

An International ICT R&D Journal Sponsored by ZTE Corporation

ISSN 1673-5188

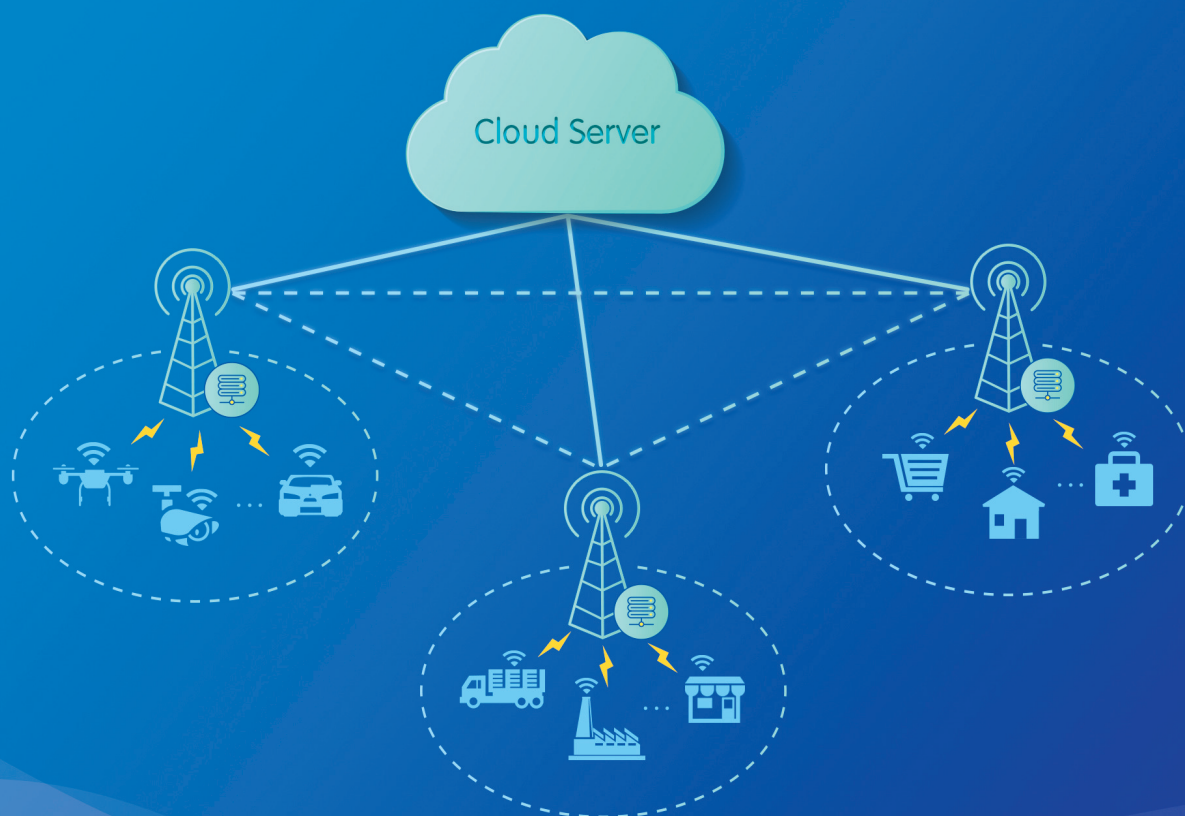
CN 34-1294/ TN

ZTE COMMUNICATIONS

中兴通讯技术(英文版)

June 2021, Vol. 19 No. 2

Special Topic: Edge Intelligence for Internet of Things



9 771673 518215



The 8th Editorial Board of ZTE Communications

Chairman	GAO Wen , Peking University (China)
Vice Chairmen	XU Ziyang , ZTE Corporation (China) XU Chengzhong , University of Macau (China)

Members (Surname in Alphabetical Order)

AI Bo	Beijing Jiaotong University (China)
CAO Jiannong	Hong Kong Polytechnic University (China)
CHEN Chang Wen	The State University of New York at Buffalo (USA)
CHEN Yan	Northwestern University (USA)
CHI Nan	Fudan University (China)
CUI Shuguang	UC Davis (USA) and The Chinese University of Hong Kong, Shenzhen (China)
GAO Wen	Peking University (China)
GAO Yang	Nanjing University (China)
GE Xiaohu	Huazhong University of Science and Technology (China)
HWANG Jenq-Neng	University of Washington (USA)
Victor C. M. LEUNG	The University of British Columbia (Canada)
LI Guifang	University of Central Florida (USA)
LI Xiangyang	University of Science and Technology of China (China)
LI Zixue	ZTE Corporation (China)
LIN Xiaodong	ZTE Corporation (China)
LIU Chi	Beijing Institute of Technology (China)
LIU Jian	ZTE Corporation (China)
LIU Ming	Institute of Microelectronics of the Chinese Academy of Sciences (China)
MA Jianhua	Hosei University (Japan)
MA Zheng	Southwest Jiaotong University (China)
NIU Zhisheng	Tsinghua University (China)
PAN Yi	Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences (China)
REN Fuji	Tokushima University (Japan)
REN Kui	Zhejiang University (China)
SHENG Min	Xidian University (China)
SONG Wenzhan	University of Georgia (USA)
SUN Huifang	Mitsubishi Electric Research Laboratories (USA)
SUN Zhili	University of Surrey (UK)
TAO Meixia	Shanghai Jiao Tong University (China)
WANG Haiming	Southeast University (China)
WANG Xiang	ZTE Corporation (China)
WANG Xiaodong	Columbia University (USA)
WANG Xiyu	ZTE Corporation (China)
WANG Yongjin	Nanjing University of Posts and Telecommunications (China)
WANG Zhengdao	Iowa State University (USA)
XU Chengzhong	University of Macau (China)
XU Ziyang	ZTE Corporation (China)
YANG Kun	University of Essex (UK)
YUAN Jinhong	University of New South Wales (Australia)
ZENG Wenjun	Microsoft Research Asia (China)
ZHANG Chengqi	University of Technology Sydney (Australia)
ZHANG Honggang	Zhejiang University (China)
ZHANG Jianhua	Beijing University of Posts and Telecommunications (China)
ZHANG Yueping	Nanyang Technological University (Singapore)
ZHOU Wanlei	City University of Macau (China)
ZHUANG Weihua	University of Waterloo (Canada)

CONTENTS

ZTE COMMUNICATIONS June 2021 Vol. 19 No. 2 (Issue 74)

Special Topic

Edge Intelligence for Internet of Things

Editorial 01

LIU Chi

RecCac: Recommendation-Empowered Cooperative Edge Caching for Internet of Things 02

Cooperative edge caching jointing the neighbor edge server is regarded as a promising technique to improve cache hit and reduce congestion of the networks. The authors investigate the issue of joint cooperative edge caching and recommender systems to achieve additional cache gains by the soft caching framework. To measure the cache profits, the optimization problem is formulated as a 0-1 ILP, which is NP-hard. Specifically, the method of processing content requests is defined as server actions, the authors determine the server actions to maximize the QoE and propose a cache-friendly heuristic algorithm to solve it.

*HAN Suning, LI Xiuhua, SUN Chuan, WANG Xiaofei,
Victor C. M. LEUNG*

Cost-Effective Task Scheduling for Collaborative Cross-Edge Analytics 11

To explicitly leverage the price heterogeneity for WAN cost minimization, the authors propose to schedule analytic tasks based on both price and bandwidth heterogeneities. Unfortunately, the problem of WAN cost minimization under performance constraint is shown NP-hard, thus computationally intractable for large inputs. To address this challenge, PPGA, an efficient task scheduling heuristic that improves the cost-efficiency of IoT data analytic jobs across edge data-centers is proposed. The authors implement PPGA based on Apache Spark and conduct extensive experiments on Amazon EC2 to verify the efficacy of PPGA.

ZHAO Kongyang, GAO Bin, ZHOU Zhi

20 BPPF: Bilateral Privacy-Preserving Framework for Mobile Crowdsensing

In this paper, the authors study privacy protection in MCS. The main challenge is to assign the most suitable worker to a task without knowing the task and the actual location of the worker. The authors propose a bilateral privacy protection framework based on matrix multiplication, which can protect the location privacy between task and worker, and keep their relative distance between task and worker unchanged.

LIU Junyu, YANG Yongjian, WANG En

29 Maximum-Profit Advertising Strategy Using Crowdsensing Trajectory Data

The authors propose some effective advertising strategies for selecting an effective set of billboards under the advertising budget to maximize commercial profit for the advertiser. First, the authors extract potential customers' implicit information and then study the billboard selection problem under two situations. Extensive experiments based on three real-world data sets verify that the proposed advertising strategies can achieve the superior commercial profit compared with the state-of-the-art strategies.

LOU Kaihao, YANG Yongjian, YANG Funing, ZHANG Xingliang

44 Speed Estimation Using Commercial Wi-Fi Device in Smart Home

A direction independent indoor speed estimation system in terms of EM wave statistical theory is proposed. Based on the statistical characteristics of EM waves, the authors establish the deterministic relationship between the ACF of CSI and the speed of moving target. Extensive experiments show that the system achieves a median error of 0.18 m/s for device-free single target walking speed estimation.

TIAN Zengshan, YE Chenglin, ZHANG Gongzhui, HE Wei, JIN Yue

Submission of a manuscript implies that the submitted work has not been published before (except as part of a thesis or lecture note or report or in the form of an abstract); that it is not under consideration for publication elsewhere; that its publication has been approved by all co-authors as well as by the authorities at the institute where the work has been carried out; that, if and when the manuscript is accepted for publication, the authors hand over the transferable copyrights of the accepted manuscript to *ZTE Communications*; and that the manuscript or parts thereof will not be published elsewhere in any language without the consent of the copyright holder. Copyrights include, without spatial or timely limitation, the mechanical, electronic and visual reproduction and distribution; electronic storage and retrieval; and all other forms of electronic publication or any other types of publication including all subsidiary rights.

Responsibility for content rests on authors of signed articles and not on the editorial board of *ZTE Communications* or its sponsors.

All rights reserved.

CONTENTS

ZTE COMMUNICATIONS June 2021 Vol. 19 No. 2 (Issue 74)

Review

Analysis of Industrial Internet of Things and Digital Twins **53**

In this paper, the efforts of 3GPP are introduced for the development of uRLLC in reducing delay and enhancing reliability, as well as the research on little jitter and high transmission efficiency. The enhanced key technologies required in the IIoT are also analyzed. Finally, digital twins are analyzed according to the actual IIoT situation.

TAN Jie, SHA Xiubin, DAI Bo, LU Ting

Research Paper

Higher Speed Passive Optical Networks for Low Latency Services **61**

Latency sensitive services have attracted much attention lately and imposed stringent requirements on the access network design. PON provides a potential long-term solution for the underlying transport network supporting these services. This paper discusses latency limitations in PON and recent progress in PON standardization to improve latency. Experimental results of a low latency PON system are presented as a proof of concept.

ZHANG Weiliang, YUAN Liquan

Differentially Authorized Deduplication System Based on Blockchain **67**

To further refine the usage scenarios for various user permissions and enhance user's data security, the authors propose a blockchain-based differential authorized deduplication system, which optimizes the traditional PoV consensus algorithm and simplifies the existing differential authorization process to realize credible management and dynamic update of authority. Based on the decentralized property of

blockchain, the authors overcome the centralized single point fault problem of traditional differentially authorized deduplication system. Besides, the operations of legitimate users are recorded in blocks to ensure the traceability of behaviors.

ZHAO Tian, LI Hui, YANG Xin, WANG Han, ZENG Ming, GUO Haisheng, WANG Dezheng

77 A Novel De-Embedding Technique of Packaged GaN Transistors

This paper presents a novel de-embedding technique of packaged high-power transistors. Different from the conventional technique of parasitic extraction, the proposed technique only requires external measurements. The frequency independent characteristic of DID is verified and the IPN is modeled and calibrated for a 50 W GaN transistor. At last, a broadband Doherty PA is fabricated with the de-embedding technique. According to the measured results, the PA exhibits its satisfactory power and efficiency performance.

WEI Xinghui, CHEN Xiaofan, CHEN Wenhua, ZHOU Junmin

82 Flexible Multiplexing Mechanism for Coexistence of URLLC and EMBB Services in 5G Networks

A dynamic 2-dimension bitmap resource indication is proposed to cancel eMBB services with a finer uplink cancellation granularity and a lower probability of false cancellation. Meanwhile, a resource indication based power control method is introduced to dynamically indicate different power control parameters to the UE based on different time-frequency resource groups and the proportion of overlapping resources. Furthermore, a dynamic selection mechanism is proposed to accommodate the varying cases in different scenarios. Extensive system level simulations are conducted and the results show that about 10.54% more URLLC UEs satisfy the requirements, and the perceived throughput of eMBB UEs is increased by 23.26%.

XIAO Kai, LIU Xing, HAN Xianghui, HAO Peng, ZHANG Junfeng, ZHOU Dong, WEI Xingguang

Serial parameters:CN 34-1294/TN*2003*q*16*90*en*P*¥20.00*2200*11*2021-06

Statement

This magazine is a free publication for you. If you do not want to receive it in the future, you can send the "TD unsubscribe" mail to magazine@zte.com.cn. We will not send you this magazine again after receiving your email. Thank you for your support.



Editorial: Special Topic on Edge Intelligence for Internet of Things



Guest Editor

LIU Chi received the Ph.D. degree from Imperial College London, UK in 2010, and the B.Eng. degree from Tsinghua University, China in 2006. He is currently a full professor and Vice Dean at the School of Computer Science and Technology, Beijing Institute of Technology, China. He is also the Director of IBM Mainframe Excellence Center (Beijing) and Director of IBM Big Data Technology Center. Before moving to academia, he joined IBM Research—China as a staff researcher and project manager, after working as a postdoctoral researcher at Deutsche Telekom Laboratories, Germany and a visiting scholar at IBM T. J. Watson Research Center, USA. His current research interests include the big data analytics, mobile computing, and deep learning. He received the Distinguished Young Scholar Award in 2013, IBM First Plateau Invention Achievement Award in 2012, and was interviewed by EEWeb.com as the Featured Engineer in 2011. He has published more than 200 prestigious conference and journal papers and owned more than 14 EU/US/UK/China patents, with Google H index 26. He currently serves as the symposium chair for IEEE ICC 2020 Next Generation Networking, an area editor for *KSII Transactions on Internet and Information Systems* and the book editor for six books published by Taylor & Francis Group, USA and China Machinery Press. He has also served as the general chair of IEEE SECON'13 workshop on IoT Networking and Control, IEEE WCNC'12 workshop on IoT Enabling Technologies, and ACM UbiComp'11 Workshop on Networking and Object Memories for IoT. He served as the consultant to Asian Development Bank, Bain & Company, and KPMG, USA, and the peer reviewer for Qatar National Research Foundation, and National Science Foundation, China. Dr. LIU is a fellow of IET and a fellow of Royal Society of the Arts.

With the fast development of the Internet of Things (IoT), increasing numbers of devices are connected to the IoT network and generate massive amounts of data. The traditional centralized cloud computing architecture cannot meet the requirements of both latency and data security. Within this context, edge computing and edge intelligence can shift data processing, computing applications and services from centralized cloud servers to the edge of a network. Depending on the processing requirements, the data generated by an edge device is either processed at the device itself or the local node deployed at the periphery of the network. This enables the analysis to be performed near the data source to avoid the delay caused by moving data to the cloud. It also allows sensitive data to be processed locally, and only non-sensitive data will be transmitted to the data center for centralized processing. On the other hand, the application of edge intelligence still faces many challenges, such as inconsistent IoT technical standards, serious energy consumption and pollution, heterogeneous IoT network and data, and data security and privacy protection.

To overcome these challenges, the first paper “RecCac: Recommendation-Empowered Cooperative Edge Caching for Internet of Things” by LI Xiuhua et al. focuses on the joint problem of cooperative edge caching and recommender systems to achieve additional cache gains and increase “effective” cache size. This paper supports massive content access in mobile edge networks and addresses rapidly growing IoT

services and content applications.

The second paper is dedicated to reducing the cost of cross-edge analysis. “Cost-Effective Task Scheduling for Collaborative Cross-Edge Analytics” by ZHAO Kongyang et al. empirically demonstrates that reducing either analytics response time or network traffic volume cannot necessarily minimize the wide area network (WAN) traffic cost, due to price heterogeneity of WAN links, so they propose to schedule analytic tasks based on both price and bandwidth heterogeneities.

The remaining three papers mainly study the application of edge intelligence. “BPPF: Bilateral Privacy-Preserving Framework for Mobile Crowdsensing” by LIU Junyu et al. studies privacy protection in Mobile Crowdsensing (MCS) and proposes a bilateral privacy protection framework (BPPF) based on matrix multiplication for protecting the location privacy between the task and the worker, and keep their relative distance unchanged. “Maximum-Profit Advertising Strategy Using Crowdsensing Trajectory Data” by LOU Kaihao et al. also works with MCS and proposes some effective advertising strategies to help maximize commercial profit for the advertiser by attracting potential customers using out-door billboard advertising. “Speed Estimation Using Commercial Wi-Fi Device in Smart Home” by TIAN Zengshan et al. studies Wi-Fi-based approaches to measure the speed of the moving target and proposes a direction independent indoor speed estimation system in terms of electromagnetic wave statistical theory.

