayesian neural networks for internet traffic classification. IEEE Trans. Neural Netw. **18**(1), 223–239 (2007)
9. Li, A., Dong, C., Tang, S., Wu, F., Tian, C., Tao, B., Wang, H.: Demodulation-free protocol identification in heterogeneous wireless networks. Comput. Commun. **55**, 102–111 (2015)
10. Dainotti, A., de Donato, W., Pescapè, A., Rossi, P.S.: Classification of network traffic via packet-level hidden Markov models. In: GLOBECOM, pp. 2138–2142. IEEE (2008)

11. Hu, S., Yao, Y., Yang, Z.: MAC protocol identification using support vector machines for cognitive radio networks. IEEE Wirel. Commun. **21**(1), 52–60 (2014)
12. LeCun, Y., Bengio, Y., Hinton, G.E.: Deep learning. Nature **521**(7553), 436–444 (2015)
13. O'Shea, T.J., Corgan, J., Clancy, T.C.: Unsupervised representation learning of structured radio communication signals. In: SPLINE, pp. 1–5. IEEE (2016)
14. Bengio, Y., Courville, A.C., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
15. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.: Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**, 3371–3408 (2010)
16. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2**(2), 121–167 (1998)
17. Yuan, R., Li, Z., Guan, X., Xu, L.: An SVM-based machine learning method for accurate internet traffic classification. Inf. Syst. Front. **12**(2), 149–156 (2010)



A Novel Path Selection Algorithm Based on Graph-Cut to Ensure Security in SDN

Shuquan Fan[1,2(✉)], Rongheng Lin[1,2], Shiqi Zhou[1,2], Hua Zou[1], Budan Wu[1], and Yun Su[3]

[1] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
blurain@l26.com
[2] Science and Technology on Infomation Transmission and Dissemination Networks Laboratory, The 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang 050081, China
[3] State Grid Shanghai Municipal Electric Power Company, Shanghai 200122, China

Abstract. To ensure the safety of data transmission in SDN network, a practical approach is to forward data by time division in multiple reachable paths. Considering multiple data transmission processes between the sender and the receiver based on the mechanism above, the overlap of the paths is inevitable, which may make the utilization rate of some switches or critical paths too high and other switches are relatively idle, then the transmission efficiency would be too low. This paper put forward a path selection algorithm based on graph-cut to solve the problem. Also, taking switch weight under consideration. In addition, effectiveness is taken into account. Considering to the large scale of actual networks and uneven occupancy rate of switches, this algorithm reflects its advantage.

Keywords: SDN · Path selection · Graph cut · Weight of switch
Multiplied topology extension

1 Introduction

Software Defined Networks (SDN), an evolving new-type network architecture, use OpenFlow Protocol to separate control plane from traditional network forwarding equipment, to realize more flexible and global control of networks [1]. Centralized control makes acquiring information of the whole network possible, and according to individualized requirements we can optimize global scheduling of network resources.

Most of the SDN architecture has three layers: the bottom is the network infrastructure to support the SDN function, the middle layer is SDN network controller has the control power, the upper layer including SDN configuration management applications and services. In the SDN network, attack generally includes data layer attacks, control layer attack and application layer attack. This paper mainly discusses the attack in the data layer, the attacker access to a certain or more SDN switches through some

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 546–556, 2018.
https://doi.org/10.1007/978-3-319-78130-3_56

means, and further access to the data. In this paper, we develop a security policy based on this situation.

To improve security of data transmission, we propose a forwarding rule based on time-slot, which divides unit time into intervals n, each of which is a time-slot. The SDN controller assigns a fixed path for each time-slot. A certain number of packets are transmitted in every time-slot, that is to say that at every point of time there is only one path transmitting data. When the number of packets is less than n, there will be some of time-slots in idle. For these idle time-slots, system could schedule other transmission processes to transmit or transmit some fake data to confuse attackers. When the number of packets is equal or greater than n, packets queue up and are transmitted in turn. Time-slot forwarding rule prevent attackers from easily monitoring data forwarding patterns and greatly improve transmission security.

For the time-slot forwarding scheme above, traditional path algorithm cannot calculate enough number of befitting paths to transmit data, so this paper put forward a path selection algorithm based on graph-cut to solve the problem. Also, taking switch weight under consideration, the algorithm elects optimal paths n as optional. In addition, effectiveness is taken into account.

2 Related Work

Software Defined Networks is raised by Stanford University, which is based on Open Flow and designed for constructing network innovation platform. Present goals of SDN researches can be classified into simplifying network operation, improving network performance, developing new-type network platform and providing technology of network innovation.

Path selection in data transmission is an important issue in SDN. Having acquired global topology using LLPD Protocol, routing can be calculated by source address and destination address of packets. Traditional path algorithms lead to the shortest path and classic ones of them are Floyd algorithm and Dijkstra algorithm. In addition, ONOS, a SDN controller, adopts Suurballe shortest-path graph search algorithm. Suurballe shortest-path graph search algorithm is capable of finding both the shortest path, as well as a backup shortest path, between source and destination so that the sum of the path lengths is minimized. However, factors such as how busy the switch also affect the efficiency of data transmission, so the algorithm mentioned above may not be the optimal path regardless of many factors.

Most SDN controllers communicate with the data plane using Open Flow Protocol. Information of SDN switches is uploaded to the controller by Open Flow Protocol. According to these information data, the controller works out path allocation scheme. As for path selection, literature [2] reaches a conclusion of questing optimal Open Flow controllers and has no concern with algorithms within controllers. Literature [3] proposes a heuristic method as path allocation algorithm. Literature [4] proposes a centralized control scheme based on Open Flow to make path calculation in waveband switching networks which is considered as replacement of GMPLS. Literature [5] makes a comparison among existing MPLS network path algorithms. Some researchers use optimized equivalent multi-path algorithm to realize traffic optimization in

networks, as in literature [6]. Literature [7] evaluates path selection with transmission response time and transmission rate. Literature [8] proposes a SDN architecture supporting flow monitoring, multicast management and multicast routing, which calculates the shortest path using multicast routing technique and compared with traditional shortest path it heightens link acceptance. Literature [9] optimizes Dijkstra algorithm. It configures QoS link weight, satisfies the meet of k pairs of end-to-end connection at maximum, and puts forward a heuristic algorithm with multiple constraints including delay, bandwidth, length of link and link spending to find optimal path. There are many standards to evaluate a path selection algorithm: goals of optimization can be maximizing link load [10]; or maximizing request volume of link [9]; or the most optimized path [8]; or improving transmission response time and rate [11]; or optimizing Dijkstra algorithm [9].

Aiming at time-slot forwarding scheme, this paper puts forward a path selection algorithm based on graph-cut to take the utmost advantage of idle switches to maximize utilization ratio of switches in a network.

3 Application Scenario

In SDN, data plane is separated from control plane. Control plane manages networking protocols and switch strategies, while data plane is in charge of forwarding.

From the view of a local switch, the detailed procedure of time-slot forwarding is as Fig. 1. A unit time has been divided into multiple time-slots. The controller gets topology information from TopoService. We suppose that there are three paths reaching to the destination. When data passes Switch 0, in the first time-slot, the controller distributes flow table entry to Switch 0 and configures forwarding port to transmit data to Switch 1 only. In the second time-slot, the controller distributes flow table entry to Switch 0 overwriting the previous one and alters forwarding port to transmit data to Switch 2 only. In the third time-slot, the controller behaves the similar

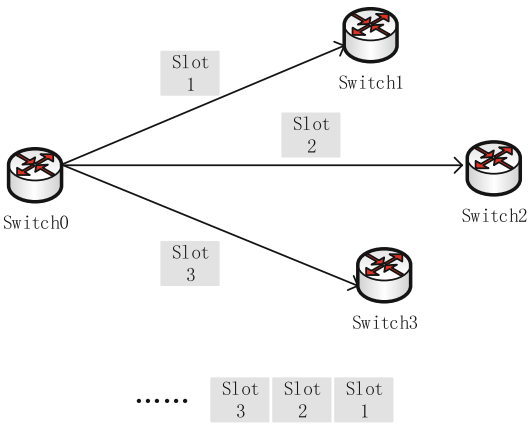


Fig. 1. Application scenario of data transmission based on time-slot scheduling

way to transmit data to Switch 3 only. In the forth time-slot, data is transmitted to Switch 1 again. And loop on until data transmission is over. In this way, even though attackers hijack one or more switches on a link, they could not get informative original data.