Finally, we would like to thank all the authors for contributing their papers to this special issue and the external reviewers for volunteering their time to review the submissions.

DOI: 10.12142/ZTECOM.202102001

Citation (IEEE Format): C. Liu, “Editorial: special topic on edge intelligence for Internet of Things,” *ZTE Communications*, vol. 19, no. 2, pp. 01–01, Jun. 2021. doi: 10.12142/ZTECOM.202102001.

RecCac: Recommendation-Empowered Cooperative Edge Caching for Internet of Things



HAN Suning¹, LI Xiuhua¹, SUN Chuan¹, WANG Xiaofei², Victor C. M. LEUNG^{3,4}

(1. Chongqing University, Chongqing 400000, China;

2. Tianjin University, Tianjin 300072, China;

3. Shenzhen University, Shenzhen 518000, China;

4. The University of British Columbia, Vancouver V6T 1Z4, Canada)

Abstract: Edge caching is an emerging technology for supporting massive content access in mobile edge networks to address rapidly growing Internet of Things (IoT) services and content applications. However, the edge server is limited with the computation/storage capacity, which causes a low cache hit. Cooperative edge caching jointing neighbor edge servers is regarded as a promising technique to improve cache hit and reduce congestion of the networks. Further, recommender systems can provide personalized content services to meet user's requirements in the entertainment-oriented mobile networks. Therefore, we investigate the issue of joint cooperative edge caching and recommender systems to achieve additional cache gains by the soft caching framework. To measure the cache profits, the optimization problem is formulated as a 0-1 Integer Linear Programming (ILP), which is NP-hard. Specifically, the method of processing content requests is defined as server actions, we determine the server actions to maximize the quality of experience (QoE). We propose a cache-friendly heuristic algorithm to solve it. Simulation results demonstrate that the proposed framework has superior performance in improving the QoE.

Keywords: IoT; recommender systems; cooperative edge caching; soft caching

DOI: 10.12142/ZTECOM.202102002

<https://kns.cnki.net/kcms/detail/34.1294.TN.20210525.1320.002.html>, published online May 26, 2021

Manuscript received: 2021-03-01

Citation (IEEE Format): S. N. Han, X. H. Li, C. Sun, et al. "RecCac: recommendation-empowered cooperative edge caching for Internet of Things," *ZTE Communications*, vol. 19, no. 2, pp. 02-10, Jun. 2021. doi: 10.12142/ZTECOM.202102002.

1 Introduction

As the development trend of future networks, the Internet of things (IoT) has become a hot research topic in the industry and academia in recent years^[1]. The emergence of "IoT" paradigm makes accessi-

bility of various IoT sensors (e.g., smart cameras and temperature sensors) universal, and thus enables intelligent services to improve the life quality of humans^[2]. Billions of IoT devices (IDs) generate a tremendous number of monitoring data while a great many end-users are consuming these data. However, countless electronic devices are anticipated to generate a sheer volume of traffic loads, and the aggregate load on core networks is expected to be large. Therefore, it is important to reduce congestion and transmission delay for network providers^[3-5].

As we have stated above, mobile edge networks are faced with the challenge of the explosive growth of IoT data requests from the IDs, especially in the current backhaul net-

This work is supported in part by National Key R&D Program of China under Grant Nos. 2018YFB2100100 and 2018YFF0214700, National NSFC under Grant Nos. 61902044 and 62072060, Chongqing Research Program of Basic Research and Frontier Technology under Grant No. CSTC2019-jcyjmsxmX0589, Key Research Program of Chongqing Science and Technology Commission under Grant Nos. CSTC2017jcyjBX0025 and CSTC2019jcsx-zdztzxX0031, Fundamental Research Funds for the Central Universities under Grant No. 2020CDJQY-A022, Chinese National Engineering Laboratory for Big Data System Computing Technology, and Canadian NSERC.

works^[6]. According to the research, most of the high load in the mobile networks is generated by downloading the same content and data. To solve this problem, it is necessary to put forward new revolutionary methods in network structure and data transmission^[7]. As one of the rapidly developing technologies, edge caching has drawn growing attention. Edge caching technology can reduce repeated downloading and transmission by caching contents in advance^[8]. However, as content providers (CPs) provide growing content, and the storage and computing capacity of a cell (e.g., edge server) are limited, we still face great challenges to solve the above problems. Many researchers are looking for additional cache gains in this area. Some current research (e.g., FemtoCache^[9]) focuses on caching contents in the edge server of base stations (BSs). However, it only focuses on the basic cell cache, and the understanding of inter-cell cooperation is not deep.

Besides, how to use the cached contents to achieve more cache gains is also a problem we have to consider. It is difficult to improve the caching performance only by focusing on the content popularity in the entertainment-oriented mobile networks. To solve this problem, the recommender systems provide an effective method that can provide personalized content recommendations through historical behavior, e.g., users may have evaluated or scored different contents. However, some related content, such as two similar comedy movies or two short videos of the same type, might have similar utility for a user. We use the term—soft caching^[10], which means that if the local BS doesn't cache the requested content, the BS can send other relevant contents available locally. If the user likes or accepts the relevant contents (under a certain threshold) instead of the content which was originally requested, a soft cache hit will occur. This scheme may give up some content relevance, but it avoids the “expensive” connection of the IDs to get the requested content from the backhaul network. Actually, some recent experimental evidence suggests that IDs may be willing to trade off some content relevance for a better quality of experience (QoE)^[11].

More specifically in this paper, the cooperative edge caching and recommender systems are used to alleviate the pressure of the backhaul network and get related contents to achieve soft caching, respectively. We combine cooperative edge caching with recommender systems to improve the QoE. Recently, some researchers consider the interaction between edge caching and recommender systems to optimize cache or recommender systems^[10–17]. However, most of the research only focuses on one side of the problem, e.g., caching-friendly recommendations^[10, 12–13, 15, 17] or recommendation-aware caching policies^[16]. The real joint treatment of both is tried in Refs. [11] and [14], but their studies on hierarchical mobile edge networks are not deep enough.

To sum up, different from the existing studies on edge caching and recommender systems, we focus on improving the QoE by judiciously selecting server actions. Our main contributions are summarized as follows:

1) We combine cooperative edge caching with soft caching for IoT systems. To measure cache profits, we propose a generic metric of QoE that depends on the quality of service (QoS) and the quality of recommendation (QoR).

2) We formulate the problem of optimally choosing the server actions towards maximizing the QoE. While such joint caching and recommendation problems have been proved to be NP-hard, we have proposed a cache-friendly hierarchical heuristic algorithm.

3) Trace-driven evaluation results demonstrate that our proposed scheme has superior performance on improving the cache hits and QoE finally.

The remainder of this paper is organized as follows. Section 2 discusses the proposed hierarchical cooperative edge caching model and formulates the optimization problem. Section 3 introduces a cache-friendly hierarchical heuristic algorithm to solve the problem. Section 4 evaluates the performance of the proposed framework and Section 5 concludes this paper.

2 System Model and Problem Formulation

In this section, we introduce the system model of edge caching. Specifically, we present the hierarchical cooperative edge caching architecture and topology in Section 2.1. Section 2.2 introduces the recommendation-aware content request processing model. Then we propose a QoE model, considering delay and recommendation in Section 2.3. Finally, Section 2.4 gives the problem formulation. Some key parameters are listed in Table 1.

▼Table 1. Key Parameters

Notation	Definition
N	Number of BSs
M	Number of IDs
F	Total number of contents
C	Cache size of a BS
D_f	Size of the content f
p_f	The content f requested probability
$\alpha_{m,n}$	The association of the ID and the BS
$s_{n,f}$	The cache state of the content f in the BS n
$w_{m,f}$	Rating (i.e., preference) for the content f of the ID m
$v_{m,n}$	The wireless transmission rate between the ID and the BS
$\pi_{m,n,f}$	System action
$d_{m,n,f}^L, d_{m,n,f}^E, d_{m,n,f}^C$	Transmission delay of the content f between the BS n and the ID m , BS and BS, BS and CS, respectively
$r_{m,n,f}^L, r_{m,n,f}^E, r_{m,n,f}^C$	Content satisfaction in different server actions

BS: base station CS: cloud server

2.1 Hierarchical Cooperative Edge Caching Model

The proposed system is a cooperative Cloud-Edge-End computing system with a cloud server (CS), some discrete BSs, and IDs. As shown in Fig. 1, we consider a cooperative edge caching scenario for IoT networks. The CS has enough computing and caching capacity, consisting of all data and contents. Each BS is equipped with an edge server, which has the limited ability to cache and compute. Each ID as a content requester generates a request at each time slot. In our proposed system, each BS communicates with the CS through the backhaul links. To enhance the usage of the BSs and alleviate the pressure of the backhaul networks, each BS can communicate with all cooperative BSs through fronthaul links instead of working individually^[18]. Besides, as the contents are cached in the BSs or the CS, IDs can fetch their requested contents either from edge servers via wireless links or directly by downloading the contents from the CS to the BSs.

The proposed system consisting of $\mathcal{N} = \{1, 2, \dots, N\}$ fully connected BSs with a finite cache size C and $\mathcal{M} = \{1, 2, \dots, M\}$ IDs are distributed in the service area of the BSs. In addition, we denote $a_{m,n} \in [0, 1]$ as the association probability between the BS n and the ID m . We assume each ID requests content or a set of data from a catalogue $\mathcal{F} = \{1, 2, \dots, F\}$ at each time slot, and we denote the size of each content as D_f .

We assume that the ID m requests the content f with the standard content probability p_f^m . Hence, we could obtain the content popularity p_f from p_f^m ^[9]. Furthermore, we assume that the content popularity p_f changes slowly, and $\sum_{f=1}^F p_f = 1$.

For the cache state, we focus on whether the content has been cached in the BSs. The content cache state is denoted

as $s_{nf} \in \{0, 1\}$, $\forall n \in \mathcal{N}$, $\forall f \in \mathcal{F}$. Here, $s_{nf} = 1$ represents that the BS n has cached the content f , otherwise $s_{nf} = 0$.

2.2 Recommendation-Aware Request Processing Model

We define a score w_{mf} to represent the ID's preference for the content or data f . As for p_f , it denotes the probability of the ID m requesting the content f . Specifically, given the scores w_{mf} , a reasonable choice could be their normalized values:

$$p_f = \frac{w_{mf}}{\sum_{i \in \mathcal{F}} w_{m,i}}. \quad (1)$$

Since soft caching is to replace the requested content with related contents or data available in the local BS, we rank the scores in a descending order to get a recommendation list K_m of the ID m . When a content request f generated by the ID m arrives at the local BS, there are three types of situation:

1) Local hits: Local hits denote that the local BS processes content requests. The local hits are divided into direct cache hits and soft cache hits.

2) Neighboring hits: the request generated by an ID can be obtained from its cooperative BSs, and the transmission delay is relatively small compared with downloading from the CS.

3) CS hits: The ID obtains the requested content from the CS. The transmission in this situation is known as "expensive".

We model the server actions of the content request with three sub-decisions models, denoted as $\pi_{m,nf} = (\pi_{m,nf}^L, \pi_{m,nf}^E, \pi_{m,nf}^C)$, where $\pi_{m,nf}^L, \pi_{m,nf}^E, \pi_{m,nf}^C \in \{0, 1\}$ are the indicators for whether the request is processed in the local BS, cooperative BSs, or the CS. Three sub-decisions can jointly determine how the request is processed. Different decisions will affect the transmission delay and content satisfaction.

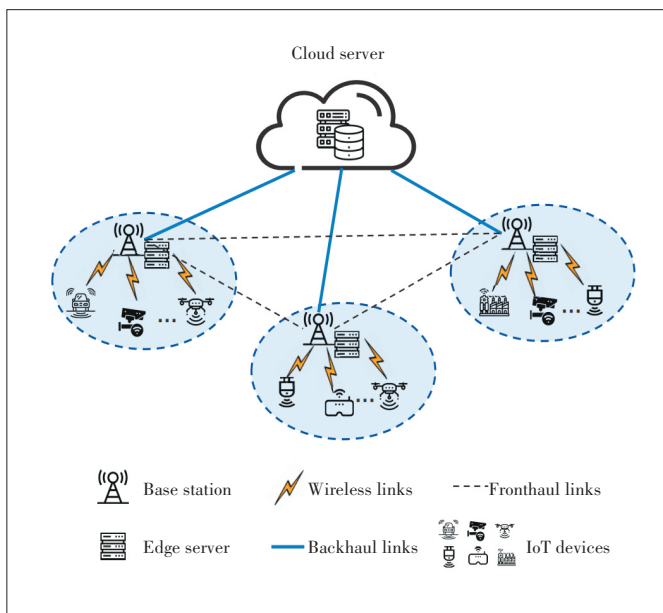
As the content is indivisible, so for $\forall m \in \mathcal{M}$, only one of $\pi_{m,nf}^L, \pi_{m,nf}^E$ and $\pi_{m,nf}^C$ can be 1. Similar to Ref. [19], the decision variable $\pi_{m,nf}$ is constrained by

$$\pi_{m,nf}^L + \pi_{m,nf}^E + \pi_{m,nf}^C = 1. \quad (2)$$

2.3 QoE Model

We define the QoE as a combination of the QoS and the QoR. The QoS and QoR are measured by the transmission delay and content satisfaction, respectively. In the following, we will discuss the two parts with different decisions in detail:

1) Delay: We consider the transmission delay as the time for an ID to receive the contents or data. In the proposed system, there are three delay parts: $d_{m,nf}^L$ denotes the transmission delay that the ID m receives the content from the local BS n , d_f^E denotes the transmission delay of the BSs' co-



▲ Figure 1. Cooperative edge caching supporting IoT architecture

operation, and d_f^C denotes the transmission delay between the BS and the CS.

Specifically, we assume that the wireless channel has been deployed. Similar to Ref. [20], we can get the transmission rate between the ID m and the local BS n as follow:

$$v_{m,n} = B \log_2 \left(1 + \frac{P_m g_{m,n}}{\sigma^2} \right), \quad (3)$$

where B denotes the channel bandwidth; σ^2 denotes the background noise power; P_m denotes the power consumption of the BS n transmission to the ID m . The channel gain $g_{m,n}$ is estimated by the distance $l_{m,n}$ between the local BS n and the ID m .

Thus, the delay of transferring the content m between the ID f and the local BS n is denoted as:

$$d_{m,n,f}^L = a_{m,n} s_{n,f} \frac{D_f}{v_{m,n}}, \quad (4)$$

The transmission among the cooperative BSs is through fronthaul links with high bandwidth. In terms of the transmission between the CS and the BSs, the CS is usually deployed at a further distance, and a large amount of traffic is transmitted through multiple intermediate nodes; We express these two parts in terms of the average rate; v_e denotes the average transmission rate between two BSs. Therefore, the transmission delay between cooperative BSs can be expressed as follow:

$$d_f^E = \frac{D_f}{v_e}. \quad (5)$$

Similarly, v_c denotes the average transmission rate between the BSs and the CS. The transmission delay between the BSs and the CS can be expressed as:

$$d_f^C = \frac{D_f}{v_c}. \quad (6)$$

2) Recommendation: If the content requested by the ID is not cached locally, the similar contents cached locally could be alternated.

Specifically, for local hits, considering the soft caching, we define the content satisfaction as:

$$r_{m,n,f}^L = a_{m,n} s_{n,f} w_{m,f}. \quad (7)$$

Similarly, for neighboring BSs cache hits, we define the content satisfaction as:

$$r_{m,n,f}^E = s_{n,f} w_{m,f}. \quad (8)$$

For downloading the content f from the CS, we define the content satisfaction as:

$$r_{m,f}^C = w_{m,f}. \quad (9)$$

2.4 Problem Formulation

In the proposed system, our goal is to find the best server actions to improve the QoE. As we have discussed above, transmission delay and content satisfaction are major factors. We express these two parts as follows:

$$d_{m,n,f} = \frac{\pi_{m,n,f}^L}{d_{m,n,f}^L} + \frac{\pi_{m,n,f}^E}{d_{m,n,f}^L + d_f^E} + \frac{\pi_{m,n,f}^C}{d_{m,n,f}^L + d_f^C}, \quad (10)$$

$$r_{m,n,f} = r_{m,n,f}^L \pi_{m,n,f}^L + r_{m,n,f}^E \pi_{m,n,f}^E + r_{m,f}^C \pi_{m,n,f}^C, \quad (11)$$

where Eq. (10) denotes the QoS, which is expressed as the reciprocal of the delay of content transmission (i.e., when the delay of the content transmission is small, the larger QoS can be obtained), $(d_f^E + d_{m,n,f}^L)$ denotes the transmission delay when the content is sent through the cooperative BSs, and $(d_f^C + d_{m,n,f}^L)$ denotes the transmission delay when the content is downloaded from the CS. Eq. (11) denotes the QoR.

To improve the QoE, we need to trade off the QoS and the QoR (i.e., find the balance between low transmission delay and high content satisfaction) by optimizing the server actions $\pi_{m,n,f}$. To maximize the QoE, we formulate the optimization problem as:

$$\mathcal{P}: \max \sum_m \sum_n \sum_{f \in \mathcal{F}} p_f (\alpha d_{m,n,f} + \beta r_{m,n,f}) \quad (12a)$$

$$\text{s.t.} \quad \alpha + \beta = 1, \quad (12b)$$

$$s_{n,f} \in \{0,1\}, \forall n \in \mathcal{N}, \forall f \in \mathcal{F}, \quad (12c)$$

$$\forall \pi_{m,n,f}^L \in \{0,1\}, \forall \pi_{m,n,f}^E \in \{0,1\}, \forall \pi_{m,n,f}^C \in \{0,1\}, \quad (12d)$$

$$\sum_n \sum_{f \in \mathcal{F}} \pi_{m,n,f} = 1, \forall m \in \mathcal{M}, \quad (12e)$$

$$\sum_{f \in \mathcal{F}} \pi_{m,n,f} D_f \leq C, \forall m \in \mathcal{M}, \forall n \in \mathcal{N}, \quad (12f)$$

where p_f denotes the probability of the content or data f requested. In Eq. (12b), α and β are the scalar parameters to balance transmission delay and content satisfaction. Eq. (12c) denotes the cache state. Eqs. (12d) and (12e) denote the constraints of the server actions. Eq. (12f) denotes the cache ability.

To represent the server actions of the system, we denote $\Pi = (\Pi^L, \Pi^E, \dots, \Pi^C)$ as the entire selection, where $\Pi^L = (\Pi_1^L, \Pi_2^L, \dots, \Pi_M^L)$, $\Pi^E = (\Pi_1^E, \Pi_2^E, \dots, \Pi_M^E)$, and $\Pi^C = (\Pi_1^C, \Pi_2^C, \dots, \Pi_M^C)$. And we denote $\Pi_m^L = \{\pi_{m,n,f}^L | \forall n \in \mathcal{N}, \forall f \in \mathcal{F}\}$,

$\Pi_m^E = \{\pi_{m,nf}^E | \forall n \in \mathcal{N}, \forall f \in \mathcal{F}\}$, and $\Pi_m^C = \{\pi_{m,nf}^C | \forall n \in \mathcal{N}, \forall f \in \mathcal{F}\}$.

Lemma 1: The QoE problem is equivalent to the 0 - 1 Integer Linear Programming (ILP) problem.

Proof: As mentioned above, different server actions will affect the transmission delay and content satisfaction. We

denote positive constants $A_{m,nf}^1 = \alpha / \frac{a_{m,n} s_{nf} D_f}{\log_2 \left(1 + \frac{P_m g_{m,n}}{\sigma^2} \right)} +$

$$\beta a_{m,n} s_{nf} w_{mf}, \quad A_{m,nf}^2 = \left[\alpha / \left(\frac{a_{m,n} s_{nf} D_f}{\log_2 \left(1 + \frac{P_m g_{m,n}}{\sigma^2} \right)} + \frac{D_f}{v_c} \right) \right] + \beta s_{nf} w_{mf},$$

$$\text{and } A_{m,nf}^3 = \left[\alpha / \left(\frac{a_{m,n} s_{nf} D_f}{\log_2 \left(1 + \frac{P_m g_{m,n}}{\sigma^2} \right)} + \frac{D_f}{v_c} \right) \right] + \beta w_{mf} \quad \text{To com-}$$

bine optimization objectives with decision variables, the optimization objective of the problem in Eq. (12) can be expressed as:

$$\mathcal{P}: \max_{\Pi} \sum_{m \in \mathcal{M}} \sum_{n \in \mathcal{N}} \sum_{f \in \mathcal{F}} \text{pr} \left(A_{m,nf}^1 \pi_{m,nf}^L + A_{m,nf}^2 \pi_{m,nf}^E + A_{m,nf}^3 \pi_{m,nf}^C \right), \quad (13a)$$

s.t. the same as Eq.s (12b), (12c), (12d), (12e), and (12f).

(13b)

Thus, the problem can be described as selecting optimal server actions for processing requests with jointing transmission delay and content satisfaction. This is a 0 - 1 ILP problem, which is NP-hard. Because the number of IDs, BSs, and contents can be large, it is of high complexity to get the optimal solution by using exact methods.

3 Proposed Framework Design

The proposed system is a hierarchical cooperation orchestrated computing topology. We focus on improving the QoE by judiciously selecting the server actions. Different server and content selections affect the final server actions. Thus, to address the above complex optimization Eq. (13), we decompose it into two simpler subproblems as below.

1) Inner algorithm for recommendation list. First, we obtain the recommendation list K_m for the ID m from the content or data catalog, which is implemented by the collaborative filtering algorithm based on items-Inverse User Frequency (ItemCF-IUF). The inner algorithm is mainly divided into two steps: calculating the similarity between two contents and generating the recommendation list. When calculating the similarity, we consider the influence of the ID

m activity on content similarity. We use the improved cosine formula to calculate the similarity between the content i and f as:

$$\text{sim}_{if} = \frac{\sum_{m \in N_i \cap N_f} \frac{1}{\log^{1+|N(m)|}}}{\sqrt{|N_i| \cup |N_f|}}, \quad (14)$$

where N_i denotes the number of IDs that like the content i , N_f denotes the number of IDs that like the content f , and $|N_i| \cup |N_f|$ denotes the number of IDs that both like content i and f . Then the score of the content f will be calculated.

Then we sort w_{mf} in a descending order to generate the final recommendation list of the ID m . The details of the proposed method for solving the inner problem are shown in Algorithm 1. The internal of the loop consists of $|\mathcal{F}|$ calculations. Next, the complexity of the sorting step is $O(\log |\mathcal{F}|)$ in a pre-ordered list. Since these steps are repeated for every ID m , the total complexity of the algorithm is $O(|\mathcal{M}| |\mathcal{F}|)$.

Algorithm 1: Inner Algorithm for Recommendation.

Input: \mathcal{M} , \mathcal{F} , and all IDs' information.

Output: $\{K\}_{M \times R}$.

1 Initialization: $\{K\}_{M \times R} \leftarrow 0_{M \times R}$;

2 **for** the ID $m \in \mathcal{M}$ **do**

3 **for** each content pair of (i, f) , $\forall i, f \in \mathcal{F}$ **do**

4 Calculate sim_{if} and w_{mf} ;

5 Sort w_{mf} in decreasing order;

6 Choose the top R contents into the K_m ;

7 Add K_m to $\{K\}_{M \times R}$;

8 **end**

9 **end**

2) Server actions. We optimize the server actions. As mentioned above, Π has $3MNF$ possible selections. It may be easy to find the optimal solution in a small scenario. Since the number of IDs, BSs, and contents can be large, it will take abundant time to converge if we use the general exhaustive methods (e.g., checking each combination of variables with a value of 0 or 1, and comparing the value of the objective function to obtain the optimal solution). To solve the problem, we propose a cache-friendly heuristic algorithm with the branch and bound (BNB) strategy.

Lemma 2: Eq. (13) can be divided into M independent subproblems as:

$$\mathcal{P}: \max Z_m = \sum_{n \in \mathcal{N}} \sum_{f \in \mathcal{F}} p_f \left(A_{m,nf}^1 \pi_{m,nf}^L + A_{m,nf}^2 \pi_{m,nf}^E + A_{m,nf}^3 \pi_{m,nf}^C \right), \quad (15a)$$

s.t. the same with Problems (12b), (12c), (12d), (12e), and (12f).
(15b)

Obviously, we have $Z^* = \sum_{m \in \mathcal{M}} Z_m$.

Proof: For each ID m , we seek the best strategy to satisfy its request and then it can benefit the whole cache system. Therefore, Eq. (13) can be separated, i.e., the sub-decision for each ID does not affect other IDs because there is no relevance between them.

Specifically, for a content or a data request generated by the ID m , we search server and content selections \mathbf{II} layer by layer. After initialization, we first determine whether a local direct hit occurs according to the cache state. If it does not happen, we consider whether the soft cache hits occur. If neither of the above two situations occurs, request processing will be completed through cooperative BSs or the CS. This procedure is repeated until the cache is full. To reduce unnecessary searches, we use the BNB strategy. In Eq. (15), when a feasible solution is determined by using the heuristic algorithm, the value of Z_m is calculated and denoted as Z_m^ψ . Thus, Z_m^ψ will be added to the constraint as the lower bound of the target value. Any solution with $Z_m < Z_m^\psi$ can be deleted without verifying whether it meets other constraints. By continuously improving the lower bound of the target value, the constraint conditions can be improved and the amount of calculation can be reduced.

The details of the proposed method for solving the whole problem are shown in Algorithm 2. And the computation complexity of Algorithm 2 is $O(|\mathcal{M}| |\mathcal{N}| |\mathcal{K}|)$.

Algorithm 2: Cache-Friendly Hierarchical Heuristic Algorithm.

Input: $C, \mathcal{N}, \mathcal{M}, \mathcal{F}, \mathcal{K}$, content request probability $\{p_f\}$, content cache status $\{s_{n,f}\}$.

Output: $\{\mathbf{II}^*\}$.

```

1 Initialization:  $\mathbf{II}^L = 1$ ;
2 while the ID  $m = 1, 2, \dots, M$  do
3   for the BS  $n \in \mathcal{N}$  do
4     for the content  $f \in \mathcal{K}_m$  do
5       Calculate  $Z_m(\pi_{m,n,f}^L)$  by Algorithm 1;
6       Store it as  $Z_m^\psi$  in a sorted list;
7       According to the cache state  $s_{n,f}$ , update  $\pi_{m,n,f}$ ;
8       Calculate  $Z_m(\pi_{m,n,f})$ ;
9       if  $Z_m(\pi_{m,n,f}) > Z_m^\psi$  then
10        Swap and update  $\pi_{m,n,f}$ ;
11        Add  $\pi_{m,n,f}$  to  $\mathbf{II}$ ;
12      end
13    end
14  if  $\pi_{m,n,f} D_f > C$  then
```

```

15    break;
16  end
17 end
18 end
19  $\mathbf{II} \leftarrow \text{argmax} Z^*(\mathbf{II})$ ;
20  $\mathbf{II}^* \leftarrow \mathbf{II}$ .
```

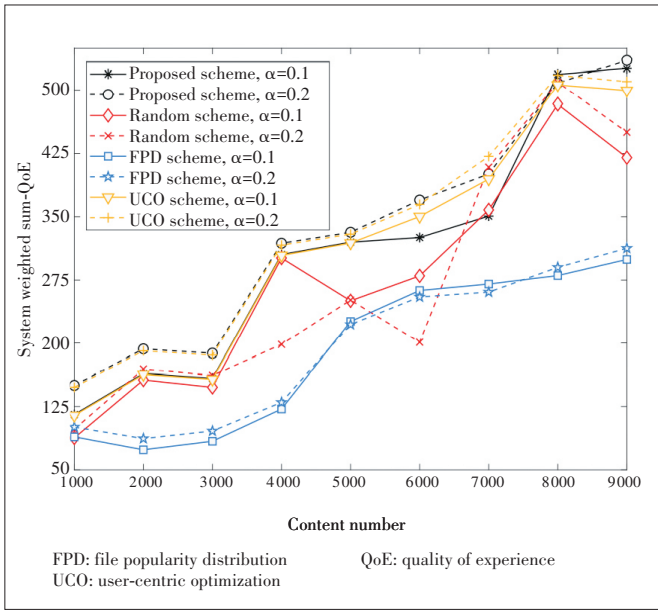
4 Simulation Results

For simulation purposes, all parameters are selected according to the real-world scenario. Numerical experiments are provided to evaluate the performance of the proposed scheme. We consider several BSs, each of which has the maximum coverage of a circle with a radius of 250 meters. And more than 400 IDs are randomly distributed within the coverage area of the BSs. We determine the local BS of each ID according to the association probability $a_{m,n}$. The channel gain is modeled as $g_{m,n} = 30.6 + 36.7 \log(l_{m,n})$ dB, where $l_{m,n}$ is the distance between the ID m and the BS n . The distance is randomly set as $[0, 250]$ m. The wireless bandwidth, transmit power of each ID, and noise power is set as 20 MHz, $[1.0, 1.5]$ W, and 10^{-13} W, respectively.

For IoT data, we consider a real data set consisting of 457 users and more than 9 000 video contents. And these contents are randomly cached in the BSs. The content size is randomly set as $[2, 5]$ Mbit. Further, the cache constraint of the BS is set to a percentage θ of the total storage size. Besides, we use itemCF-IUF to get the recommendation list for each ID, and we get the corresponding score $w_{m,f}$. The parameter of Algorithm 1 is set as $R = 2$. To verify the experimental effect of the recommendation algorithm, we calculate the accuracy rate, recall rate, and their weighted harmonic average. And the results are respectively 0.4, 0.1311, and 0.1975.

To evaluate our proposed framework, we consider the following three baseline schemes: 1) File popularity distribution (FPD) strategy. As mentioned in Ref. [21], when a content request is generated by the ID, the cache system will distribute popular contents according to the popularity of contents. However, this strategy processes requests without considering content preferences and soft caching; 2) User-centric optimization (UCO) strategy. Similar to our paper, a simple QoE metric has been proposed for combining content caching with the recommender systems in Ref. [11]. They weigh the QoS and QoR, but the work of cooperative edge caching is missing; 3) Random scheme. The content request is randomly processed at the local BS, cooperative BSs, or the CS. \mathbf{II} is randomly set under the constraints in Eqs. (12d), (12e), and (12f).

In Fig. 2, we study different server selection schemes under contents ranging from 1 000 to 9 000, and eight independent simulations are considered (in this case, we set the $N = 2$). For each scheme, we set the balance constraint α to



▲ Figure 2. QoE versus different numbers of contents

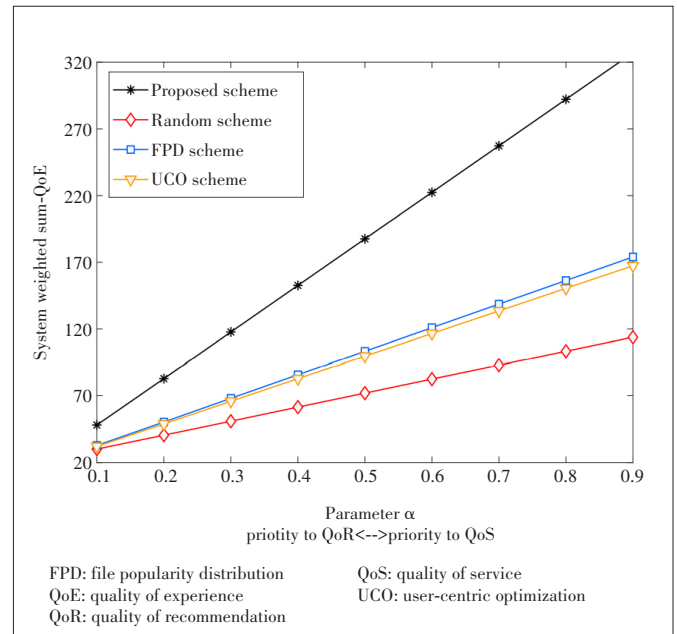
0.1 and 0.2 respectively. We observe that the QoE increases rapidly with the increase of the contents in our proposed scheme, mainly because a tremendous amount of contents can provide more accurate references for recommendation (e.g., more historical behaviors). In the random scheme, the result fluctuates obviously because the decision is random. The experimental effect of our proposed scheme is also better than other schemes. In particular, the proposed scheme has an overall performance improvement of about 30% compared with the FPD scheme. The reason is that the soft cache fully considers content preferences, ensuring that content preferences are controllable and the distortion is minimized.

Next, we investigate whether the proposed scheme has better performance in QoS-QoR trade-off, as shown in Fig. 3. The balance factor α is in the range of 0.1 to 0.9. According to the simulation, the QoE increases linearly with the increase of α . When $\alpha = 0.1$ (i.e., QoR is given priority), we observe that the performance of the FPD scheme and the UCO scheme is similar to the proposed scheme, mainly because cooperative caching has little effect on additional cache gain. When α increases gradually (i.e., a part of QoR is sacrificed and QoS is given priority), and the performance of the proposed scheme is greatly improved compared with the FPD scheme and UCO scheme. Due to the strong randomness of the random scheme, the performance improvement is not obvious.