The key switch said sender to the destination must pass the switch. In general, the number of switches in the entire network is m, the number of switches on the alternative path is n, the number of key switches is $1(l \leq n)$, the probability of capturing all data in the case of a switch has been compromised is expressed as:

$$P_1 = \frac{l}{m} \leq \frac{n}{m} \tag{1}$$

So, the probability of capturing all data in the case of k switches have been compromised is expressed as:

$$P_k = 1 - C_k^1\left(\frac{m-n}{m}\right) \tag{2}$$

It can be seen that the time slot forwarding can reduce the probability that all data will be intercepted in the case of k switches have been compromised, that is to say, the security performance can be improved by time slot forwarding.

In time-slot forwarding strategy described above, there may be multiple reachable paths between sender and receiver. As shown in Fig. 2, there are multiple reachable paths between H1 and H2. In actual networks, we often select the first two shortest paths H1—>S5—>S7—>S10—>H2 and H1—>S5—>S8—>S10—>H2 to transmit data. Then if H3 and H4 need to transmit data, selecting an appropriate path can improve efficiency of forwarding. We should choose relatively idle switches as optional. Because switches on transmission path of H1 and H2 are working, it would be best to avoid using them. So we propose a path selection algorithm to select ideal paths and to improve efficiency of forwarding.

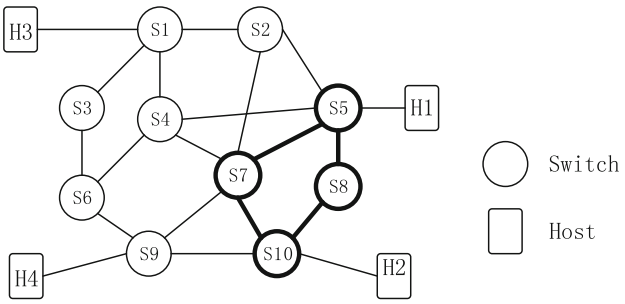


Fig. 2. Schematic diagram of topology

4 Algorithm Description

The algorithm can be described as 3 parts. First, path Selection Based on Graph Cut, second, add weight flag of Switch, last, the way of multiplied topology extension.

4.1 Path Selection Based on Graph Cut

As shown in Fig. 2, in the entire network topology, S5, S7, S8, S10 can be seen as busy switches, when the data needs to be transferred between H3 and H4. It should be considered to use remaining switches to forward data. Equivalently, delete busy switches and its associated links and then select the appropriate forwarding path in the remaining topology.

As shown in Fig. 3, a part of the original topology is used as initial candidate paths, where the data transmission path between H3 and H4 is selected from. In this case, we use Dijkstra algorithm to get shortest paths between H3 and H4. There are two shortest paths: H3->S1->S3->S6->S9->H4 and H3->S1->S4->S6->S9->H4. If there are only two time-slots in the time-slot forwarding procedure, the paths selected can meet the requirements. If there are three time-slots, we need to select another path in the network topology.

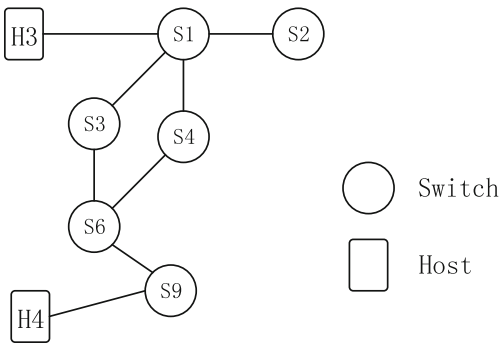


Fig. 3. Initial candidate path

4.2 Weight of Switch

First of all, for the problem mentioned above, there is no third path between H3 and H4. Then, to select a third path, you have to restore the busy switches that was deleted before, but if you restore them all, the previous graph-cut process will make no sense and the path selection will not be optimized. Here we assign weight to the switches. Weight of each switch is set to 1 initially and is increased by 1 when the switch is used by a new link. And then we restore the switches in accordance with the weight, so the relatively idle switches are chosen to forward data. As shown in Fig. 4, S5, S7, S8, S10 are switches supposed to be restored. Because of data being transmitting between H1 and H2, the weight of S5 and S10 is increased by 2, that of S7 and S8 are increased by

1, so S7 and S8 will be restored to the topology. Obviously, after S7 added to the topology, the third path can be selected.

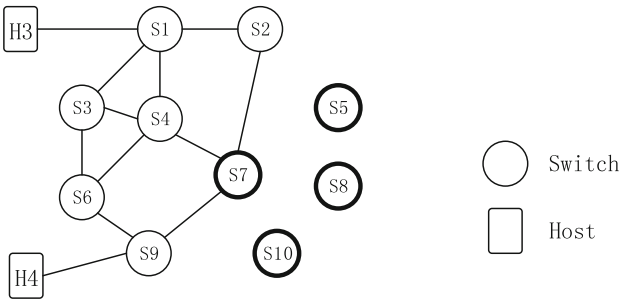


Fig. 4. Weight of Switch

4.3 Multiplied Topology Extension

The number of switches in an actual network is huge. In fact, if you add a switch to the topology, and then run the shortest path algorithm every time, the whole process is very inefficient. In order to improve efficiency and achieve the goal, here we restore twice the number of switches according to weight, so that we can quickly find the path meeting the requirements. This may make the selected path is not the optimal one though, efficiency is greatly improved. As shown in Fig. 5, in the first expansion, a switch with weight of 2 is added; in the second expansion, switches with weight of 2 and 3 is added, and so on until all switches are added to the topology. An extreme case is when all the switches are added to the topology, the number of paths is still not enough, it will return the found path.

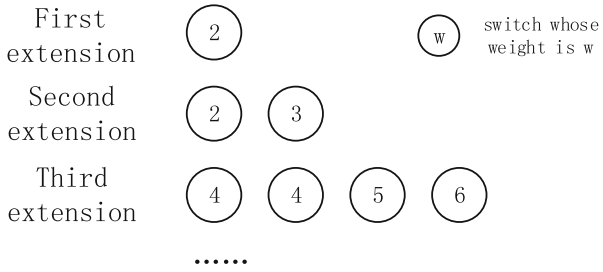


Fig. 5. Multiplied topology extension

4.4 Algorithmic Flow

The entire process can be represented by a flow chart, as shown in Fig. 6:

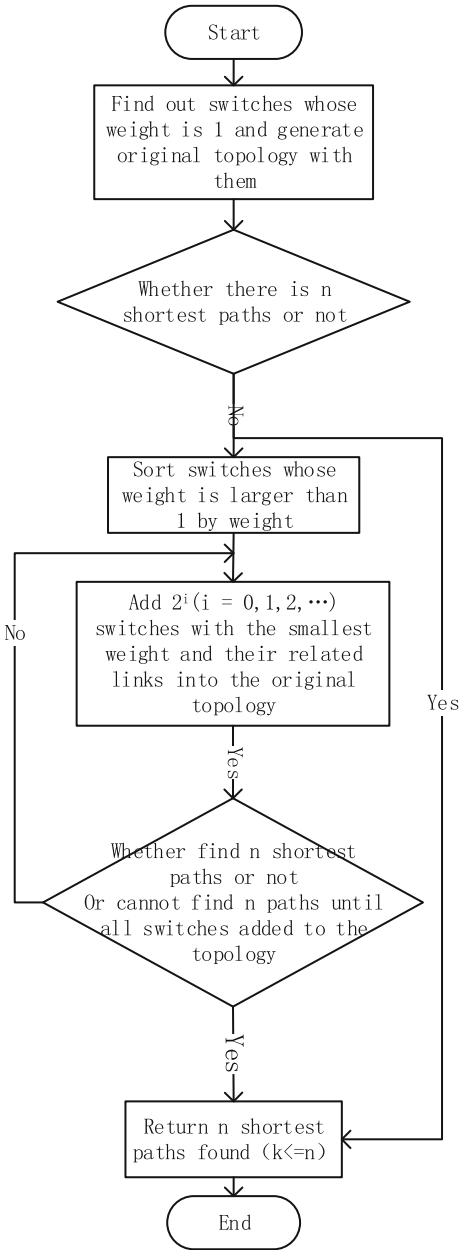


Fig. 6. Algorithm flowchart

5 Experiment

5.1 Algorithm to Achieve

The above algorithm can be represented by the following pseudocode:

Input: *The original topology(topology) , Starting point(src), End point(dst), Number of paths to be selected(n)*
Output: *The selected path(*paths)

```
Topology subTopology = getsubTopology(topology)
Topology compTopology =getcompTopology(topology, ubTopology)
Set<Path> paths = getPaths(subTopology, src, dst)
If paths.size()>=n Then
    return paths
Else
    While isNotEmpty(compTopology) Then
        updateTopology(subTopology)
        paths.addAll(getPaths(subTopology, src, dst))
        If paths.size()>=n Then
            return paths
        End If
    End
    return paths
End If
```

Description of algorithm function:

getsubTopology(): Get the switches whose weight is 1 in the original topology and the initial topology "subTopology";

getcompTopology(): Get the switches whose weight is greater than 1 and the complementary graph of the initial topology "compTopology";

getPaths(): Get the shortest path from the source to the destination in a topology;

updateTopology(subTopology): Update subTopology and add 2^i (i = 0, 1, 2, …) switches with the smallest weight and their related links into subTopolog

paths.addAll() Add the newly acquired shortest path to the collection.

5.2 Algorithm Evaluation

In the study of data transmission scheduling in SDN network, a network environment is needed. Mininet can be used to simulate the main network environment. Mininet provides a range of tools to test the connectivity between network hosts and the rate of network links and so on. Therefore, the experimental network was created using Mininet. The Open Network Operating System [18] (ONOS) is a software defined networking (SDN) OS for service providers that has scalability, high availability, high performance and abstractions to make it easy to create apps and services. The platform is based on a solid architecture and has quickly matured to be feature rich and production ready.

First of all, we use the algorithm above to select the transmission path for data transmission, then control the transmission path by ONOS. Send 10,000 packets and count the time required and compare it to the ONOS default routing algorithm. Increase the size of the network topology by increasing the number of switches from 10 to 100, respectively, record the time used for the completion of the transfer task, in seconds (Fig. 7).

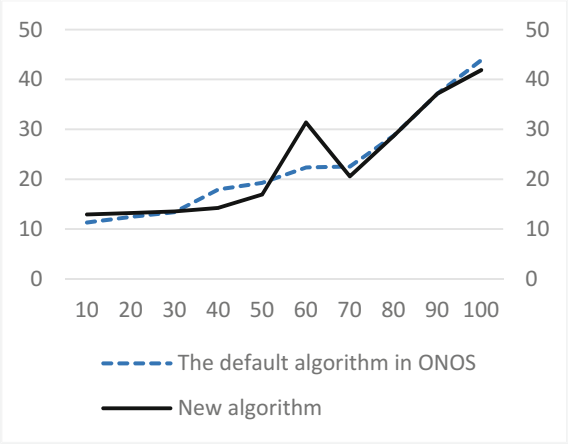


Fig. 7. Efficiency comparison

The path selected by this algorithm is very similar to the one selected by the ONOS when the network is idle, but there is a big difference when the network is busy. In fact, this phenomenon can be understood, simply increase the size of the network and most of the switches are in the idle state, the weight of idle switch is 1, a large number of idle switches lead to the algorithm is similar to find the shortest path.

Then, in order to reflect the characteristics of the algorithm, we do another experiment, Increase the number of transmission tasks, from one task to 10 tasks, thus increasing the degree of network congestion. Statistics packet transmission time, and compare to the ONOS default situation.

It can be concluded that the algorithm can find relatively idle switches for data transmission. There may be some cases where the number of hops increases, but it still can achieve good results. When the busy degree of the switch in the network increases, this algorithm takes precedence over the idle switch, which is different from the ONOS default shortest path algorithm. It can be seen from Fig. 8 that the time taken to transmit data using our algorithm is reduced, and this difference is bound to increase as the amount of data in the network increases.

In addition, the characteristics of the algorithm can be explained from another point of view. If part of switches are extremely busy, while others are in the idle state, then the path is not good, the share of the switch is high, and the utilization of the network is low. On the contrary, if the busy degree of all the switches in the network is not a big

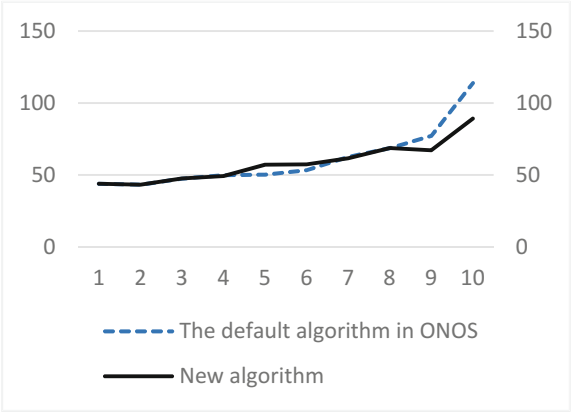


Fig. 8. Efficiency comparison

difference, then the network resources is better used. Using the path selection algorithm mentioned in this paper, due to the use of the graph-cut algorithm and the switch weight, the weight of idle switches is low, and idle switches will be firstly used, resulting in the busy degree of all the switches in the network is not a big difference, and low share of the switch.

6 Conclusion

In SDN, path selection has great significance in data transmission. Time-slot forwarding strategy adopted can enhance security of data transmission, but it requests multiple reliable paths to transmit data. For this problem, we introduce a path selection algorithm capable of selecting multiple befitting paths based on graph-cut. Considering to the large scale of actual networks and uneven occupancy rate of switches, this algorithm reflects its advantage.

Acknowledgments. This work is supported by Beijing Natural Science Foundation (4174099) and National High Technology Research and Development Program (863 Program) of China (2015AA050203).

References

1. Software-defined-networking. https://en.wikipedia.org/wiki/Software-defined_networking
2. Kanagavelu, R., Mingjie, L.N., Mi, K.M., Lee, B.-S., Francis, Heryandi: OpenFlow based control for re-routing with differentiated flows in data center networks. In: International Conference on Networks (ICON), pp. 228–233. IEEE (2012)
3. Kai, W.: The load balancing research of SDN based on ant colony algorithm. Jilin University (2015)

4. Giorgetti, A., Cugini, F., Paolucci, F., Castoldi, P.: OpenFlow and PCE architectures in wavelength switched optical networks. In: 2012 16th International Conference on Optical Network Design and Modeling (ONDM), pp. 1–6. IEEE (2012)
5. Gargi, B., Deepinder, S.: Comparative analysis of path computation techniques for MPLS traffic engineering. Comput. Netw. **40**(1), 149–165 (2002)
6. Li, Y., Pan, D.: OpenFlow based load balancing for Fat-Tree networks with multipath support. In: Proceedings of 12th IEEE International Conference on Communications (ICC 2013), Budapest, Hungary, pp. 1–5 (2013)
7. Kaur, S., Singh, J., Kumar, K., et al.: Round-robin based load balancing in Software Defined Networking. In: Proceedings of 2nd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 2136–2139 (2015)
8. Craig, A., Nandy, B., Lambadaris, I., et al.: Load balancing for multicast traffic in SDN using real-time link cost modification. In: Proceedings of International Conference on Communications (ICC), pp. 5789–5795. IEEE (2015)
9. Abe, J.O., Mantar, H.A., Yayimli, A.G.: k-maximally disjoint path routing algorithms for SDN. In: International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, pp. 499–508. IEEE Computer Society (2015)
10. Guo, Y., Wang, Z., Yin, X., et al.: Traffic engineering in SDN/OSPF hybrid network. In: 2014 IEEE 22nd International Conference on Network Protocols (ICNP), pp. 563–568. IEEE (2014)
11. Luo, C.: Research of the optimal path allocation algorithm oriented load balancing in SDN. Hubei University Of Technology (2016)
12. Li, K., Wang, S., Xu, S., et al.: ERMAO: an enhanced intradomain traffic engineering approach in LISP-capable networks. In: 2011 IEEE Global Telecommunications Conference (GLOBECOM 2011), pp. 1–5. IEEE (2011)
13. Bazgan, C., Toubaline, S., Vanderpooten, D.: Efficient algorithms for finding the *k* most vital edges for the minimum spanning tree problem. In: Wang, W., Zhu, X., Du, D.-Z. (eds.) COCOA 2011. LNCS, vol. 6831, pp. 126–140. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22616-8_11
14. Nardelli, E., Proietti, G., Widmayer, P.: A faster computation of the most vital edge of a shortest path. Inf. Process. Lett. **79**(2), 81–85 (2001)
15. Guo, C., Lu, G., Li, D., et al.: BCube: a high performance, server-centric network architecture for modular data centers. SIGCOMM **39**(4), 63–74 (2009)
16. Farrington, N., Porter, G., Radhakrishnan, S., et al.: Helios: a hybrid electrical/optical switch architecture for modular data centers. ACM SIGCOMM Comput. Commun. Rev. **40**(4), 339–350 (2010)
17. Al-Fares, M., Loukissas, A., Vahdat, A.: A scalable, commodity data center network architecture. ACM SIGCOMM Comput. Commun. Rev. **38**(4), 63–74 (2008)
18. The Open Network Operating System. http://onosproject.org



Secure and Reliable Transmission for Dense HetNets Based on Secure Energy Efficiency Optimization

Yunjia Xu(✉), Kaizhi Huang, and Yi Zou

National Digital Switching System Engineering and Technological Research Center, Zhengzhou 450002, China
136564061@qq.com

Abstract. Due to dense and random deployment of dense heterogeneous cellular networks (HetNets), the systems are more susceptible to channel estimation errors, which effect secure and reliable transmission. To solve this problem, we propose a secure and reliable transmission scheme for dense HetNets based on secure Energy Efficiency (EE) Optimization. Firstly, the system model with channel estimation error is established. Then, transmission threshold selection algorithm is designed to maximize secure EE, and emission control is achieved by comparing the receiving signal-to-interference-plus-noise ratio (SINR) with the transmission threshold. Finally, secure EE constrained by the connection and secrecy probabilities is derived. Numerical results verify that the transmission scheme can increase secure EE by at least 50%.