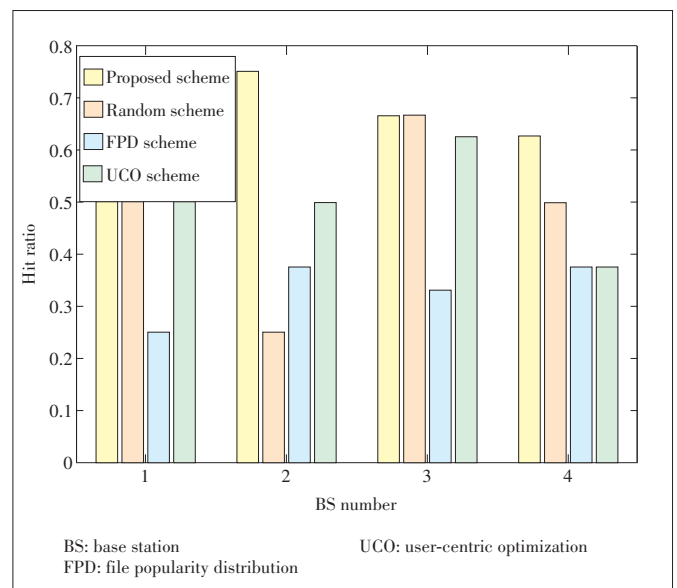
We also evaluate the hit ratio under different BS numbers, as shown in Fig. 4. In the proposed scheme, cache hits are defined as local hits and neighboring hits. We study different server selection schemes under the N range of 1 to 4. The hit ratio of the proposed scheme fluctuates depending

on the number of BSs. For instance, it achieves the best hit ratio when the BS number is 2. But when the numbers of BSs are equal to 3 and 4, the hit ratio decreases gradually, mainly because more BSs will receive more content requests. In terms of improving the hit ratio, the performance of the proposed scheme is obviously better than the other three baseline schemes, mainly because the proposed scheme provides more cache hit possibilities.

The proposed scheme considers soft caching and the cooperation between the BSs. Compared with other baseline schemes, our proposed scheme considers the content prefer-



▲ Figure 3. QoE versus different balance parameters



▲ Figure 4. Hit Ratio versus different numbers of BSs

ences of the IDs to meet their needs and the BSs' cooperation to reduce the transmission delay of contents in the networks. Therefore, our scheme is superior to other schemes in the above comparative experiments.

5 Conclusions

In this paper, we have investigated the joint problem of cooperative edge caching and recommender systems for IoT systems. We have used the concept of soft caching by shifting from satisfying requests of IDs to satisfying their needs. Under the constraints of resources, computing conditions, etc., we choose the appropriate server actions to improve the QoE, which is defined as a 0 - 1 ILP problem. To solve it, we have proposed an uncomplicated and cache-friendly hierarchical heuristic algorithm with the BNB strategy. Simulation results have revealed the superior performance of the proposed scheme on increasing the QoE.

References

- [1] MA H D, LIU L, ZHOU A F, et al. On networking of Internet of Things: explorations and challenges [J]. IEEE Internet of Things journal, 2016, 3(4): 441 - 452. DOI: 10.1109/JIOT.2015.2493082
- [2] HE Y, YU F R, ZHAO N, et al. Software-defined networks with mobile edge computing and caching for smart cities: a big data deep reinforcement learning approach [J]. IEEE communications magazine, 2017, 55(12): 31 - 37. DOI: 10.1109/MCOM.2017.1700246
- [3] GONG C, LIN F H, GONG X W, et al. Intelligent cooperative edge computing in Internet of Things [J]. IEEE Internet of Things journal, 2020, 7(10): 9372 - 9382. DOI: 10.1109/JIOT.2020.2986015
- [4] CHEN B, LIU L, SUN M X, et al. IoT cache: toward data-driven network caching for Internet of Things [J]. IEEE Internet of Things journal, 2019, 6(6): 10064 - 10076. DOI: 10.1109/JIOT.2019.2935442
- [5] ZHANG F, HAN G J, LIU L, et al. Joint optimization of cooperative edge caching and radio resource allocation in 5G-enabled massive IoT networks [J]. IEEE Internet of Things journal, 2021, (99): 1. DOI: 10.1109/JIOT.2021.3068427
- [6] LI X H, WANG X F, XIAO S J, et al. Delay performance analysis of cooperative cell caching in future mobile networks [C]//2015 IEEE International Conference on Communications. London, UK: IEEE, 2015: 5652 - 5657. DOI: 10.1109/ICC.2015.7249223
- [7] LI X H, WANG X F, WAN P J, et al. Hierarchical edge caching in device-to-device aided mobile networks: modeling, optimization, and design [J]. IEEE journal on selected areas in communications, 2018, 36(8): 1768 - 1785. DOI: 10.1109/JSAC.2018.2844658
- [8] YANG P, ZHANG N, ZHANG S, et al. Content popularity prediction towards location-aware mobile edge caching [J]. IEEE transactions on multimedia, 2019, 21(4): 915 - 929. DOI: 10.1109/TMM.2018.2870521
- [9] SHANMUGAM K, GOLREZAEI N, DIMAKIS A G, et al. FemtoCaching: wireless content delivery through distributed caching helpers [J]. IEEE transactions on information theory, 2013, 59(12): 8402 - 8413. DOI: 10.1109/TIT.2013.2281606
- [10] SERMPEZIS P, GIANNAKAS T, SPYROPOULOS T, et al. Soft cache hits: Improving performance through recommendation and delivery of related content [J]. IEEE journal on selected areas in communications, 2018, 36(6): 1300 - 1313. DOI: 10.1109/JSAC.2018.2844983
- [11] TSIGKARI D, SPYROPOULOS T. User-centric optimization of caching and recommendations in edge cache networks [C]//2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks". Cork, Ireland: IEEE, 2020: 244 - 253. DOI: 10.1109/WoWMoM49955.2020.00052
- [12] CHATZIELEFTHERIOU L E, KARALIOPOULOS M, KOUTSOPOULOS I. Caching-aware recommendations: nudging user preferences towards better caching performance [C]//IEEE Conference on Computer Communications. Atlanta, USA: IEEE, 2017: 1 - 9. DOI: 10.1109/INFOCOM.2017.8057031
- [13] SERMPEZIS P, SPYROPOULOS T, VIGNERI L, et al. Femto-caching with soft cache hits: Improving performance with related content recommendation [C]//IEEE Global Communications Conference. Singapore: IEEE, 2017: 1 - 7. DOI: 10.1109/GLOCOM.2017.8254035
- [14] LIU D, YANG C Y. A learning-based approach to joint content caching and recommendation at base stations [C]//IEEE Global Communications Conference. Abu Dhabi, United Arab Emirates: IEEE, 2018: 1 - 7. DOI: 10.1109/GLOCOM.2018.8647827
- [15] GIANNAKAS T, SERMPEZIS P, SPYROPOULOS T. Show me the cache: optimizing cache-friendly recommendations for sequential content access [C]//2018 IEEE 19th International Symposium on "A World of Wireless, Mobile and Multimedia Networks". Chania, Greece: IEEE, 2018: 14 - 22. DOI: 10.1109/WoWMoM.2018.8449731
- [16] ZHENG D S, CHEN Y Y, YIN M X, et al. Cooperative cache-aware recommendation system for multiple Internet content providers [J]. IEEE wireless communications letters, 2020, 9(12): 2112 - 2115. DOI: 10.1109/LWC.2020.3014266
- [17] COSTANTINI M, SPYROPOULOS T, GIANNAKAS T, et al. Approximation guarantees for the joint optimization of caching and recommendation [C]//IEEE International Conference on Communications. Dublin, Ireland: IEEE, 2020: 1 - 7. DOI: 10.1109/ICC40277.2020.9148740
- [18] WANG F X, WANG F, LIU J C, et al. Intelligent video caching at network edge: a multi-agent deep reinforcement learning approach [C]//IEEE Conference on Computer Communications. Toronto, Canada: IEEE, 2020: 2499 - 2508. DOI: 10.1109/INFOCOM41043.2020.9155373
- [19] SUN C, HUI L, LI X H, et al. Task offloading for end-edge-cloud orchestrated computing in mobile networks [C]//IEEE Wireless Communications and Networking Conference. Seoul, South Korea: IEEE, 2020: 1 - 6. DOI: 10.1109/WCNC45663.2020.9120496
- [20] WANG X F, WANG C Y, LI X H, et al. Federated deep reinforcement learning for Internet of Things with decentralized cooperative edge caching [J]. IEEE Internet of Things journal, 2020, 7(10): 9441 - 9455. DOI: 10.1109/JIOT.2020.2986803
- [21] SUN R J, WANG Y, CHENG N, et al. QoE-driven transmission-aware cache placement and cooperative beamforming design in cloud-RANs [J]. IEEE transactions on vehicular technology, 2020, 69(1): 636 - 650. DOI: 10.1109/TVT.2019.2952726

Biographies

HAN Suning received the bachelor's degree in software engineering from Tiangong University, China in 2020. He is currently pursuing the master's degree with the School of Big Data and Software Engineering, Chongqing University, China. His current research interests include mobile edge computing, big data and recommender system.

LI Xiuhua (lixihua1988@gmail.com) received the B.S. degree from the Honors School, Harbin Institute of Technology, China in 2011, the M.S. de-

gree from the School of Electronics and Information Engineering, Harbin Institute of Technology, in 2013, and the Ph.D. degree from the Department of Electrical and Computer Engineering, The University of British Columbia, Canada in 2018. He joined Chongqing University through One-Hundred Talents Plan of Chongqing University, China in 2019. He is currently a tenure-track Assistant Professor with the School of Big Data & Software Engineering, and the Dean of the Institute of Intelligent Network and Edge Computing at Key Laboratory of Dependable Service Computing in Cyber Physical Society, Chongqing University. His current research interests are 5G/B5G mobile Internet, mobile edge computing and caching, big data analytics, and machine learning.

SUN Chuan is a Ph.D. student with the School of Big Data & Software Engineering, Chongqing University, China. He received his B.S. degree from Wuhan University of Science and Technology, China in 2017. His current research interests include multi-access edge computing, recommender systems, and machine learning.

WANG Xiaofei is currently a professor with the Tianjin Key Laboratory of Advanced Networking, School of Computer Science and Technology, Tianjin University, China. He got his master's and doctoral degrees from Seoul

National University, South Korea in 2006 and 2013, and was a postdoctoral fellow at The University of British Columbia, Canada from 2014 to 2016. Focusing on research of social-aware cloud computing, cooperative cell caching, and mobile traffic offloading, he has authored over 100 technical papers in *IEEE JSAC*, *IEEE TWC*, *IEEE Wireless Communications*, *IEEE Communications Magazine*, *IEEE TMM*, *IEEE INFOCOM*, and *IEEE SEC-ON*. He was a recipient of the National Thousand Talents Plan (Youth) of China. He received the Scholarship for Excellent Foreign Students in the IT Field from NIPA of South Korea from 2008 to 2011, the Global Outstanding Chinese Ph.D. Student Award of the Ministry of Education of China in 2012, and the Peiyang Scholar of Tianjin University. In 2017, he received the Fred W. Ellersick Prize from the IEEE Communication Society.

Victor C. M. LEUNG is currently a Distinguished Professor of computer science and software engineering with Shenzhen University, China. He is also an Emeritus Professor of electrical and computer engineering and the Director of the Laboratory for Wireless Networks and Mobile Systems, The University of British Columbia, Canada. His research is in the broad areas of wireless networks and mobile systems. He has published widely in archival journals and refereed conference proceedings in these areas; several of his papers have won Best Paper Awards. He is a fellow of the Royal Society of Canada, Canadian Academy of Engineering, and Engineering Institute of Canada.



Cost-Effective Task Scheduling for Collaborative Cross-Edge Analytics

ZHAO Kongyang¹, GAO Bin², ZHOU Zhi¹

(1. Sun Yat-sen University, Guangzhou 510275, China;
2. National University of Singapore, Singapore 119077, Singapore)

Abstract: Collaborative cross-edge analytics is a new computing paradigm in which Internet of Things (IoT) data analytics is performed across multiple geographically dispersed edge clouds. Existing work on collaborative cross-edge analytics mostly focuses on reducing either analytics response time or wide-area network (WAN) traffic volume. In this work, we empirically demonstrate that reducing either analytics response time or network traffic volume does not necessarily minimize the WAN traffic cost, due to the price heterogeneity of WAN links. To explicitly leverage the price heterogeneity for WAN cost minimization, we propose to schedule analytic tasks based on both price and bandwidth heterogeneities. Unfortunately, the problem of WAN cost minimization underperformance constraint is shown non-deterministic polynomial (NP)-hard and thus computationally intractable for large inputs. To address this challenge, we propose price- and performance-aware geo-distributed analytics (PPGA), an efficient task scheduling heuristic that improves the cost-efficiency of IoT data analytic jobs across edge datacenters. We implement PPGA based on Apache Spark and conduct extensive experiments on Amazon EC2 to verify the efficacy of PPGA.

Keywords: collaborative cross-edge analytics; Internet of Things; task scheduling

DOI: 10.12142/ZTECOM.202102003

<https://kns.cnki.net/kcms/detail/34.1294.TN.20210607.1503.002.html>, published online June 9, 2021

Manuscript received: 2021-04-09

Citation (IEEE Format): K. Y. Zhao, B. Gao and Z. Zhou. "Cost-effective task scheduling for collaborative cross-edge analytics," *ZTE Communications*, vol. 19, no. 2, pp. 11 - 19, Jun. 2021. doi: 10.12142/ZTECOM.202102003.

1 Introduction

With the fast burgeoning as well as accelerating convergence of artificial intelligence (AI) and the Internet of Things (IoT), unprecedented prosperity of AI of Things or AI-empowered IoT applications has emerged recently. This new trend is coined as AIoT, which pushes the frontier of AI from the centralized

cloud to the ubiquitous IoT devices, paving the last mile delivery of AI capabilities. Recently, AIoT has gained mounting attention from industrial giants and boosted many intelligent applications as exemplified by intelligent personal assistants, personalized shopping recommendations, video surveillance and smart home appliances, which are significantly changing our daily life. For example, by bringing live video analytics to urban traffic management, Microsoft^[1], Baidu, and Alibaba remarkably improve commuting efficiency and safety in Bevellue (a city outside of Seattle), Beijing, and Hangzhou, respectively.

Specifically, for many emerging collaborative AIoT appli-

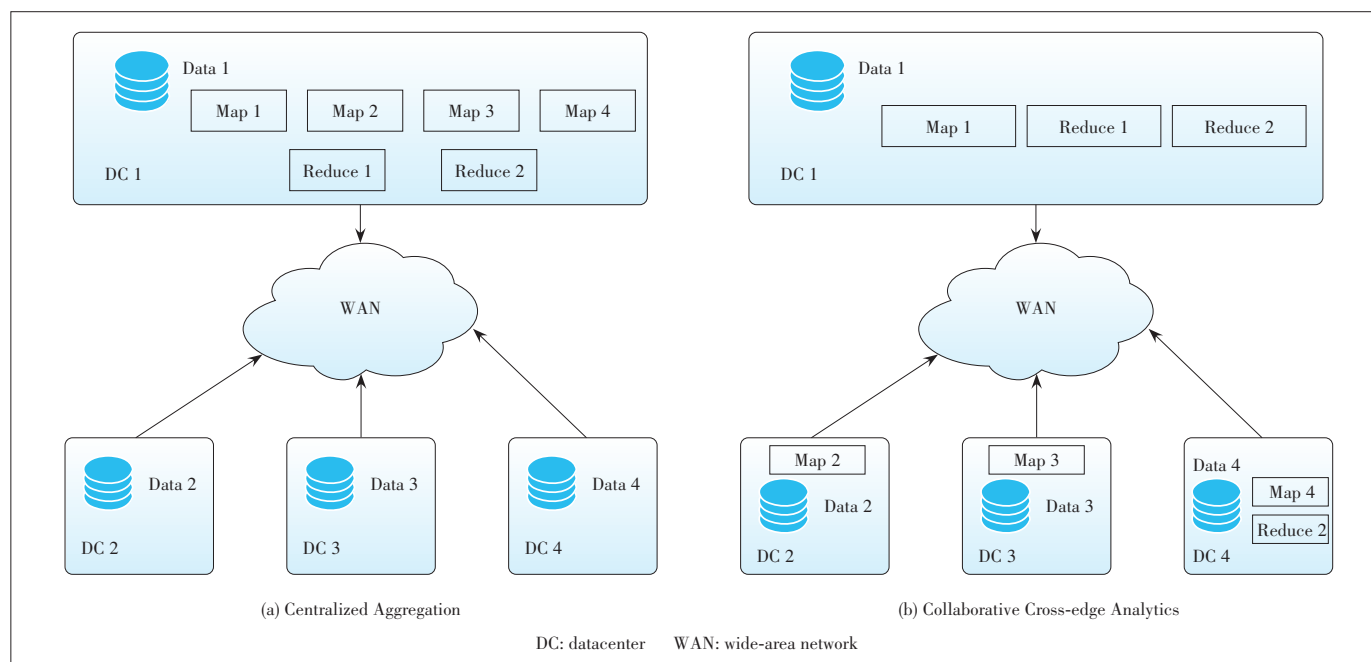
This work was supported in part by the National Natural Science Foundation of China under Grant No. 61802449 and the Guangdong Natural Science Funds under Grant No. 2021A1515011912.

cations spanning across multiple regions and edge datacenters^[2-3], the raw data generated at each datacenter can be huge, redundant and unlabeled^[4]. For example, for urban traffic control and management, the raw data born at each edge datacenter contains only the ID and timestamp of the vehicles traversing this region, but not the global trajectory of each vehicle, which can be used to learn the vehicle behavior. While for the smart retail application, the raw data of each unmanned store only contains the shopping record of each user and item information of each good, rather than the user preference and item popularity data that can be used to learn the user's shopping behavior. To extract the labeled data (e.g., vehicle trajectory, user preference and item popularity data over multiple stores) that can be used to train an AI model in such collaborative cross-region scenarios, we need to perform preprocessing to the raw data stored at cross-region edge datacenters to obtain some statistical query results (i.e., labeled data, such as the item popularity over multiple stores), employing data-parallel frameworks such as Map-Reduce, Spark and Flink.

In this paper, we propose to empower AIoT by optimizing data preprocessing across multiple edge datacenters, which is referred to as collaborative cross-edge analytics^[5-6]. With collaborative cross-edge analytics, we push the computation tasks of both the input stage (e.g., Map) and the output stage (e.g., Reduce) of the analytics framework (e.g., Map-Reduce) to the widespread edge datacenters where the data was born, instead of collecting raw data to the master datacenter as shown in Fig. 1a. Fig. 1b shows an example of collaborative cross-edge analytics, in which both Map and Reduce tasks are moved to the place where the data is stored. Since in the

output stage tasks are distributed across edge datacenters, the intermediate data after the input stage still needs to be shuffled across edge datacenters for final analytic results. Unfortunately, due to the aforementioned bandwidth heterogeneity of the cross-edge wide-area network (WAN) links, the duration of different cross-edge shuffle transfers can be greatly diverse. To cut down the finish time of the shuffle phase by mitigating the stragglers (i.e., the slowest ones), a WAN bandwidth-aware task placed for the output stage has been proposed.

Apart from the performance, another primary objective for collaborative cross-edge analytics is the WAN bandwidth cost. Compared with the sufficient intra-edge network bandwidth, cross-edge network bandwidth is far scarcer and more expensive. Motivated by the exacerbating shortage of WAN bandwidth, recent efforts have been made to minimize the amount of data that traverses the expensive WAN. For example, Pixida^[7], Geode^[8] and Iridium^[6] reduce the WAN bandwidth usage by cost-efficient task placement, i.e., placing more tasks at datacenters with more input data. However, for collaborative cross-edge analytics, is the reduction of WAN bandwidth usage really translated into cost savings? The answer is probably "No", unfortunately. The rationale is that the price of WAN bandwidth usage (in US dollar per GB, in terms of the bill for one unit of data transfer) for different WAN links also shows diversity. Currently, Infrastructure-as-a-Service (IaaS) cloud service providers such as Amazon EC2, Microsoft Azure and Google Cloud Engine charge outgoing bulk data transfer from different datacenters with various prices, as illustrated in Table 1. The price diversity clearly indicates that using less WAN bandwidth does not



▲ Figure 1. Illustration of collaborative cross-edge analytics

▼Table 1. Price heterogeneity (in US dollar per GB) of outgoing WAN bandwidth usage at different regions of Amazon EC2

Location	Price
Frankfurt	0.02
Iceland	0.02
London	0.02
Northern California	0.02
Northern Virginia	0.02
Ontario	0.02
Oregon	0.02
Tokyo	0.02
Mumbai	0.086
Seoul	0.09
Singapore	0.09
Sydney	0.09
Sao Paulo	0.16

WAN: wide-area network

necessarily mean less usage cost. Even worse, our statistical study on the Amazon AWS platform in Section 2 shows that, for datacenters connected to faster WAN links, the WAN bandwidth usage price is usually higher. Intuitively, with such a performance-cost tradeoff, existing performance driven task scheduling for collaborative cross-edge analytics would greatly increase the usage cost of WAN bandwidth.

Motivated by the price heterogeneity as well as the performance-cost tradeoff for the WAN bandwidth, we argue that price-awareness should be brought to task scheduling of collaborative cross-edge analytics. Given the non-coincidence between performance and bandwidth usage cost, we propose to minimize WAN bandwidth usage cost under performance guarantee, i. e., response time constraint. This problem is proven NP-hard, and we design PPGA, an efficient greedy-based scheduler that explicitly leverages both price and bandwidth heterogeneity, to place tasks across geo-distributed edge datacenters. PPGA is implemented based on Apache Spark and deployed across five Amazon EC2 regions. Extensive experiments using realistic benchmarks and workloads have shown that PPGA can reduce the WAN cost of collaborative cross-edge by up to 31.6%.

2 Background and Motivation

2.1 Collaborative Cross-Edge Analytics

We consider a cross-edge infrastructure, in which multiple edge datacenters at different regions are interconnected by a WAN and to host AIoT services locally. With such a setup, data are naturally born at each edge cloud in a distributed fashion, while a dataset (e.g., shopping records for a smart retail service) generally contains multiple data partitions that are originated across different edge datacenters. Traditional-

ly, to analyze such a dataset, the centralized aggregation approach is adopted as shown in Fig. 1a, i.e., all the data partitions are transferred to a master datacenter to be processed.

With collaborative cross-edge analytics^[6], a job query script is first converted into a direct acyclic graph (DAG) of consecutive stages, each stage consisting of many parallel tasks that can be executed at different DCs. For the example of the Map-Reduce query illustrated in Fig. 1, the corresponding DAG contains two stages: Map and Reduce. Tasks of the input stages (Map in Fig. 1) are pushed to the place where the data partitions are stored, so that data can be processed locally. The output of input stages, called intermediate data, is then shuffled to output stages (e.g., Reduce in Fig. 1) to compute the final query results. Since consecutive stages are linked by data dependency, a stage (e.g., Reduce) can be started only after it has received all the intermediate data from parent stages (e.g., Map). For input stage tasks, the commonly adopted approach is “site-locality”, i.e., their locations are the same as the locations of their input data. As a result of data locality and in-memory caching, input stages can be finished extremely quickly.

The scarcity of WAN bandwidth has resulted in the high usage price of WAN bandwidth. Currently, IaaS cloud providers such as Amazon AWS, Microsoft Azure and Google Cloud Engine charge the WAN bandwidth usage while leave the intra-datacenter bandwidth usage for free. Moreover, the price of outgoing WAN bandwidth usage at different datacenters also exhibits moderate diversity, due to various capital expenditure (Cap. Ex) and operational expenditure (Oper. Ex) at different regions. As illustrated in Table 1, for the case of Amazon EC2, the outgoing WAN bandwidth usage price at different datacenters varies from \$0.02 per GB to \$0.16 per GB.

2.2 Motivation

The tradeoff between the performance and the cost is shown in this section. The above example motivates us to take advantage of the heterogeneities of usage price and bandwidth of the WAN, and then further optimize the cost and performance of the shuffle phase. Clearly, to minimize the cost of WAN bandwidth usage, more Reduce tasks should be placed at datacenters with higher usage cost of out-going WAN bandwidth, so that more intermediate data would be transferred out from datacenters with lower usage price of WAN bandwidth. On the other hand, to minimize the finish time of the shuffle phase, more Reduce tasks should be placed at datacenters connected to faster inter-datacenter WAN links. Thus, if it is the idle case that datacenters with higher usage price of WAN bandwidth are connected to faster WAN links, placing more Reduce tasks at these datacenters optimizes the cost and performance simultaneously.

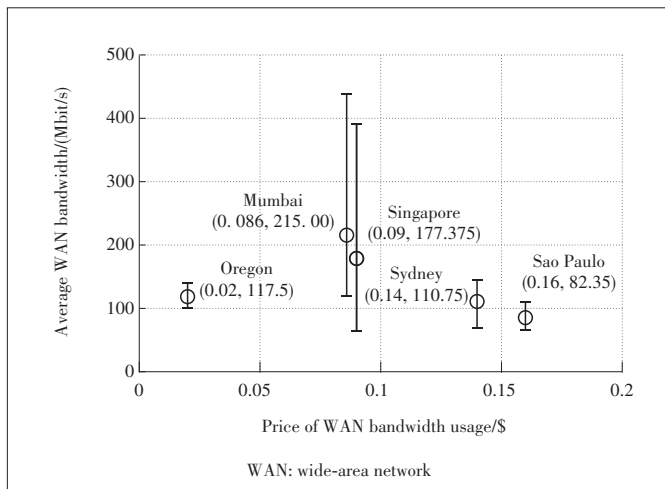
Unfortunately, however, our empirical measurements

based on Amazon EC2 show that datacenters with higher usage price of WAN bandwidth are not necessarily connected to faster WAN links. Specially, we first measure the available inter-datacenter WAN bandwidth between 5 regions of Amazon EC2: Oregon (North America), Singapore (Asia-Pacific), Sao Paulo (South America), Sydney (Oceania) and Mumbai (Asia-Pacific). We then plot the usage price of WAN bandwidth, as well as the average bandwidth of the WAN links connected to each region in Fig. 2. Interestingly, we observe that there is an approximate negative correlation between the average WAN bandwidth and the usage price. Specifically, for the 4 locations of Mumbai, Singapore, Sydney and Sao Paulo, there is a clear negative correlation between the average WAN bandwidth and the usage price. This observation convincingly demonstrates the inherent tradeoff between performance and cost of the shuffle phase.

To navigate the cost-performance tradeoff, we propose to minimize the cost of WAN bandwidth usage while enforcing the shuffle phase to be finished within a pre-defined deadline. The rationale of bounding the finish time rather than minimizing it is that, for some realistic analytic jobs, the analytic result is used by a future event or decision making with a time lag. Therefore, finishing the analytic job within the time lag would not compromise future events or decision making. Towards the goal of minimizing WAN bandwidth usage with bounded finish time of the shuffle phase, we propose PPGA, a task scheduler that leverages the heterogeneities of usage price as well as the bandwidth of the cross-edge WAN, in the next section.

3 Model and Optimization for Collaborative Cross-Edge Analytics

In this section, we present the formulation and optimization for WAN price and performance-aware task scheduling



▲ Figure 2. Usage price of outgoing WAN bandwidth versus the average outgoing WAN bandwidth of 5 Amazon EC2 regions

of collaborative cross-edge analytics.

3.1 Infrastructure

We consider an AIoT service provider running AIoT services and originate data on a set of N geographically dispersed edge datacenters, denoted by $\mathcal{D} = \{1, 2, \dots, N\}$. Here each edge datacenter can be a private datacenter, a public multi-tenant datacenter (e.g., Amazon EC2), or a public colocation datacenter (e.g., Equinix). The computational capacity of each datacenter $i \in \mathcal{D}$ is denoted as C_i , in terms of the maximal number of computational tasks that can be executed in a parallel manner.

With collaborative cross-edge analytics, a job query script is first converted into a direct acyclic graph (DAG) of consecutive stages^[9]. Then, an intuition globally schedules the stages in the DAG all together. However, such global scheduling requires prior knowledge of the task characteristics of all stages, as well as the long-term bandwidth availability. Though they are predictable, an exact prediction without error is difficult. Therefore, scheduling the DAG is not practical^[10]. Instead, we schedule tasks stage by stage in an online fashion, i.e., we choose to schedule the tasks within the same stage to geo-distributed datacenters, rather than considering all the tasks in the DAG. Note that the default task scheduler of Spark also adopts the stage-by-stage method. Though such online scheduling may not be globally optimal, it enables adjustments that can be made on the fly to better cater to the dynamic job progress and network environment.

For a given stage of the job DAG, we use $\mathcal{J} = \{1, 2, \dots, M\}$ to denote the set of M Reduce tasks which can be scheduled to different edge datacenters and executed in parallel. To denote the placement scheduling of those Reduce tasks, we introduce binary variables $x_{ij} \in \{0, 1\}$, $\forall i \in \mathcal{D} = \{1, 2, \dots, N\}$, $\forall j \in \mathcal{J} = \{1, 2, \dots, M\}$. Specifically, if the task $j \in \mathcal{J}$ is allocated to edge datacenter $i \in \mathcal{D}$, then $x_{ij} = 1$, otherwise $x_{ij} = 0$. Since each task $j \in \mathcal{J}$ can be scheduled to one and only one edge datacenter, we have the following placement constraint:

$$\sum_{i \in \mathcal{D}} x_{ij} = 1, \forall j \in \mathcal{J}. \quad (1)$$

Besides the above placement constraint, the task scheduling is also constrained by the computational capacity of each edge datacenter. That is, the number of Reduce tasks allocated to each edge datacenter i cannot exceed the computational capacity C_i of datacenter i :

$$\sum_{j \in \mathcal{J}} x_{ij} \leq C_i, \forall i \in \mathcal{D}. \quad (2)$$

For Reduce task $j \in \mathcal{J}$, it has to gather intermediate data from other edge datacenters to reduce operation. Here we use S_{ij} to denote the amount of intermediate data stored at data-

center i and to be collected by Reduce task j . For the shuffle phase where intermediate data are transferred across edge datacenters and given the source edge datacenter i and destination edge datacenter k , the amount of intermediate data transferred from edge datacenter i to edge datacenter k can be denoted as $\sum_{j \in \mathcal{D}} S_{ij} x_{kj}$.

3.2 Performance of Shuffle Phase

Given the available pair-wise bandwidth of the WAN link from edge datacenter i to edge datacenter k , B_{ik} , and together with the amount of intermediate data traversing this WAN link, the duration of the corresponding intermediate data transfer can be denoted as $\frac{\sum_{j \in \mathcal{D}} S_{ij} x_{kj}}{B_{ik}}$. Since the performance, in terms of the finish time of the intermediate data shuffle phase, is determined by the slowest intermediate data transfer, it can be formulated as:

$$z = \max_{i \in \mathcal{D}} \max_{k \in \mathcal{D}} \frac{\sum_{j \in \mathcal{D}} S_{ij} x_{kj}}{B_{ik}}. \quad (3)$$

Here if $i = k$, it means that the corresponding intermediate data transfer is an intra-edge transfer rather than a cross-edge WAN transfer. Since the intra-edge network bandwidth typically far more overweighs the cross-edge WAN bandwidth, the former can be finished very soon.

If we enforce the shuffle phase to be finished before the deadline W , each intermediate data transferred from edge datacenter i to edge datacenter k should be finished within this deadline W , that is:

$$\frac{\sum_{j \in \mathcal{D}} S_{ij} x_{kj}}{B_{ik}} \leq W, \forall i, k \in \mathcal{D}. \quad (4)$$

3.3 Cost of WAN Bandwidth Usage

Unlike the moderately sufficient intra-edge bandwidth, the cross-edge WAN bandwidth represents a scarce resource that incurs high capital and operational expenditure. For this reason, internet service providers (ISP) and IaaS edge cloud providers typically charge WAN bandwidth usage according to the bytes transferred, i.e., the amount of data transferred across the WAN. Specifically, for IaaS edge cloud providers such as Amazon AWS, Microsoft Azure and Google Cloud Engine (GCE), the price for inter-datacenter WAN transfer is dependent on the source datacenter of the transfer, and different source datacenters usually have various prices. Here we use P_i to denote the WAN bandwidth usage price for traffic going out of datacenter i , and P_i exhibits geographical diversity across datacenters. Given the amount of $\sum_{j \in \mathcal{D}} S_{ij} x_{kj}$ of intermediate data transferred from edge datacenter i to edge

datacenter k , the WAN bandwidth usage cost is given by $P_i \sum_{j \in \mathcal{D}} S_{ij} x_{kj}$. Considering all the inter-datacenter WAN transfers, the total cost of WAN bandwidth usage can be computed by:

$$\sum_{i \in \mathcal{D}} \sum_{k \in \mathcal{D}, k \neq i} P_i \sum_{j \in \mathcal{D}} S_{ij} x_{kj}. \quad (5)$$

3.4 Problem Formulation

The objective of optimizing the task scheduling of collaborative cross-edge analytics is twofold: shortening the finish time of the shuffle phase and reducing the cost of WAN bandwidth usage. Here comes the question that whether the above two objectives are coincident. To answer this question, we first look at the performance formulation and cost formulation. Intuitively, to reduce the finish time of the shuffle phase, we should place more Reduce tasks at edge datacenters connected to higher bandwidth (i.e., B_{ik}) WAN links. On the other hand, to minimize the cost of WAN bandwidth usage, we should place more Reduce tasks at datacenters with lower bandwidth usage prices (i.e., P_i). Unfortunately, as we have empirically demonstrated in Section 2, the bandwidth usage price at the datacenter with faster WAN links is typically higher and thus the performance goal is consequently not aligned to the cost goal.