Keywords: Channel estimation errors · Physical layer security
Dense heterogeneous networks · Security energy efficiency

1 Introduction

As an important content of 5G communication technology, dense heterogeneous cellular networks (HetNets) [1] enhance the system capacity, but also left huge security risks to the system, and information is more vulnerable to be eavesdropped. Traditional encryption algorithm is faced with difficulties of key distribution and computational complexity when it is used in dense HetNets, which seriously limits the performance of the system. As an important complement to traditional security means, physical layer security technology uses the basic attributes of wireless channels to ensure secure communications. The security performance of dense HetNets is analyzed in [2,3]. On the basis, researchers have designed many methods to improve the security performance of the system by using large-scale MIMO [4], cooperative transmission [5], cooperative interference [6], resource allocation [7] and so on.

However, the above studies [2–7] are based on the ideal channel hypothesis. In practice, the structure of dense HetNets determines the system is vulnerable to

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
B. Li et al. (Eds.): ChinaCom 2017, LNICST 236, pp. 557–566, 2018.
https://doi.org/10.1007/978-3-319-78130-3_57

be influenced by channel estimation errors. To ensure secure communication with imperfect CSI, secure on-off scheme [9], robust beamforming [10,11] and artificial noise assisted transmission [12] and cooperative transmission [13,14] have been considered in traditional network. As dense HetNets nodes are randomly deployed, the traditional transmission scheme is no longer applicable. To analyze the effect of multilink channel estimation errors on the system, it is necessary to design a secure and reliable transmission scheme taking the characteristics of dense HetNets into consideration.

In view of the above problems, this paper propose a secure transmission scheme for dense HetNets based on secure Energy Efficiency (EE) Optimization. Firstly, the system model with channel estimation errors is established. Then, the transmission threshold selection algorithm is designed to maximize secure EE of the system, and the emission control is achieved by comparing the received signal-to-interference-plus-noise ratio (SINR) with the threshold. Finally, the influence of channel estimation accuracy and transmission threshold on secure EE of the system is analyzed. Numerical results verify that the secure transmission scheme can increase the system secure EE by at least 50%.

2 System Model

In K-tier dense HetNets, BSs in different layers are distinguished by transmission power, path loss coefficient, antennas number and spatial distribution density. We model positions of BSs, legitimate users and eavesdroppers in the ith tier as homogeneous Poisson Pointed Processes (PPPs) Φ_i, Φ_u and Φ_e, with intensity λ_i, λ_u and λ_e. Taking the m th BS in tier i as an example, i.e. B_{im}, the total transmission power and antennas number of B_{im} are $P_{i,total}$ and N_i. We consider large scale and quasi-static Rayleigh fading, and the large scale fading coefficient is α_i, $\alpha_i > 2$.

In K-tier open access networks, users can choose arbitrary BSs to access. Users choose the BS providing maximum receiving power to associate, and the association probability can be expressed as [15]:

$$\mathcal{A}_i = 2\pi\lambda_i \int_0^\infty x \exp\{-\pi \sum_{j=1}^{K} \lambda_j (\frac{P_j}{P_i})^{2/\alpha_j} x^2\} dx \tag{1}$$

The probability density function (PDF) of the distance from the user to the serving base station can be expressed as [15]:

$$f_{X_i}(x) = \frac{2\pi\lambda_i}{\mathcal{A}_i} x \exp\left\{-\pi \sum_{j=1}^{K} \lambda_j (\frac{P_j}{P_i})^{2/\alpha_j} x^2\right\} \tag{2}$$

In the channel estimation phase, BSs transmit the pilot signals at each time slot. Users receive the pilot signals to estimate the current channel, and feed the estimated value back to BSs through the ideal backhaul link. Then BSs

design the precoding matrix $\tilde{\mathbf{w}}_i = \tilde{\mathbf{h}}_{ui}^H \Big/ \left\| \tilde{\mathbf{h}}_{ui} \right\|$ according to the current CSI. The random deployment and overlapping coverage ranges of BSs causes channel estimation errors easily. We consider $\mathbf{h}_{ui}$ as the legitimate communication link, and $\tilde{\mathbf{h}}_{ui}$ is the estimated value of $\mathbf{h}_{ui}$, and the relationship between $\tilde{\mathbf{h}}_{ui}$ and $\mathbf{h}_{ui}$ can be expressed as:

$$\mathbf{h}_{ui} = \tilde{\mathbf{h}}_{ui} + \mathbf{n}_i \tag{3}$$

where $\mathbf{n}_i \sim \mathcal{CN}(0,\ 1-\rho_i)$, $1-\rho_i$ is variance of the channel estimation errors. We can consider $\rho_i\ (0 \le \rho_i \le 1)$ as channel estimation accuracy, and a high ρ_i indicates accurate channel estimation.

According to Slivnyak theorem, we consider a typical user accessing to base station B_{im}, the estimated SINR can be expressed as:

$$\tilde{\gamma}_U^i = \frac{P_i \left\| \tilde{\mathbf{h}}_{ui} \tilde{\mathbf{w}}_i \right\|^2 |X|^{-\alpha}}{I_{ui} + {\sigma_{ui}}^2} \tag{4}$$

Assuming non-colluding eavesdroppers can estimate the channel accurately. Considering the worst case, we study the eavesdropping link with the largest SINR, which can be expressed as:

$$\gamma_E^i = \max_{e \in \Phi_e} \left\{ \frac{P_i \|\mathbf{h}_{ei} \tilde{\mathbf{w}}_i\|^2 |Y|^{-\alpha}}{I_{ei} + {\sigma_{ei}}^2} \right\} \tag{5}$$

where $i \in [1, K]$, $\|\mathbf{h}_{ei}\tilde{\mathbf{w}}_i\|^2 \sim gamma(N_i, \rho_i)$, $\|\mathbf{h}_{ei}\tilde{\mathbf{w}}_i\|^2 \sim \exp(1)$. $I_{ui} = \sum_{j=1}^{K} \sum_{q \in \Phi_j \setminus B_{im}} P_j \|\mathbf{h}_{uj}\|^2 |Z|^{-\alpha}$ and $I_{ei} = \sum_{j=1}^{K} \sum_{q \in \Phi_j \setminus B_{im}} P_j \|\mathbf{h}_{ej}\|^2 |M|^{-\alpha_j}$ denotes interference received by the legitimate user and eavesdropper respectively, where $\mathbf{h}_{uj}$ and $\mathbf{h}_{ej}$ indicate the channel from interference BS to the legitimate user and eavesdropper, $\|\mathbf{h}_{uj}\|^2 \sim \exp(1)$, $\|\mathbf{h}_{ej}\|^2 \sim \exp(1)$. X, Y and Z, M stand for the distance from the service BS and interference BSs to the intended user and eavesdropper respectively and are power of gauss white noise received by the user and eavesdropper. σ_{ui}^2 and σ_e^2 are power of gauss white noise received by the user and eavesdropper.

Due to the impact of the channel estimation errors, the user's real received SINR can be expressed as:

$$\gamma_U^i = \frac{P_i \left\| \tilde{\mathbf{h}}_{ui} \tilde{\mathbf{w}}_i \right\|^2 |X|^{-\alpha}}{P_i \|\mathbf{n}_i \tilde{\mathbf{w}}_i\|^2 |X|^{-\alpha} + I_{ui} + {\sigma_{ui}}^2} \tag{6}$$

where $P_i \|\mathbf{n}_i \tilde{\mathbf{w}}_i\|^2 |X|^{-\alpha}$ denotes the interference caused by the channel estimation errors, $\|\mathbf{n}_i \tilde{\mathbf{w}}_i\|^2 \sim \exp(1-\rho_i)$. We define $\tilde{\gamma}_N = P_i \|\mathbf{n}_i \tilde{\mathbf{w}}_i\|^2 \left(I_{ui} + {\sigma_{ui}}^2\right)^{-1} x^{-\alpha}$, and the relationship between estimated SINR and actual SINR can be expressed as:

$$\gamma_U^i = \frac{\tilde{\gamma}_U^i}{\tilde{\gamma}_N + 1} \tag{7}$$

3 Secure and Reliable Transmission Design

In this section, we design a secure and reliable transmission scheme with good secure EE performance, whilst satisfying certain constraints on the reliability and security constraints. In particular, BSs decides whether to transmit signals according to instantaneous SINR estimated by the legitimate user. The transmission threshold selection algorithm is designed to maximize secure EE, and emission control is implemented by comparing the received SINR with the transmission threshold.

The secure EE is defined as average secrecy rate with unit energy consumption. Mathematically, the secure EE of a typical link in tier i is given by:

$$\xi_i = \frac{\lambda_i p_{tx}^i \left(1 - p_{co}^i\right) \left(1 - p_{so}^i\right) R_s^i}{P_{i,total}} \tag{8}$$

BSs in different layers have the same security constraint and reliability constraint since users have the same communication requirements. The optimization transmission threshold selection algorithm can be expressed as:

$$\begin{aligned} \max_{\mu_b^1, \cdots \mu_b^K} \quad & \xi = \sum_{i=1}^{K} \mathcal{A}_i \cdot \frac{\lambda_i p_{tx}^i \left(1-p_{co}^i\right)\left(1-p_{so}^i\right) R_s^i}{P_{i,total}}, \\ s.t. \quad & p_{so}^i \leq \varepsilon, p_{co}^i \leq \delta, i = 1, 2 \cdots K \end{aligned} \tag{9}$$

where $\varepsilon \in [0, 1]$, $\delta \in [0, 1]$. p_{tx}^i, p_{co}^i and p_{so}^i are transmission probability (TP), connection outage probability (COP) and security outage probability (SOP) in tier i.

Since ξ_i is only related to μ_b^i, we can convert the multi-parameters $\mu_b^1 \cdots \mu_b^K$ optimization problem to the optimal single parameter μ_b^i optimization problem for each ξ_i. The secure EE optimization problem of tier i is expressed as:

$$\begin{aligned} \max_{\mu_b^i} \quad & \xi_i = \frac{\lambda_i p_{tx}^i \left(1-p_{co}^i\right)\left(1-p_{so}^i\right) R_s^i}{P_{i,total}}, \\ s.t. \quad & p_{so}^i \leq \varepsilon, p_{co}^i \leq \delta \end{aligned} \tag{10}$$

The maximum secure EE of the system can be expressed as:

$$\xi = \sum_{i=1}^{K} \mathcal{A}_i \cdot \xi_i \tag{11}$$

To solve the above optimization problem, we have:

$$\begin{aligned} &\frac{d\xi\left(\mu_b^i\right)}{d\mu_b^i} = 0 \\ &\frac{d^2\xi\left(\mu_b^i\right)}{d^2\mu_b^i} < 0 \end{aligned} \tag{12}$$

The unique solution of (12) is $\mu_b^i = \Xi\left(R_b^i\right)$. To solve the constraints, $p_{co}^i \leq \delta$, $p_{so}^i \leq \varepsilon$, we have:

$$\mu_b^i \geq \nu_b^i \tag{13}$$

where $\nu_b^i = \max\left\{F^i\left(\delta\right), G^i\left(\varepsilon\right)\right\} F^i\left(\delta\right)$ and $G^i\left(\delta\right)$ are solutions of equations $p_{co}^i = \delta$ and $p_{so}^i = \varepsilon$, respectively.

Proposition 1. *The optimal transmission threshold μ_b^i of the secure EE maximizing is given as follows:*

$$\mu_b^i = \begin{cases} \Xi\left(R_b^i\right), & \nu_b^i \le \Xi\left(R_b^i\right) \\ \nu_b^i, & \nu_b^i > \Xi\left(R_b^i\right) \end{cases} \tag{14}$$

The above threshold optimization algorithm can be summarized as shown in Table 1:

Table 1. Threshold optimization algorithm

Step1	Calculate p_{tx}^i, p_{co}^i, p_{so}^i and ξ_i respectively
Step2	Solve the transmission threshold $\mu_b^i = \Xi\left(R_b^i\right)$ according to Eq. (12)
Step3	Solve ν_b^i according to constraints, $p_{co}^i \le \delta$, $p_{so}^i \le \varepsilon$
Step4	Substituting ν_b^i into Eq. (14) to find the optimization threshold

Plainly, ε and δ have impacts on secure EE, and bigger constraints make it easier to achieve maximum secure EE. It can be understood that the system does not need to sacrifice secure EE to satisfy the loose security and reliability constraints. But when the constraints become strict, the system may sacrifice its secure EE to meet the needs of secure and reliable transmission.

BSs implement secure emission control by comparing the current SINR with transmission threshold. BSs transmit signals when the received SINR of the legitimate user is greater than transmission threshold. Otherwise, BSs suspend transmission as the current channel quality is not suitable for secure and reliable transmission.

To avoid secure EE decreasing with low encoding rate and transmission interruption with high encoding rate in transmission phase, we set the encoding rate in each layer as the lower bound of average achievable rate of the legitimate user, which can be expressed as [5]:

$$R_U^{L_i} = \log_2\left(1 + \mathbb{E}\left[\left(SINR_U^{i\ -1}\right)\right]^{-1}\right) \tag{15}$$

Proposition 2. *The encoding rate R_b^i is given as follows:*

$$R_b^i = \log_2\left(1 + N_i\rho_i P_i\left[\begin{array}{l}(1-\rho_i)\,P_i + \\ \int_0^\infty \left(\sum_{j=1}^{K}\left(\frac{2\pi\lambda_j P_j (M_j)^{2-\alpha_j}}{\alpha_j - 2}\right) + \sigma^2\right) x^{\alpha_i} f_{X_i}(x)dx\end{array}\right]^{-1}\right) \tag{16}$$

where $M_j = \left(P_j/P_i\right)^{1/\alpha_j} X_i$ is the closest distance of interferer BSs in tier j.

Proof. We derive $\mathbb{E}\left(SINR_U^{i\ -1}\right)$ first according to (15), $\mathbb{E}\left(SINR_U^{i\ -1}\right) = \frac{\int_0^\infty \left((I_{ui})+{\sigma_u}^2\right)x^{\alpha_i} f_{X_i}(x)dx}{\left(P_i\|\mathbf{n_i}\tilde{\mathbf{w}}_\mathbf{i}\|^2\right)+\left(P_i\left\|\tilde{\mathbf{h}}_{ui}\right\|^2\right)}$, where $\left(P_i\|\mathbf{n_i}\tilde{\mathbf{w}}_\mathbf{i}\|^2\right) = (1-\rho_i)\,P_i$, where $\mathbb{E}\left(P_i\|\mathbf{n_i}\tilde{\mathbf{w}}_\mathbf{i}\|^2\right)$ $= (1-\rho_i)\,P_i$, $\mathbb{E}\left(P_i\left\|\tilde{\mathbf{h}}_{ui}\right\|^2\right) = N_i\rho_i P_i$, $\mathbb{E}\left(I_{ui}\right) = \sum_{j=1}^{K} 2\pi\lambda_j P_j (M_j)^{2-\alpha}(\alpha-2)^{-1}$. Substituting them into (15), R_b^i is obtained.

In summary, the secure and reliable transmission scheme of dense HetNets with imperfect CSI can be summarized as shown in Table 2.

Table 2. Secure and reliable transmission scheme

Step1	Estimate the channel, and access the BS
Step2	Calculate the coding rate R_b^i according to the Eq. (16)
Step3	Calculate the μ_b^i according to the optimization algorithm shown in Table 1
Step4	Compare the current SINR with the preset threshold. If $SINR^i \geq \mu_b^i$, transmit signals. Otherwise, turn back to Step2

Since channel estimation errors may occur in different layers of dense HetNets, secure and reliable transmission scheme can improve the reliability and security performance of the system with channel estimation errors by ameliorating secure EE performance of each layer.

4 Performance Analysis

In this section, we analyze the effects of channel estimation and transmission threshold on secure EE. We analyze the performance of secure EE by deriving TP, COP and SOP of the typical links in the considered K-tier dense HetNets.