To address the challenge of contradictory performance and cost objectives, we propose to minimize the cost of WAN bandwidth usage while enforcing the shuffle phase to be finished within a pre-defined deadline. Formally, such a performance-cost tradeoff problem can be formulated as the following integer programming (IP):

$$\begin{aligned} \min \quad & \sum_{i \in \mathcal{D}} \sum_{k \in \mathcal{D}, k \neq i} \sum_{j \in \mathcal{D}} P_i S_{ij} x_{kj} \\ \text{s.t.} \quad & \sum_{k \in \mathcal{D}} x_{kj} = 1, \forall j \in \mathcal{J} \\ & \sum_{j \in \mathcal{J}} x_{kj} \leq C_k, \forall k \in \mathcal{D} \\ & \frac{\sum_{j \in \mathcal{D}} S_{ij} x_{kj}}{B_{ik}} \leq W, \forall i, k \in \mathcal{D} \\ & x_{kj} \in \{0, 1\}, \forall k \in \mathcal{D}, j \in \mathcal{J}. \end{aligned} \quad (6)$$

Remark: A widely adopted alternative approach to navigating the performance-cost tradeoff is to transform the finish time into monetary cost, and then minimize the total cost. It is however nontrivial to precisely map the finish time of the shuffle phase to economic cost. In contrast, our model is

more amenable to practical implementation, since based on the stringency of the job query it is ready to set a reasonable deadline W to ensure moderate performance.

Theorem 1: The cost-performance tradeoff problem is NP-hard.

Proof: We construct a polynomial-time reduction from the cost-performance tradeoff problem to the Generalized Assignment Problem (GAP), a classic combinatorial optimization problem which is proven NP-hard:

$$\begin{aligned}
 & \min \sum_{i=1}^m \sum_{j=1}^n c_{ij} x_{ij} \\
 & \text{s.t. } \sum_{i=1}^m x_{ij} = 1, \forall j = 1, \dots, n \\
 & \sum_{j=1}^n w_{ij} x_{ij} \leq t_i, \forall i = 1, \dots, m \\
 & x_{ij} \in \{0, 1\}, \forall i = 1, \dots, m, j = 1, \dots, n.
 \end{aligned} \tag{7}$$

Given an instance $A = (m, n, c_{ij}, w_{ij}, t_i)$ of the GAP, we map it to an instance of the cost-performance tradeoff problem with $A' = \left(|\mathcal{D}| = m, |\mathcal{J}| = n, c_{ij} = \sum_{k \neq i, k \in \mathcal{D}} P_k S_{kj}, w_{ij} = 1, t_i = C_i, W = +\infty \right)$. Clearly, the above mapping can be done in polynomial time. Then, if there exists an algorithm that solves the cost-performance tradeoff problem A' , it solves the corresponding GAP A as well. As a result, the GAP can be treated as a special case of the cost-performance tradeoff problem. Given the NP-hardness of GAP, the cost-performance tradeoff problem must be NP-hard as well.

Theorem 1 reveals that solving the cost-performance tradeoff problem is NP-hard and it is computationally infeasible for large input. Thus, we propose to develop a heuristic that seeks a good approximate solution to the cost-performance tradeoff problem.

3.5 Greedy-Based Heuristic for Task Scheduling

Before deriving the heuristic for task scheduling, we first rewrite the objective function of the cost-performance tradeoff problem as follows:

$$\begin{aligned}
 \sum_{i \in \mathcal{D}} \sum_{k \in \mathcal{D}, k \neq i} \sum_{j \in \mathcal{J}} P_i S_{ij} x_{kj} &= \sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{J}} P_i S_{ij} \sum_{k \in \mathcal{D}, k \neq i} x_{kj} = \\
 & \sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{J}} P_i S_{ij} (1 - x_{ij}).
 \end{aligned} \tag{8}$$

The above equation indicates that, to minimize the cost of WAN bandwidth usage, we need to maximize the term

$\sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{J}} P_i S_{ij} (1 - x_{ij})$. Intuitively then, task $j \in \mathcal{D}$ should be scheduled to edge datacenters with a larger $P_i S_{ij}$. With this insight, we first list the tasks in \mathcal{D} in a non-increasing order of the value $\sum_{j \in \mathcal{J}} P_i S_{ij}$, and then the next task on the list is scheduled to the available datacenter with the largest $P_i S_{ij}$ without exceeding the computing capacity or compromising the performance goal. The detail is given in Algorithm 1.

Algorithm 1: Greedy-based heuristic for task scheduling

Input:

Edge datacenter capacity, C_k ;
WAN bandwidth, B_{ik} ;
Price of WAN bandwidth, P_k ;
Deadline, W ;
Amount of intermediate data, S_{kj} ;

Output:

Task placement, x_{kj} ;

1: list the tasks in \mathcal{D} in a non-increasing order of the value $\sum_{k \in \mathcal{D}} P_k S_{kj}$;
2: Define $u_{kj} = \max_{i \in \mathcal{D}} \frac{S_{ij}}{B_{ik}}, \mathcal{M}_k = \emptyset, \forall k, j$;
3: **for** $j = 1, 2, \dots, M$ **do**
4: $k^* = \operatorname{argmin}_{k \in \mathcal{D}} \left\{ P_k S_{kj} \mid \sum_{l \in \mathcal{M}_i} x_{kl} \leq C_k - 1, \sum_{l \in \mathcal{M}_i} x_{kl} u_{kl} \leq W - u_{kj} \right\}$
5: $x_{k^*j} = 1, \mathcal{M}_{k^*} = \mathcal{M}_{k^*} \cup \{j\}$
6: **end for**
7: **return** x_{kj} ;

4 Implementation and Performance Evaluation

4.1 Implementation

The implementation of PPGA is on top of the Apache Spark framework. We override Spark's default scheduler and build our existing solution together with the default scheduler. When a taskset is submitted, we choose the scheduling method according to the dependency type of the taskset. When the taskset has shuffle dependency, we try to use PPGA to optimize the task placement. If PPGA is not suitable for the taskset, we choose default scheduler to finish task scheduling.

4.2 Experimental Setup

1) Experimental platform: Our experiment cluster is deployed on 5 datacenters with 10 instances. The 5 datacenters we choose are Singapore, Mumbai, Sao Paulo, Oregon, and Sydney. All instances used in our experiment are M4.4 XLARGE, which contains 16 vCPUs and 64 GB memory. We choose the instance at the Singapore region as Spark's master

and Hadoop Distributed File System's (HDFS) name node. The pair-wise bandwidth between the different EC2 regions is shown in Table 2. The price of outgoing WAN bandwidth usage across geo-distributed datacenters is shown in Table 3.

2) Software Settings: Our instances are running on CentOS 7. The Spark version number of our implementation system is 2.1. We use HDFS from Apache Hadoop 2.7.1 as our distributed file system and start all instances as data nodes and worker nodes. The HDFS's block size that we use is 128 MB. The replication of HDFS is 3. Spark runs in stand-alone mode and does not require additional resource management intervention.

3) Workload Specifications: To evaluate the effectiveness of our system to various kinds of computation tasks, we measure the performance metrics of three applications: WordCount, PageRank^[11], and TeraSort^[12]. Computation tasks can be divided into two categories (i.e., computation intensive and I/O intensive). In our benchmarks, WordCount is computation intensive while PageRank and TeraSort are I/O intensive.

- WordCount: WordCount aims to calculate the number of every single word in passages. WordCount first produces map tasks to calculate the frequency of words in every partition. Then Reduce tasks will collect the results of map tasks to get the final result. This application represents a typical data processing job. We use 11 GB data from Wikipedia as the input file.

- PageRank: PageRank computes the weights of the website using the amount and quality of links. It is a fundamental data processing application. It is used to calculate PageRank for a website. Our experiment uses a dataset that has 1 632 803 nodes and 30 622 564 edges.

- TeraSort: TeraSort is also a benchmark that measures computing ability of the big data framework. It is used to sort the sequences of results in the distributed situation. We generate 1 GB raw data to TeraSort for measurement.

4) Baseline: We compare our scheduler with the situation setting the price factor of each node to 1. According to the objective function of our problem, setting the price to 1 actually minimizes the amount of bandwidth usage. In the case of deadline insensitivity, we will contrast the different performances between minimizing the cost of WAN Bandwidth usage and minimizing WAN bandwidth usage.

4.3 Evaluation Results

1) Cost of WAN bandwidth usage: The primary performance metric is the cost of WAN Bandwidth usage. As we can see in Fig. 3, PPGA reduces the cost of WAN Bandwidth usage of WordCount, PageRank, and TeraSort by 30.26%, 25.82% and 31.60%, respectively. There are two reasons behind the cost reduction. Firstly, PPGA mainly considers minimizing the cost of WAN Bandwidth usage of an application rather than the volume of the WAN band-

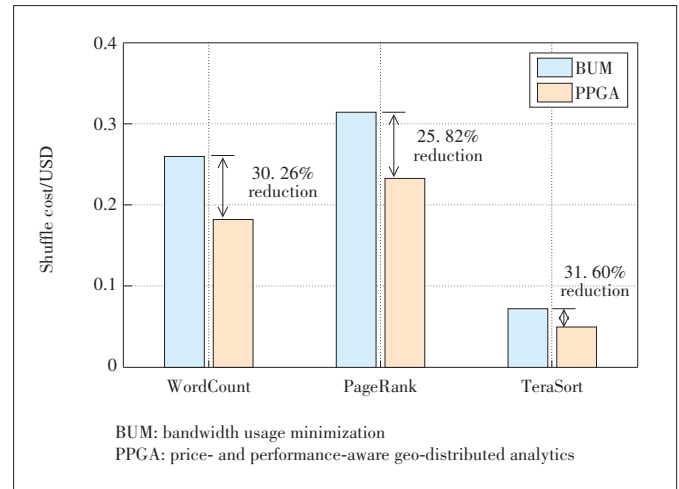
▼Table 2. Pair-wise bandwidth (in Mbit/s) between 5 different EC2 regions

City	Singapore	Oregon	Sydney	Sao Paulo	Mumbai
Singapore	46 028.8	116	140	63.5	390
Oregon	123	46 284.8	137	110	100
Sydney	144	124	46 592	69.0	106
Sao Paulo	66.6	109	74.5	46 899.2	79.3
Mumbai	390	103	106	73.7	46 694.4

▼Table 3. Price (in US dollar per GB) of outgoing WAN bandwidth usage

City	Singapore	Oregon	Sydney	Sao Paulo	Mumbai
Price	0.09	0.02	0.14	0.16	0.086

WAN: wide-area network

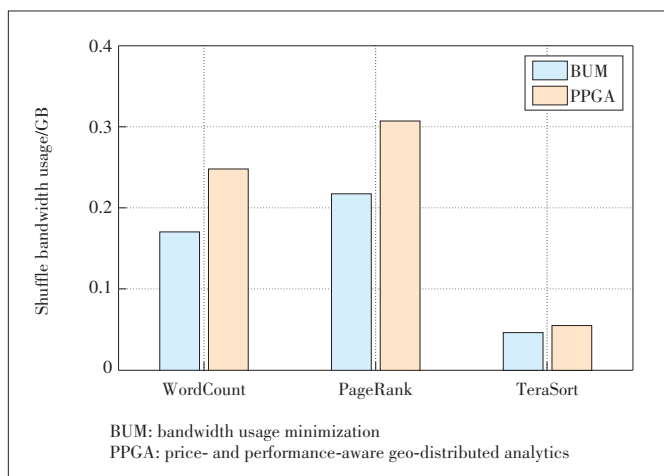


▲Figure 3. Data transfer costs of WordCount, PageRank and TeraSort

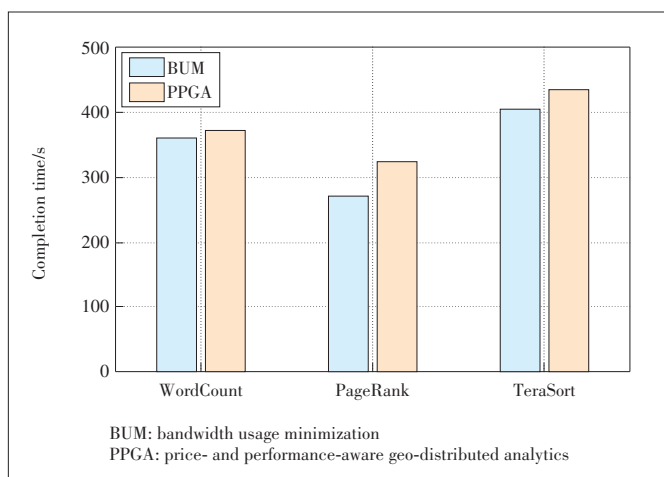
width usage. When the price is set to 1, the target of PPGA transforms to minimize the bandwidth usage. The second reason is that PPGA entirely considers all reduce tasks in a taskset. When a taskset is submitted to the schedule, PPGA makes a scheduling plan by analyzing overall tasks' distribution, data transferred price and bandwidth speed factors. So, we can get a smaller cost to complete the whole taskset. Further, the cost of WAN bandwidth usage of the whole application will be reduced.

2) Volume of data transferred across datacenters: Fig. 4 shows the results of bandwidth usage. It is clear that the bandwidth usage minimization task scheduling incurs less bandwidth usage. This corresponds to our model in which we do not consider the price. At the same time, the results suggest that PPGA has a larger bandwidth usage, but it has a smaller data transfer cost, which also confirms our assumptions in Section 2.2.

3) Application completion times: Fig. 5 demonstrates that all applications' completion time with different prices is increased relative to the price situation (set to 1). This can be seen as a tradeoff between bandwidth usage and the cost of bandwidth usage. The price set to 1 just minimizes the bandwidth usage. When PPGA considers the price and the



▲ Figure 4. WordCount, PageRank and TeraSort's shuffle bandwidth usage across datacenters



▲ Figure 5. Application completion times of WordCount, PageRank and TeraSort

bandwidth, a task may be scheduled to a datacenter with a smaller cost, which needs more data transfer time. That is why applications running with PPGA have longer completion time.

5 Related Work

Collaborative cross-edge analytics, also known as geo-distributed analytics, has received great attention recently, due to the unprecedented increase in data volume. Most of the existing work focuses on optimizing a single objective of either query response time or WAN bandwidth usage. For reducing the usage of expensive WAN bandwidth, Pixida^[7] works on dividing the DAG graph of a job into several parts to be processed in a datacenter. JetStream^[13], a stream processing for data structured as online analytical processing (OLAP) cubes, relies entirely on aggregation and approximation to reduce bandwidth. However, JetStream does not optimize data and task placement. Geode^[8] jointly catches intermediate re-

sults and optimizes task placement for Structured Query Language (SQL) queries. It optimizes WAN bandwidth usage but may lead to a poor performance. Flutter^[10] carefully orchestrates task placement by exploiting bandwidth heterogeneity of the WAN. Note that none of the above work considers the cost of WAN bandwidth usage. On the other hand, Iridium^[6] and the most recent work^[5] jointly optimize the task and input data placement to reduce both response time and WAN traffic. Our work is inherently different from them in at least three important aspects. First, we leverage the heterogeneities of both the usage price and the bandwidth of the WAN, while Iridium and the work in Ref. [5] only consider the bandwidth diversity of the WAN. Second, rather than assuming that the network bottleneck exists in the up/down links of edge sites, we assume the cross-edge links as the bottleneck. Third, Iridium places tasks via solving the NP-hard problem with solvers like Gurobi, while we apply a simple heuristic which has better computational efficiency and scalability.

6 Conclusions

In this paper, we study the task scheduling problem for collaborative cross-edge analytics to jointly optimize cost and performance. We first demonstrate that, the commonly adopted approach of WAN bandwidth usage does not necessarily minimize the cost of WAN bandwidth usage, due to the price heterogeneity of WAN bandwidth usage. To fully explicit the price heterogeneity, we propose PPGA, a price and performance-aware task scheduler for collaborative cross-edge analytics. Unfortunately, the problem of WAN cost minimization under performance constraint is shown to be NP-hard, and thus computationally intractable for large inputs. To address this challenge, we propose an efficient greedy-based heuristic to improve the cost-efficiency of collaborative cross-edge analytics. The implementation of PPGA is based on Apache Spark, and extensive experiments across 5 Amazon EC2 regions demonstrate the cost-efficiency of PPGA.