TP quantifies the probability that the message can be transmitted at the service BS. Mathematically, TP of a typical link in tier i is given by:

$$\begin{aligned} p_{tx}^i &= \Pr\left(\tilde{\gamma}_U^i > \mu_b^i\right) \\ &= \int_0^{D_i} \sum_{k=0}^{N-1} \frac{\left[\left(\mathbb{E}(I_{ui})+{\sigma_{ui}}^2\right)\mu_b^i x^\alpha\right]^k}{(\rho_i P_i)^k k!} \mathcal{L}_{I_i}\left(\frac{\mu_b^i}{\rho_i P_i}x^\alpha\right) \exp\left(-\frac{\mu_b^i {\sigma_{ui}}^2}{\rho_i P_i}x^\alpha\right) f(x)\,dx \end{aligned} \tag{17}$$

where

$$\begin{aligned} \mathcal{L}_{I_i}(s) &= \mathbb{E}_{\Phi_j, g_{uj}}\left[\exp\left(-s\sum_{j=1}^{K}\sum_{i\in\Phi_j\setminus B_{ko}} P_j g_{uj}|Z_j|^{-\alpha}\right)\right] \\ &= \prod_{j=1}^{K} \exp\left(-2\pi\lambda_j \int_{M_j}^{\infty} \frac{z}{1+s^{-1}z^{\alpha_j}}dz\right) \end{aligned} \tag{18}$$

Taking $\frac{\mu_b^i}{P_i}x^\alpha$ into $\mathcal{L}_{I_i}(s)$, $\mathcal{L}_{I_i}\left(\frac{\mu_b^i}{\rho_i P_i}x^\alpha\right) = \prod_{j=1}^{K} \exp\left\{-\pi\lambda_j (P_j/\rho_i P_i)^{\frac{2}{\alpha}}\left(\mathbb{Z}\mu_b^i, \alpha 1\right)x^2\right\}$, where $\mathbb{Z}\left(\mu_b^i, \alpha 1\right) = \frac{2\mu_b^i}{\alpha-2}{}_2F_1\left[1, 1-\frac{2}{\alpha}; 2-\frac{2}{\alpha}; -\mu_b^i\right]$, and ${}_2F_1\left[1, 1-\frac{2}{\alpha}; 2-\frac{2}{\alpha}; -\mu_b^i\right]$ is the Gauss hypergeometric function.

COP measures reliability of communication link, which can be expressed as:

$$\begin{aligned} p_{co}^{i} &= \Pr\left(C_{ui}^{L} < R_b^i | \tilde{\gamma}_b^i > \mu_b^i\right) \\ &= 1 - \left(p_{tx}^{i}\right)^{-1} \cdot \int_0^{\infty} \sum_{k=0}^{N_i-1} \frac{\Theta^k \exp(-\Theta/\rho_i)}{k! {\rho_i}^k} f(x)\, dx \end{aligned} \tag{19}$$

where

$$\begin{aligned} \Theta &= \max\left(\left(\mathbb{E}(I)+\sigma^2\right) {P_i}^{-1} \mu_b^i/\rho_i,\ \left(\mathbb{E}(I)+\sigma^2\right) {P_i}^{-1}\left(2^{R_b^i}-1\right)\Big/\rho_i\right) x^{\alpha} \\ &= \begin{cases} \left(\mathbb{E}(I)+\sigma^2\right)\mu_b^i (P_i\rho_i)^{-1} x^{\alpha}, & \mu_b^i \geq 2^{R_b^i}-1 \\ \left(\mathbb{E}(I)+\sigma^2\right)\left(2^{R_b^i}-1\right)(P_i\rho_i)^{-1} x^{\alpha}, & \mu_b^i < 2^{R_b^i}-1 \end{cases} \end{aligned} \tag{20}$$

SOP measures security of communication link, which can be expressed as:

$$\begin{aligned} p_{so}^{i} &= \Pr\left(C_e^i > R_b^i - R_s^i | \tilde{\gamma}_b^i > \mu_b^i\right) \\ &= 1 - \exp\left(-2\pi\lambda_e \int_0^{\infty} K\left(2^{R_b^i - R_s^i}-1, y\right) y dy\right) \end{aligned} \tag{21}$$

where

$$K(\gamma, y) = \prod_{j=1}^{K} \exp\left(-2\pi\lambda_j \int_0^{\infty}\left(1 - \frac{1}{\gamma P_j y^{\alpha_i} |z|^{-\alpha_j}/P_i + 1}\right) z dz\right) \exp\left(-\frac{\sigma^2\gamma}{P_i y^{-\alpha_i}}\right) \tag{22}$$

Combine (17), (19) and (21), we can have ξ_i. Note that μ_b^i and ρ_i have effects on COP and SOP of tier i. Since $\frac{\partial\left(p_{co}^i\right)}{\partial \mu_b^i} < 0$, COP of tier i monotonically decreases with μ_b^i. This signifies that reliable communication can be guaranteed by a high transmission threshold. In addition, $\frac{d\left(p_{co}^i\right)}{d\rho_i} < 0$, $\frac{d\left(p_{so}^i\right)}{d\rho_i} < 0$. COP and SOP of tier i monotonically decreases with ρ_i. This implies that accurate channel estimation can decrease COP and SOP, which is an important prerequisite for reliable and secure communication.

In addition, μ_b^i and ρ_i have effects on TP. Since $\frac{\partial\left(p_{tx}^i\right)}{\partial \mu_b^i} < 0$, $\frac{d\left(p_{tx}^i\right)}{d\rho_i} > 0$, TP of tier i monotonically decreases with μ_b^i, and increases with ρ_i. This implies that low transmission threshold and accurate channel estimation can increase the probability of transmission.

Since the influence on TP, COP and SOP caused by μ_b^i and ρ_i are different, it is obvious that secure EE is not a monotonic function of μ_b^i, but an increase function of ρ_i according to (8). We analyze the influence on secure EE caused by μ_b^i and ρ_i in the numerical results.

5 Numerical Results

In this section, we illustrate and analyze the numerical results for the secure and reliable transmission design, considering a 2-tier dense HetNets as an example, i.e. the macrocell base stations (MBSs) and the picrocell base stations (PBSs).

We set the confidential information rate as $R_s = 3\,\text{bit/s/Hz}$, and the main simulation parameters are $P_{1,total} = 20\text{W}$, $P_{2,total} = 2\text{W}$, $\lambda_1 = 0.01$, $\lambda_e = 0.05$, $N_1 = 10$, $N_2 = 3$, $alpha = 3$.

The effects of channel estimation accuracy on the reliability, security performance and secure EE are examined in Figs. 1 and 2 respectively. Since the effects are similar on each tier, we present TP, COP and SOP in tier 1 over different channel estimation accuracy in Fig. 1, i.e. p_{co}^1 and p_{so}^1. As can be observed, the analytical results are in quite good agreement with simulation results. This fact validates the accuracy of our analytical results. It is obvious that p_{co}^1 and p_{so}^1 are decreasing functions of ρ_1. This implies that the channel estimation is very important in reliable and secure communication. Otherwise, p_{tx}^1 is an increasing function of ρ_1, and this signifies that the transmission scheme can ensure reliability, security performance of the system by transmission control.

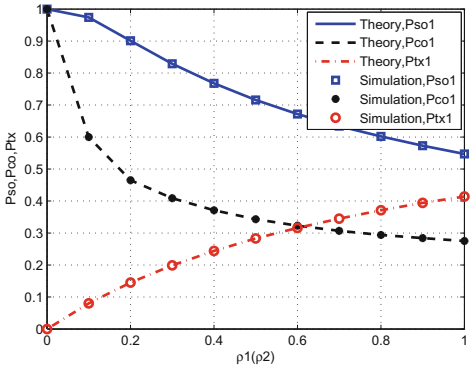


Fig. 1. TP, COP and SOP versus channel estimation accuracy.

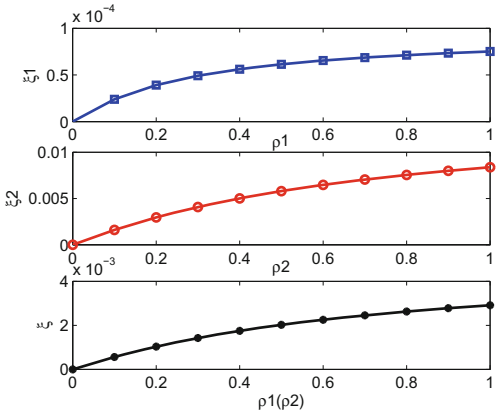


Fig. 2. TP, COP and SOP versus channel estimation accuracy.

The secure EE over different channel estimation accuracy are shown in Fig. 2. It is obvious that the secure EE is increasing function of channel estimation accuracy, and the secure EE of the system is determined by secure EE of each layer, which is consistent with the analytical results in (11). The observations from the Figs. 1 and 2 show that channel estimation effects reliability and security performance on the system. Therefore, it is always significant to improve channel estimation in every layer to increase the secure EE of dense HetNets with imperfect CSI, and the transmission scheme is effective to improve the performance of the system.

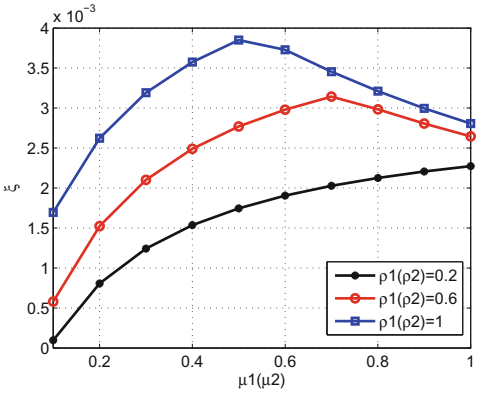


Fig. 3. Secure EE versus transmission threshold

The secure EE over varying transmission threshold with different ρ_1 are shown in Fig. 3. It is obvious that the secure EE is not a monotonic function of transmission threshold. Comparing with the system without transmission threshold, secure EE of the system with optimal transmission threshold can be increased by at least 50%. This implies that the transmission scheme can effectively improve the secure EE of dense HetNets with imperfect CSI, and it is significant to set a suitable transmission threshold. Therefore, it can help the designers of real dense HetNets to find the optimal transmission threshold according to the preference for different requirements.

6 Conclusion

In this paper, we propose a secure and reliable transmission scheme for dense HetNets based on secure EE optimization. We first establish the system model with channel estimation errors. Then, we design the transmission threshold selection algorithm to maximize the system secure EE, and perform emission control by comparing the receiving SINR with the transmission threshold. Constrained by the connection and secrecy probabilities, we evaluate the network-wide secure

EE. Numerical results show the secure and reliable transmission scheme can effectively improve the reliability and security performance of dense HetNets with imperfect CSI, which increases the secure EE by at least 50%.

Acknowledgment. This work is supported in part by Chinas High-Tech RD Program (863 Program) SS2015AA011306 and National Natural Science Foundation of China under Grants No. 61379006, 61401510, 61521003.

References

1. Yang, N., Wang, L., Geraci, G., Elkashlan, M.: Safeguarding 5G wireless communication networks using physical layer security. IEEE Commun. Mag. **53**(4), 20–27 (2015)
2. Wang, H.M., Zheng, T.X., Yuan, J., Towsley, D.: Physical layer security in heterogeneous cellular networks. IEEE Trans. Commun. **64**(3), 1204–1219 (2016)
3. Wu, H., Tao, X., Li, N.: Secrecy outage probability in multi-rat heterogeneous networks. IEEE Commun. Lett. **20**(1), 53–56 (2016)
4. Deng, Y., Wang, L., Wong, K.K., Nallanathan, A.: Safeguarding massive MIMO aided hetnets using physical layer security. In: International Conference on Wireless Communications Signal Processing, pp. 1544–1556 (2015)
5. Xu, M., Tao, X., Yang, F.: Enhancing secured coverage with comp transmission in heterogeneous cellular networks. IEEE Commun. Lett. **20**(11), 2272–2275 (2016)
6. Wang, C., Wang, H.M.: Opportunistic jamming for enhancing security stochastic geometry modeling and analysis. IEEE Trans. Veh. Technol. **65**(12), 10213–0217 (2016)
7. Gong, S., Xing, C., Fei, Z., Kuang, J.: Resource allocation for physical layer security in heterogeneous network with hidden eavesdropper. China Commun. **13**(Suppl. 3), 82–95 (2016)
8. Xu, X., Yang, W., Cai, Y., Jin, S.: On the secure spectral-energy efficiency tradeoff in random cognitive radio networks. IEEE J. Sel. Areas Commun. **34**(10), 2706–2722 (2016)
9. He, B., Zhou, X.: Secure on-off transmission design with channel estimation errors. IEEE Trans. Inf. Forensics Secur. **8**(12), 1923–1936 (2013)
10. Mukherjee, A., Swindlehurst, A.L.: Robust beamforming for security in MIMO wiretap channels with imperfect CSI. IEEE Trans. Sig. Process. **59**(1), 351–361 (2011)
11. Li, B., Fei, Z., Chen, H.: Robust artificial noise-aided secure beamforming in wireless-powered non-regenerative relay. IEEE Access **4**(4), 7921–7929 (2016)
12. Liu, C., Yang, N., Geraci, G., Yuan, J., Malaney, R.: Secrecy in MIMOME wiretap channels: beamforming with imperfect CSI. In: IEEE International Conference on Communications, pp. 4711–4716 (2014)
13. Chen, X., Chen, H.H.: Physical layer security in multi-cell MISO downlinks with incomplete CSI—a unified secrecy performance analysis. IEEE Trans. Sig. Process. **62**(23), 6286–6297 (2014)
14. Li, B., Fei, Z.: Robust beamforming and cooperative jamming for secure transmission in DF relay systems. EURASIP J. Wirel. Commun. Netw. **68**(1), 1–11 (2016)
15. Jo, H.S., Sang, Y.J., Xia, P., Andrews, J.G.: Heterogeneous cellular networks with flexible cell association: a comprehensive downlink SINR analysis. IEEE Trans. Wirel. Commun. **11**(10), 3484–3495 (2011)

Author Index

Bai, Baoming I-460, II-122
Bie, Zhisong I-146
Bouras, Taha II-383

Cai, Hui I-206, I-254
Chai, Yuan I-186
Chang, Hong II-450
Chang, Jianxin II-102
Chang, Jun II-405
Chang, Shuhao I-332
Chen, Daqing II-61
Chen, Dong II-474
Chen, Hang II-3
Chen, Jun II-132
Chen, Min-Rong II-153
Chen, Ping I-37
Chen, Qinbo I-206, I-254
Chen, Wei II-292
Chen, Wenqiu II-497
Chen, Xi I-332
Chen, Xiancai I-254
Chen, Xiang I-64, II-552
Chen, Xiaoming II-271
Chen, Yishun II-552
Chen, Zhaoxuan II-331
Cheng, Ran II-373
Cong, Li II-485
Cui, Panpan II-176

Dai, Man I-27
Dan, Zesheng I-297
Deng, Der-Jiunn II-182
Deng, Su I-492
Deng, Yichen II-82
Ding, Zhihao I-95
Dong, Jian I-524
Dong, Wenqian I-381
Dong, Zhenjiang II-312
Du, Jianchao I-443
Du, Qinghe I-84, I-95, I-341, II-283
Duan, Aoqin I-84

Fan, Shuquan I-448, I-546
Fan, Wei II-271
Fan, Weiqiang II-72
Fan, Yangyu I-274
Fang, Li I-37
Fang, Shu I-512
Feng, Hong I-195
Feng, Shixuan I-106, I-409
Fu, Zhi II-112

Gan, Jiefu II-373
Gang, Hui-Seon I-361
Gao, Ang II-474
Gao, Jinchun I-176
Gao, Ling II-95
Gao, Meng II-102
Gao, Tianrun II-3
Gao, Zibin II-323
Gong, Qianyun I-206, I-254
Guo, Lantian II-302

Han, Yuanchao I-512
Hao, Ming I-3
He, Chen I-264
He, Di II-383
He, Zunwen II-509
Hou, Bo II-132
Hu, Binjie II-563
Hu, Wei II-302
Huang, Di I-216, II-215
Huang, Hongbin I-492
Huang, Kaizhi I-557
Huang, Si I-226, II-352
Huang, Xin II-102
Huang, Xuewen II-519
Hui, Leifang I-425
Huo, Yuehua II-72

Ji, Baihua II-323
Ji, Hong I-195, I-433, II-341, II-395
Ji, Xiang I-27

Jialiang, Lu II-165
Jiang, Guofeng I-274
Jiang, Jichen II-395
Jiang, Lei II-250
Jiang, Wan I-481