Reference

- [1] ANANTHANARAYANAN G, BAHL P, BODÍK P, et al. Real-time video analytics: the killer App for edge computing [J]. Computer, 2017, 50(10): 58 - 67. DOI: 10.1109/MC.2017.3641638
- [2] JAIN S, ZHANG X, ZHOU Y H, et al. Spatula: Efficient cross-camera video analytics on large camera networks [C]//2020 IEEE/ACM Symposium on Edge Computing (SEC). San Jose, USA: IEEE, 2020: 110 - 124. DOI: 10.1109/SEC50012.2020.00016

- [3] JIANG J C, ANANTHANARAYANAN G, BODIK P, et al. Chameleon: scalable adaptation of video analytics [C]//2018 Conference of the ACM Special Interest Group on Data Communication. Budapest, Hungary: ACM, 2018: 253 – 266. DOI: 10.1145/3230543.3230574
- [4] ZHOU Z, CHEN X, LI E, et al. Edge intelligence: paving the last mile of artificial intelligence with edge computing [J]. *Proceedings of the IEEE*, 2019, 107(8): 1738 – 1762. DOI: 10.1109/JPROC.2019.2918951
- [5] JIN H, JIA L, ZHOU Z. Boosting edge intelligence with collaborative cross-edge analytics [J]. *IEEE Internet of Things journal*, 2021, 8(4): 2444 – 2458. DOI: 10.1109/JIOT.2020.3034891
- [6] PU Q F, ANANTHANARAYANAN G, BODIK P, et al. Low latency geo-distributed data analytics [C]//Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication. London, United Kingdom: ACM, 2015: 421 – 434. DOI: 10.1145/2785956.2787505
- [7] KLOUDAS K, MAMEDE M, PREGUIÇA N, et al. Pixida [J]. *Proceedings of the VLDB endowment*, 2015, 9(2): 72 – 83. DOI: 10.14778/2850578.2850582
- [8] VULIMIRI A, CURINO C, GODFREY PB, et al. Global analytics in the face of bandwidth and regulatory constraints [C]//The 12th USENIX Symposium on Networked Systems Design and Implementation (NSDI). Oakland, USA: USENIX, 2015:323 – 336. DOI: 10.1109/TST.2016.7442496
- [9] MAO H Z, SCHWARZKOPF M, VENKATAKRISHNAN S B, et al. Learning scheduling algorithms for data processing clusters [C]//The ACM Special Interest Group on Data Communication. Beijing, China: ACM, 2019: 270 – 288. DOI: 10.1145/3341302.3342080
- [10] HU Z M, LI B C, LUO J. Flutter: Scheduling tasks closer to data across geo-distributed datacenters [C]//The 35th Annual IEEE International Conference on Computer Communications. San Francisco, USA: IEEE, 2016: 1 – 9. DOI: 10.1109/INFOCOM.2016.7524469
- [11] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: bringing order to the Web. Technical report [R]. 1919
- [12] O' MALLEY O. Terabyte sort on apache Hadoop [EB/OL]. [2021-01-04]. <http://sortbenchmark.org/YahooHadoop.pdf>
- [13] RABKIN A, ARYE M, SEN S, et al. Aggregation and degradation in Jet-Stream: streaming analytics in the wide area [C]//Usenix Conference on Net-

worked Systems Design & Implementation. Seattle, USA: USENIX Association, 2014

Biographies

ZHAO Kongyang received the B.E. degree from the South China University of Technology, China in 2020. He is currently pursuing his master's degree in Sun Yat-sen University, China. His research interests include edge computing, edge intelligence, and serverless computing.

GAO Bin is now a research assistant of School of Computing in National University of Singapore (NUS). Before this, he received the master's degree and bachelor's degree from Huazhong University of Science and Technology (HUST), China in 2017 and 2020, respectively. His research interests include operation system, mobile edge computing, cloud computing, and geo-distributed data analytics.

ZHOU Zhi (zhouzhi9@mail.sysu.edu.cn) received the B.S., M.E., and Ph.D. degrees in 2012, 2014, and 2017, respectively, all from the School of Computer Science and Technology at Huazhong University of Science and Technology (HUST), China. He is currently an associate professor in the School of Computer Science and Engineering at Sun Yat-sen University, China. In 2016, he was a visiting scholar at University of Goettingen, Germany. He was nominated for the 2019 CCF Outstanding Doctoral Dissertation Award, the sole recipient of the 2018 ACM Wuhan & Hubei Computer Society Doctoral Dissertation Award, and a recipient of the Best Paper Award of IEEE UIC 2018. His research interests include edge computing, cloud computing, and distributed systems.

BPPF: Bilateral Privacy-Preserving Framework for Mobile Crowdsensing



LIU Junyu, YANG Yongjian, WANG En

(Jilin University, Changchun 130012, China)

Abstract: With the emergence of mobile crowdsensing (MCS), merchants can use their mobile devices to collect data that customers are interested in. Now there are many mobile crowdsensing platforms in the market, such as Gigwalk, Uber and Checkpoint, which publish and select the right workers to complete the task of some specific locations (for example, taking photos to collect the price of goods in a shopping mall). In mobile crowdsensing, in order to select the right workers, the platform needs the actual location information of workers and tasks, which poses a risk to the location privacy of workers and tasks. In this paper, we study privacy protection in MCS. The main challenge is to assign the most suitable worker to a task without knowing the task and the actual location of the worker. We propose a bilateral privacy protection framework based on matrix multiplication, which can protect the location privacy between the task and the worker, and keep their relative distance unchanged.

Keywords: mobile crowdsensing; task allocation; privacy preserving

DOI: 10.12142/ZTECOM.202102004

<https://kns.cnki.net/kcms/detail/34.1294.TN.20210517.1615.002.html>, published online May 17, 2021

Manuscript received: 2021-03-11

Citation (IEEE Format): J. Y. Liu, Y. J. Yang, and E. Wang. "BPPF: bilateral privacy-preserving framework for mobile crowdsensing," *ZTE Communications*, vol. 19, no. 2, pp. 20 – 28, Jun. 2021. doi: 10.12142/ZTECOM.202102004.

1 Introduction

Mobile crowdsensing (MCS), as a critical component of the Internet of Things (IoT)^[1], relies on various sensors in mobile devices to collect and transmit data through a wireless network. Nowadays, mobile devices are essential for our daily activities, including businesses, communications, and entertainments. According to Gartner statistics, the number of worldwide smartphones sales in 2018 was 1.55 billion. Anyone with a smartphone can become a participant in the MCS system with wide coverage, which has gained popularity in recent years and become an appealing paradigm for sensing and collecting data^[2].

In the traditional mobile crowdsensing, there are three entities: the crowdsensing platform, the worker, and the task requester. The task requester has some tasks to be completed in

some places and is willing to pay for the task, the worker is the person who registers on the mobile crowdsensing platform to get the reward for completing the task, and the crowdsensing platform provides services for the task requester. The purpose of the platform is to recruit suitable workers for the task requester to complete the task under the condition of the minimum compensation. The distance between workers and tasks is highly related to the cost of workers who complete tasks. Most of the existing recruitment mechanisms take the distance between tasks and workers as the cost of workers completing tasks. Therefore, the total distance between workers and tasks is taken as the optimization objective of the algorithm in hope of minimizing the total cost of workers.

In order to recruit suitable workers for the task, the platform needs the location information of the task and workers. If us-

ers upload their actual location, it will bring danger to the user's privacy and eliminate the user's enthusiasm to participate in mobile crowdsensing activities. The invasion of the privacy of workers may occur when a worker uploads his or her actual location information on the platform, which may infer the worker's identity and preference based on the worker's location information, and disclose the privacy information to a third party for profit. On the other hand, to select the right worker, the task requester publishes the task information on the platform, and the platform and all workers can see the information, which poses a risk to the privacy of the task requester, because in certain cases the location and content of the task needs to be protected for the requester. For example, people with health problems at home can seek help through mobile crowdsensing, but publishing her health problems and her home address clearly violates her privacy.

Therefore, many privacy protection methods have been proposed to deal with the problem of privacy leakage in mobile crowdsensing, such as anonymity, obfuscation and encryption^[3]. By adding noise to the user's actual location and confusing it with other locations^[4], the secret sharing technology is used to encrypt the privacy information^[5], and k-anonymity is used to cluster a group of users to protect the user's privacy information^[6]. However, most of the existing privacy protection work mainly focuses on the privacy protection of workers with little consideration of task privacy protection. Some works^[7] consider the privacy of both workers and tasks, but they need an online trusted third party (TTP) to assist in the task allocation phase, which undoubtedly leads to high communication overhead and unnecessary delay in the task allocation phase.

Inspired by the privacy protection problem, we focus on designing a bilateral privacy protection framework, which chooses the right worker to minimize the total moving distance while protecting the location information of workers and tasks without an online TTP in the task allocation phase. Specifically, the whole geographic space is divided into several sub-regions. The task requester maps the location of the task to the sub-region, uses a matrix to transfer the location information, then hides the actual location matrix, and uploads the confusing location matrix. Similarly, workers also map their positions to their sub-regions, use a matrix to transfer the position information, hide the actual location matrix, and upload the confusing location matrix. After receiving the confusing location matrix from the task requester and the worker, the platform uses the confusing location information instead of the actual location information to select the appropriate worker.

Obviously, if the location of the task and the location of the worker are both confusing, it is difficult for the platform to choose the right worker. Therefore, the first challenge of the bilateral privacy protection framework is to solve this problem: how to protect the location information of tasks and workers while keeping the relative distance constant. In this regard, we design a novel privacy protection method based on matrix mul-

tiplication, with which we can protect not only the location privacy of workers, but also the location privacy of tasks. At the same time, we can ensure that the platform can accurately measure the relative distance between tasks and workers through the confusing location matrix, so that the platform can select the right workers to minimize the total moving distance.

The contributions of our work are summarized as follows:

- 1) We propose a privacy protection framework to protect the location privacy of tasks and workers without an online TTP in the task allocation phase.
- 2) We design a novel location privacy protection method based on matrix multiplication, which can protect the location privacy of tasks and workers at the same time and retain the relative distance information after confusing the location, so that the platform can quantify the relative distance between workers and tasks through the confused location.
- 3) We have done a lot of experiments on real datasets to verify our proposed method. The experimental results show that our method outperforms the state-of-the-art method.

2 Related Work

In this section, we briefly introduce the related work of task allocation optimization and privacy protection.

2.1 Task Allocation Optimization

Task allocation aims to allocate appropriate workers to tasks based on some optimization objectives, such as the cost of recruiting tasks on the platform, the task coverage, and the quality of task completion. In Ref. [8], the authors take the number of recruited workers as the optimization objective in the task allocation stage. The fewer the number of workers, the less the recruitment cost. The platform focuses on selecting a minimum number of workers and ensures that tasks can be covered. In Ref. [9], the authors propose a task coverage oriented assignment method based on the worker analysis and worker attribute model. They propose to migrate certain qualified workers to the less popular tasks for increasing the task coverage, and meanwhile optimize other performance factors. In Ref. [10], the authors take the quality of task completion as the optimization objective of task allocation.

In the location-based mobile crowdsensing task allocation, the recruited workers often need to move from their current location to the task location to complete the data collection task. Therefore, in this kind of mobile sensing, the cost of workers completing the task is closely related to the workers' moving distance. Therefore, in order to minimize the recruitment cost, the platform often takes minimizing the total moving distance as the priority of task allocation. In Ref. [11], the authors propose a tabu search algorithm to select workers, considering various task requirements (sensors, location accessibility and reliability) and work specifications (available sensor set and speed) as well as task completion sequence, and the total mov-

ing distance of all workers is minimized. In Ref. [12], the authors study two kinds of multi-task assignment schemes which make the number of completed tasks the most and the total moving distance the shortest. In Ref. [13], the authors propose ActiveCrowd, a worker selection framework for multitask MCS environments under two situations, for time-sensitive tasks. Workers are required to move to the task venue intentionally, and the goal is to minimize the total distance moved. For delay-tolerant tasks, the goal is to minimize the total number of workers. None of the above work has considered the issue of privacy protection.

2.2 Privacy-Preserving Task allocation

The problem of privacy protection has attracted more and more attention. Until now, many privacy protection technologies have been used in MCS, such as anonymization, obfuscation, encryption, and authentication^[3], for the location-based mobile crowdsensing. In Ref. [14], the authors formulate a mixed-integer nonlinear programming problem to minimize the expected travel distance of the selected workers under the constraints of differential and distortion privacy. WANG et al.^[15] propose a probabilistic winner selection mechanism (PWSM) to minimize the total travel distance with the obfuscated information from workers, by allocating each task to the worker most likely to be closest to it. The location privacy of workers is protected by adding Laplace noise to the distance between tasks and workers. WANG et al.^[16] propose a location aggregation method, which groups users into a group to realize k -anonymity. However, all the above work only considers the location privacy of workers, while ignores the location privacy of tasks. Recently, there have been some works that consider both tasks' and workers' location privacy. For example, NI et al.^[17] propose SPOON, a strong privacy-preserving mobile crowdsensing scheme supporting accurate task allocation based on geographic information and credit points of mobile users. Although this scheme protects the location privacy of workers and tasks, it can only determine whether the tasks and workers are in the same grid. In general, if there is no worker in the grid where a task is located, the platform needs to recruit workers from the nearby grid. This scheme is not suitable for this situation. YUAN et al.^[18] devise a grid-based location protection method, which can protect the locations of workers and tasks. However, this method requires the task requester to calculate the neighbor grid code in a certain range in advance. If the workers in the range cannot meet the requirements of the task, the platform may need to recruit from outside the range. This method is not suitable for this situation. ZENG et al.^[7] utilize the multi-secret sharing scheme to preserve location privacy in the MCS task assignment. However, this scheme is only suitable for edge computing environments with fog nodes, but not universal for traditional mobile crowdsensing. Different from the above work, the bilateral privacy-preserving framework (BPPF) in this pa-

per can protect the location privacy of tasks and workers, and it is convenient for the platform to measure the relative distance between tasks and workers in any range at the same time. Our proposed BPPF has universality, which is suitable for most mobile sensing scenarios.

3 Preliminaries

In this section, we briefly introduce the models of the system including the threat model and the privacy model, and the design purpose. Table 1 lists the notations frequently used in this paper.

3.1 System Model

Our system mainly includes four entities: task requesters, workers, crowdsensing platform and offline TTP.

1) Task requesters: Task requesters can be individuals, groups and organizations. They have some location-based data collection tasks, such as collecting traffic flow data of a certain road section, monitoring air pollution of a certain area, and taking photos to investigate commodity prices of a supermarket. They do not have the resources to complete the task, so task requesters pay a certain amount of reward to upload the task to the platform for crowdsensing services.

2) Workers: Workers are users with mobile smart devices (such as mobile phones, tablets and smart-watches). They register as workers on the mobile sensing platform in advance, and use their own devices to complete the tasks assigned by the platform and get paid.

3) Crowdsensing platforms: They have enough computing resources and storage resources to provide mobile sensing services for task requesters. They receive some location-based task requests from task requesters and assign appropriate workers to the task. When a worker completes a task, the platform receives the data uploaded by the worker and sends it to the task requester.

4) Offline-TTP: The offline-TTP is responsible for generating the user's secret key, which is used to encrypt and decrypt the content of the task, as well as the disturbance matrix.

Definition 1 (location matrix): In this paper, we use a $k \times k$

▼Table 1. Notation List

Notations	Description
w	A worker
t	A task
L_w	A matrix to represent the current location of a user
L_t	A matrix to represent the sensing area of a task
R_w	Confusion matrix of workers
R_T	Confusion matrix of task requester
k	The size of the location matrix
α	The matrix assignment variable
$K_\#$	Task encryption key
SOD (R)	Sum of diagonal elements of matrix R

matrix to represent the grid of locations of tasks and workers.

The platform divides the whole geographic space covered by its services into $k \times k$ grids according to a certain granularity^[19]. It is represented by a matrix $L = \{\psi_{ij} | 0 \leq i \leq k-1, 0 \leq j \leq k-1\}$. Each element in the matrix corresponds to the divided grid in the actual geographic space. Therefore, the task requester uses a location matrix L_t to represent the location of the task. If the task is located in the i -th row and the j -th column of the grid space, the element corresponding to the task location matrix is represented by a large number, and the farther the elements of other positions in the position matrix are from the task, the smaller the number is. Similarly, workers use the same method to generate the location matrix L_w .

Definition 2 (travel distance): In the location-based MCS task, we need the worker's current location and the task's location for task allocation. Because the worker needs to cross the city block to reach the task's location, we use Manhattan distance as the travel distance between the worker and the task, i.e., if the location index of a task in the grid space is (i, j) and the worker's location index is (m, n) , the travel distance between tasks and workers is shown as Eq. (1):

$$d(t, w) = |i - m| + |j - n|. \quad (1)$$

3.2 Threat Model

The danger in MCS mainly comes from two aspects: external threats and internal threats^[20]. External threats come from the outside of the MCS system and are initiated by external entities that have no contact with the MCS system. The main way is to steal the communication information of internal entities in the MCS system by monitoring the communication channel of entities in the MCS system, thus threatening the privacy of users. In this paper, the encryption technology is used to prevent such attacks. Internal threats come from the inside of the MCS system and are initiated by entities participating in MCS system activities. In this paper, we only consider privacy threats from the platform. We assume that offline-TTP is completely trusted, and similar to the traditional MCS privacy protection mechanism^[7, 18, 21], we assume that the platform is honest-but-curious, which means that they are not only honest with the pre-defined protocol, but also curious about the privacy of the requester and worker.

3.3 Privacy Model

We divide the privacy model into two categories as follows:

1) **Bilateral location privacy:** The location of workers and tasks needs to be protected. In this paper, the location of tasks and workers is represented by a matrix L , so we need to confuse the matrix L to protect the location privacy of workers and tasks.

2) **Task content privacy:** The task content may be sensitive to task requesters in some cases, so we need to protect task content from the platform.