Kang, Jiawen II-542
Ke, Sheng I-186
Ke, Tenghui II-323
Kuang, Rongjie II-323

Lei, Miao II-312
Leng, Supeng II-239
Li, Bo II-61
Li, Dantao I-240, II-141
Li, Dong II-72
Li, Guoquan I-155
Li, Haili II-13
Li, Hongxing I-37
Li, Huaan I-460
Li, Hui II-312
Li, Jieqi I-351, I-372
Li, Lei I-388, II-176
Li, Lizhao I-381
Li, Longjiang II-239
Li, Qi II-199
Li, Qingying II-61
Li, Shuyun I-186
Li, Weibin I-443
Li, Xi I-195, I-433, II-341, II-395
Li, Xiaohuan II-497, II-519
Li, Xiaohui I-226, I-240, II-141, II-227, II-352
Li, Xiaoya II-95
Li, Xiongfei I-54
Li, Yanchao II-41, II-51
Li, Yi I-37
Li, Yong II-450
Li, Zhengxue I-351, I-372
Lian, Baowang I-297
Liang, Zhonghua I-114, I-125, I-135
Liao, Guiwen II-405
Liao, Han II-426
Lim, Tiong Hoo II-463
Lin, Chun-Cheng II-182
Lin, Jinzhao I-155
Lin, Rongheng I-448, I-546
Lin, Xiaola II-552
Lin, Yuyang II-509
Ling, Zheng I-470
Liu, Chunxu I-166
Liu, Cong I-398
Liu, Danli I-114, I-125, I-135
Liu, Haojie I-114
Liu, Hui I-284
Liu, Jinjin I-125
Liu, Lei I-166
Liu, Pan I-54
Liu, Peilin II-383
Liu, Xiao II-323
Liu, Xiaotao II-563
Liu, Yiming I-166, II-112
Liu, Yingting II-485
Liu, Yinlong II-72
Liu, ZiYao I-216
Liu, Ziyao II-215
Lu, Baicen II-13
Lu, Liang-Yi II-182
Lu, Renhao II-440
Lu, Xuming I-64
Lu, Yueming II-373
Luo, Danyan I-524
Lv, Guoyun II-13, II-23
Lv, Pin II-331
Lv, Shanshan I-307
Lyu, Chao I-341

Ma, Dong I-502
Ma, La'ning I-114, I-135
Ma, Lixiang II-250
Ma, Mingyang I-372
Ma, Wei I-425
Mao, Baolei II-302
Mao, Di I-323
Martins, Philippe II-165
Mbom, Christian Ango II-189
Mei, Shaohui I-351, I-372
Meng, Danfeng I-240, II-141
Meng, Xiangyang I-54
Mu, Dejun II-302

Niu, Jinping II-95
Niu, Meng I-21, II-207
Nyarko, James Kweku Nkrumah II-189

Pan, Bin II-532
Pan, Di I-481
Pan, Siyu II-331
Pang, Yu I-155
Park, June-Buem I-361
Peng, Yang I-524
Pyun, Jae-Young I-361

Qian, Yueyan II-260
Qin, Maoyuan II-302
Qin, Tao II-292
Qiu, Xiaofeng I-264
Qiu, Xunjie II-132
Qu, Jiahui I-381

Ren, Jie I-535, II-95
Ren, Pinyi I-95, I-341, II-283

Shen, Weiqiang II-153
Sheng, Jie I-502
Shi, Chaoxiang II-102
Shi, Jinglun II-153
Song, Haina II-563
Song, Huansheng I-135
Song, Ke-Pu I-21, II-207
Song, Tianming I-433
Song, Yingjie I-3
Song, Yingshuan II-341
Su, Luo-Hui I-21, II-207
Su, Yun I-546
Sun, Bo I-84
Sun, Chengjun II-199
Sun, Hanwen I-425
Sun, Jinhua I-307
Sun, Li I-341
Sun, Yin I-64

Tai, Yu II-302
Tan, Feiyang I-388
Tan, Hongzhou I-64
Tang, Chengkai I-46
Tang, Xin II-497, II-519
Tian, Maoxin I-206, I-254
Tian, Zhongqi I-351
Tong, Yan II-292

Wan, Shuai I-372
Wang, Binbin I-481
Wang, Daoce II-323
Wang, Guan II-519
Wang, Hai II-95
Wang, Hao I-307
Wang, He II-199
Wang, Hongzhi II-416
Wang, Huichao II-239
Wang, Jiacheng I-155
Wang, Jun I-216, II-215
Wang, Liangliang I-64
Wang, Lihong I-106, I-409
Wang, Linlin II-23
Wang, Menglu I-332
Wang, Rong I-492
Wang, Ruonan II-405
Wang, Shikai II-13
Wang, Tonghui II-13
Wang, Wei II-312
Wang, Xia II-33
Wang, Xinluo II-312
Wang, Yu I-13
Wang, Zhaowei I-166
Wang, Zhifei I-492
Wang, Zhijie I-216, II-215
Wang, Zulin I-535
Wei, Fang I-481
Wei, Juan I-106, I-409
Weitao, Pan I-470
Wen, Fei II-383
Wu, Budan I-448, I-546
Wu, Daolong I-317
Wu, Fan II-239, II-250
Wu, Hao II-532
Wu, Jibing I-492
Wu, Jung-Chao II-182
Wu, Luqi II-33
Wu, Maoqiang II-542
Wu, Peng I-297
Wu, Weigang II-552
Wu, Wenjing I-125
Wu, Xiaofei I-481
Wu, Yahui I-492
Wu, Zhibo I-524

Xiao, Shoubai II-440
Xiao, Song I-381, I-388, II-176
Xie, Gang I-176
Xu, Fan I-502
Xu, Hengzhou I-75, I-284, I-460
Xu, Jia II-331
Xu, Mai I-535

Xu, Mengmeng I-75
Xu, Mingdi II-292
Xu, Yunjia I-557

Ya, Gao I-470
Yan, Chunman II-485
Yan, Dawei I-398
Yan, Yongfei II-416
Yan, Zhengqing II-82
Yang, Chunlei II-362
Yang, Fei I-146
Yang, Fuzheng I-417
Yang, Hongwu II-485
Yang, Mao II-112
Yang, Qinghai I-75
Yang, Qiong I-46, I-297
Yang, Xiaojun I-125
Yang, Xu II-227
Yang, Xuetao I-443
Yang, Yunchuan II-199
Yang, Zhiwei II-552
Yao, Mingwu II-260
Yao, Rongbin II-497, II-519
Ye, Dongdong II-542
Ye, Jincai II-497, II-519
Yin, Wei II-362
Yong, Shaowei I-398
You, Peng I-398
You, Yang I-502
Yu, Bin II-199
Yu, Jiang II-509
Yu, Ke I-481
Yu, Rong II-542
Yu, Wenxian II-383
Yu, Yanggege II-23
Yu, Yanping II-82
Yu, Zhong-yang II-122
Yuan, Hongbo I-274
Yuan, Ruiyang I-226, II-227, II-352
Yuchen, Wang II-165

Zeng, Guo-Qiang II-153
Zeng, Yu I-512
Zhang, Anxue II-271
Zhang, Bo I-284, I-460
Zhang, Chuanlin II-153
Zhang, Heli I-433, II-341
Zhang, Jian II-292
Zhang, Ke II-239, II-250
Zhang, Liang II-260
Zhang, Meng I-3
Zhang, Mengjun I-176
Zhang, Nangen II-215
Zhang, Ningxin II-23
Zhang, Qiang II-122
Zhang, Qilin II-132
Zhang, Xing-Long I-21, II-207
Zhang, Yan II-509
Zhang, Yanqiu I-512
Zhang, Yi I-46
Zhang, Yong I-351, I-372
Zhang, Yuexia II-3
Zhao, Mengdi I-37
Zhao, Sai I-206
Zhao, Yanbin I-226, I-240, II-141, II-227, II-352
Zhao, Zhe I-425
Zhen, Jiaqi II-41, II-51
Zheng, Jie II-95
Zhong, Xiaoxiong II-440
Zhou, Bin I-226, I-240, II-141, II-227, II-352
Zhou, Gang I-274
Zhou, Hongjian II-362
Zhou, Kaixiong II-102
Zhou, Mingyue II-416
Zhou, Shiqi I-448, I-546
Zhou, Xiangyun I-155
Zhou, Xianwen II-283
Zhou, Ying I-155
Zhou, Yuan I-27
Zhu, Canyan I-502
Zhu, Chunsheng I-433, II-341
Zhu, Hai I-284
Zhu, Hongyu I-155
Zhu, Min I-460
Zhu, Zhixiao II-563
Zi, Ran II-405
Zong, Rong II-405
Zou, Hua I-448, I-546
Zou, Wenjie I-417
Zou, Yi I-557
Zuo, Xiaoya I-351