3.4 Problem Formulation

Then, our problem has emerged as follows:

Bilateral privacy-preserving task allocation problem: In MCS, there is a set of tasks distributed in different locations $T = \{t_1, t_2, \dots, t_m\}$, and a set of workers distributed in different locations $W = \{w_1, w_2, \dots, w_n\}$. The honest-but-curious platform selects a set of workers based on the confusing information of tasks and workers. As shown in Eq. (2), the objective is to minimize the total travel distance of selected workers, where (t_j, w_j) represents an assignment.

$$\sum_{j=1}^{|T|} d(t_j, w_j). \quad (2)$$

4 Overview of BPPF

In this section, we show the details of BPPF, a novel bilateral privacy protection framework, which can protect the location privacy of tasks and workers, as well as the content of tasks, while the task allocation phase does not need the participation of online TTP.

4.1 Service Setup

In order to build the MCS service, the platform first divides the whole geographic space into $k \times k$ grid space (for example, longitude and latitude), and uses the matrix $L_{k \times k}$ to indicate that each element in the matrix corresponds to a grid area in the grid space, and randomly selects value $\alpha \in (k^2, +\infty)$.

4.2 User Registration

In order to participate in MCS activities and prove that they are legal users, task requesters and workers must register in offline TTP in advance and generate authenticated ID: $Sign_{SK}(ID)$, where SK is the authenticated private key of offline-TTP, $Sign_{SK}(ID)$ can be verified by the corresponding authentication public key of the offline-TTP, i.e., if $Verify_{PK}(Sign_{SK}(ID)) = ID$, then the user is legal, and the signature of the TTP can prevent malicious attacks from simply forging ID to participate in MCS activities. At the same time, offline-TTP generates two matrices randomly, namely R_t, R_w ($R_t \times R_w = E$), and symmetric key $K_\#$, where $K_\#$ is used to encrypt and decrypt the task content, and R_t, R_w is the confusion matrix of task requester and worker to protect their location privacy.

4.3 Generation and Confusion of Location Matrix

Firstly, the whole geographic space is represented by a $k \times k$ matrix, and each element in the matrix represents a small grid in the grid space. In order to retain the relative distance information between the task and the worker after confusion, a novel assignment rule of location matrix is designed in this paper. The specific details are shown in Algorithm 1. If the task requester has a task in one of the small grids, the elements in the corresponding location matrix are represented by a maximum number, and the farther the other elements are from the

task location, the smaller the assignment is. That is, if the task is located in the position with the matrix index of (i, j) , then the position element $\psi_{x,y}$ is assigned the value of α^{2k-2} . The farther the other positions are from (i, j) , the smaller the element ψ assignment is. An example is shown in Fig. 1. After the task requester and the worker generate the corresponding location matrix L , in order to protect the location privacy of the task and the worker, they need to transform the actual location matrix L into the privacy-protected location matrix L^* . Finally, the worker and the task requester upload the confusing location matrix to the platform.

Algorithm 1. Generating privacy-preserving location matrix

Input: task Location index (i, j) , worker Location index (m, n)

Output: task confusion location matrix L_t^* , user confusion location matrix L_w^*

1: Initialization: $L_t \leftarrow 0, L_w \leftarrow 0$

2: for each $\psi_{(x,y)} \in L_t$ do

3: $\psi_{x,y} = \alpha^{2k-2-l_i-x-l_j-y}$

4: end for

5: for each $\psi_{(x,y)} \in L_w$ do

6: $\psi_{x,y} = \alpha^{2k-2-l_m-x-l_n-y}$

7: end for

8: $L_w^* = L_w \times R_w$

9: $L_t^* = R_t \times L_t$

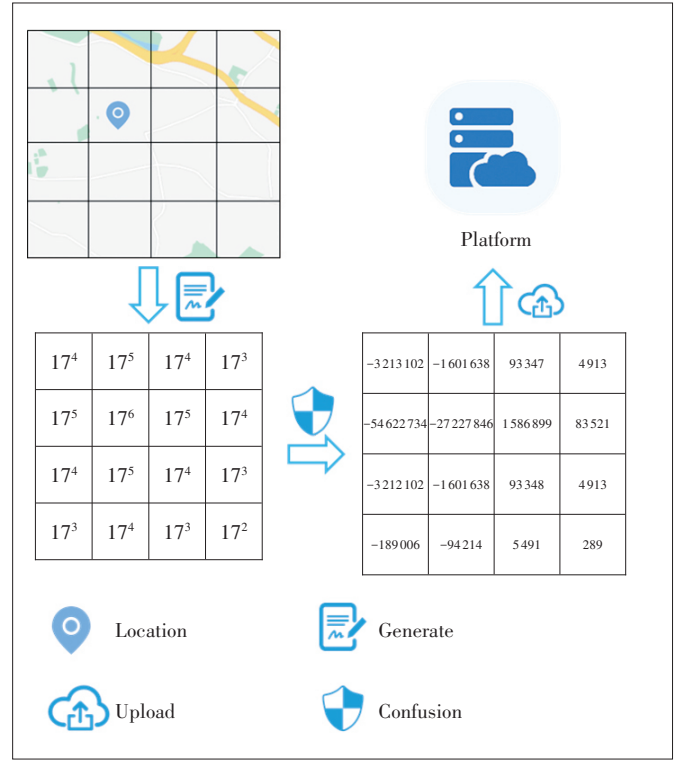
10: Return L_w^*, L_t^*

4.4 Privacy-Preserving Task Submission and Worker Participation

The task requester submits an encrypted task to the platform. For a task, the task requester submits its verifiable ID, the encrypted task content $Enc(K_{\#}, t)$, and the location matrix after confusion L_t^* , i.e., $task \leq t, S_p, Enc(K_{\#}, t), L_t^* >$, where S_t denotes verifiable ID, $Enc(K_{\#}, t)$ represents the task content encrypted with symmetric secret key, and L_t^* represents the location matrix of the task after confusion. After the task requester submits the task to the platform, the platform first passes the verification S_t to check whether the task requester is a legal user, and then the platform assigns the right worker to the task. In order to participate in MCS activities, workers registered on the platform need to upload their verifiable ID and the confusing location matrix L_w^* to the platform, i.e., $user = < w, S_w, L_w^* >$, where S_w denotes verifiable ID, and L_w^* represents the location matrix of the task after the confusion. After receiving the worker's information, the platform first checks whether the worker is legal, then stores the user's information locally and adds it to the candidate worker set.

4.5 Privacy-Preserving Task Matching and Distribution

After receiving the messages from the task and the worker, the platform allocates workers based on the confusion location matrix of task and worker L_t^*, L_w^* . First, the platform computes



▲ Figure 1. An example of generating privacy-preserving location matrix

$R = L_t^* \times L_w^*$, and then takes the sum of diagonal elements of result matrix R , $SOD(R)$ as the standard to measure the distance between a pair of tasks and workers. The larger the value of $SOD(R)$, the shorter the actual distance. The details will be explained in Theorem 1.

Each task needs one worker to complete, and the purpose of the platform is to minimize the total moving distance of the assigned workers. Because the relative distance $SOD(R)$ obtained by the platform is inversely proportional to the actual distance, we can transform the problem into a maximum weight bipartite graph matching problem. The task and the worker are two disjoint point sets in the bipartite graph. The relative distance between the task and the worker $SOD(R)$ is used as the weight of the edge between the task and the worker. Our goal is to find a set of edge sets to maximize the total weight, that is, the total moving distance of assigned workers is the minimum.

Theorem 1. The actual distance d between workers and tasks is inversely proportional to the relative distance $SOD(R)$. That is, if $d_1 < d_2$, then $SOD_1(R) > SOD_2(R)$.

Proof. Assume the task's location is (i, j) , the corresponding location matrix is L_t , the worker's location is (m, n) , the corresponding location matrix is L_w , $R = L_t^* \times L_w^*$, and $SOD(R)$ is the sum of diagonal elements of matrix R . The formulas are:

$$d_1 < d_2, \quad (3)$$

$$SOD_1(\mathbf{R}) > SOD_2(\mathbf{R}), \quad (4)$$

where d_1 is the Manhattan distance between task 1 and worker 1 and d_2 is the Manhattan distance between task 2 and worker 2. Formulas (3) and (4) can be transformed into the following form:

$$|i_1 - m_1| + |j_1 - n_1| < |i_2 - m_2| + |j_2 - n_2|. \quad (5)$$

$$\sum_{x=0}^{k-1} \sum_{y=0}^{k-1} \alpha^{4k-4-(|i_1-x|+|j_1-y|+|m_1-x|+|n_1-y|)} > \sum_{x=0}^{k-1} \sum_{y=0}^{k-1} \alpha^{4k-4-(|i_2-x|+|j_2-y|+|m_2-x|+|n_2-y|)}. \quad (6)$$

The expansion of $SOD(\mathbf{R})$ is the accumulation of k^2 items, which is recorded as ζ . For the next proof, first of all, we want to get the value range of ζ . According to the properties of the exponential function, the minimum value of ζ is not less than 0. When the index number is the largest, the value of ζ is the largest. Therefore, the problem of finding the maximum value of ζ is to find a point (x, y) in the two-dimensional space of $\{(x, y) | 0 \leq x \leq k-1, 0 \leq y \leq k-1\}$ so that the sum of the Manhattan distance from this point to point (i, j) plus the Manhattan distance to point (m, n) is the minimum. According to the image, the minimum value of $|i-x| + |j-y| + |m-x| + |n-y|$ is $|i-m| + |j-n| = d$, so the maximum value of ζ is α^{4k-4-d} , and finally the value range of ζ is $(0, \alpha^{4k-4-d}]$. So formula (6) can be transformed into formula (7).

$$\alpha^{4k-4-d_1} + \zeta_1 + \zeta_2 + \dots + \zeta_{k^2-1} > \alpha^{4k-4-d_2} + \zeta'_1 + \zeta'_2 + \dots + \zeta'_{k^2-1}. \quad (7)$$

By narrowing the left side of inequality (7) and enlarging the right side, we get inequality (8):

$$\alpha^{4k-4-d_1} > k^2 \alpha^{4k-4-d_2}. \quad (8)$$

We divide both sides of inequality (8) by α^{4k-4-d_2} , and then the inequality (9) is obtained.

$$\alpha^{d_2-d_1} > k^2, \quad (9)$$

where d_1 and d_2 is the Manhattan distance in the grid space and the minimum value of d_2 is d_1+1 . From inequality (9), we can draw the following conclusion: if $\alpha > k^2$ and $d_1 < d_2$, $SOD_1(\mathbf{R}) > SOD_2(\mathbf{R})$ must be true.

4.6 Task Extraction and Execution

After receiving the task assigned by the platform, the work-

er first performs the decryption operation to obtain the task content $Dec(K_{\#}, Enc(K_{\#}, t))$. Only authenticated legitimate users have the secret $K_{\#}$. After the worker completes the task, the collected data is encrypted with the secret key and uploaded to the platform $\langle w, S_w, Enc(K_{\#}, date) \rangle$. After receiving the information forwarded by the platform, the task requester first verifies whether w is a legitimate user and then decrypts the sensor data with the secret key.

5 Experimental Evaluation

In this section, we evaluate the performance of BPPF by theoretical analysis and extensive experiments based on real-world datasets.

5.1 Experimental Setup

1) Datasets. In the simulations, we adopt two widely used real-world datasets: Feeder^[22] and GeoLife^[23]. Feeder contains four kinds of data, i.e., the cellphone call detail records data, smartcard data, taxicab GPS data, and bus GPS data collected from Shenzhen. GeoLife is from the user's mobile phone data which records the user's mobile trajectory data. For the Feeder dataset, we preserve the tax ID, the latitude, and the longitude to construct the worker's attributes. The selected users locate in the area with a latitude ranging from 22.488899 to 22.7491, and a longitude ranging from 113.801048 to 114.241135. For the GeoLife dataset, we construct workers' attributes by extracting the user ID, the latitude and longitude. The selected users locate in the area with a latitude ranging from 40.013225 to 40.424714, and a longitude ranging from 116.327406 to 116.655039. We split the urban area of Feeder and GeoLife into 40×40 grids, each with the size of $1 \text{ km} \times 1 \text{ km}$.

2) Baseline approaches. In this paper, we compare our strategy with the following benchmarks: No-privacy, the strategy which is the no-privacy-version of our strategy without the confusion strategies; PriRadar^[18], a grid-based bilateral privacy protection framework. For the convenience of comparison, we set the hash table $H(t) = 1, H(w) = 1$. This means that a task needs to be completed by one worker, and a worker can only complete one task, meanwhile we do not set the noise hash code.

3) Evaluation metrics. We use the following metrics to evaluate the compared algorithms: the number of completed tasks and the travel distance, which are the main metrics to evaluate our strategy.

4) Setting. All the algorithms were implemented in Java. All the experiments were evaluated on a notebook with Intel Core i7-4720HQ central processing unit (CPU), of which the clock rate is 2.60 GHz and the memory is 8.00 GB. The operation system is 64-bit Windows 10.

5.2 Number of Completed Tasks Evaluation

We first compare the performance of No-privacy, BPPF,

and Pripadar in the number of tasks completed. In the specific experiment, we randomly select the user information from the dataset to construct the worker attributes. As shown in Fig. 2, we can see the No-privacy method and our BPPF method can allocate all tasks by changing the number of tasks from 50 to 130, because both methods can find a maximum matching scheme in task allocation, while the number of tasks completed by Pripadar method increases with the increase of search distance under the same number of tasks. This is because the effect of Pripadar is related to the search distance (d). With the increase of search distance, the platform has more workers to choose and a task is more likely to be assigned to a worker. Then we compare the numbers of completed tasks with different algorithms with respect to different numbers of workers, and we change the number of workers from 50 to 130. As shown in Fig. 3, we can see that the No-privacy and BPPF methods can assign all tasks even when the number of workers is equal to the number of tasks. While in the case of the same search distance, the number of tasks completed by Pripadar method increases with the increase of the number of workers, because with the increase of the number of workers, there are more workers to choose within the search distance of a task.

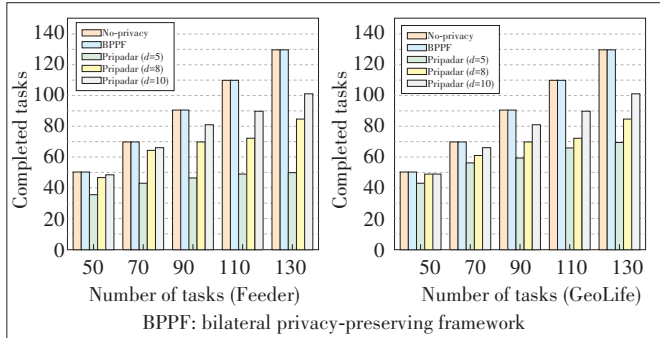
5.3 Travel Distance Evaluation

We then evaluate the travel distance. The travel distance is defined as the sum of the Manhattan distances between all assigned workers and task pairs. For consistency, we set the search distance (d) of Pripadar so that all tasks can be as-

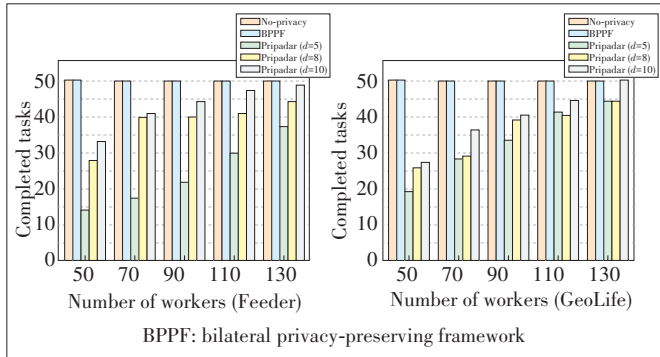
signed. As shown in Fig. 4, the moving distances of the three methods increase with the number of tasks. We can observe that the No-privacy method and our BPPF method are always better than Pripadar. As shown in Fig. 5, the travel distance of the three methods decreases with the increase of the number of workers. This is because with the increase of the number of workers, a task is more likely to be assigned to a closer worker. The result of GeoLife dataset does not change obviously. Because the user location distribution of GeoLife dataset is relatively concentrated, while the user location distribution of Feeder is relatively scattered, the workers assigned by GeoLife dataset are already the optimal workers when the number of workers is small. Even if the number of workers increases, the change of moving distance will not be greatly reduced. From Figs. 4 and 5, we can observe that the No-privacy method and our BPPF method are always better than Pripadar, and our BPPF method has almost no difference with No-privacy, which indicates that our method can achieve accurate task allocation under the condition of protecting tasks' and workers' positions.

5.4 Location Privacy

We also do experiments to verify our proposed bilateral privacy protection mechanism based on the matrix. We set the task location as $(0,0)$, then 100 workers are set in different grids, and the relative distance SOD (R) calculated by the platform is shown in Fig. 6. We can observe that the farther the worker is from the task, the smaller the relative distance is,



▲ Figure 2. Completed tasks vs. number of tasks



▲ Figure 3. Completed tasks vs. number of workers

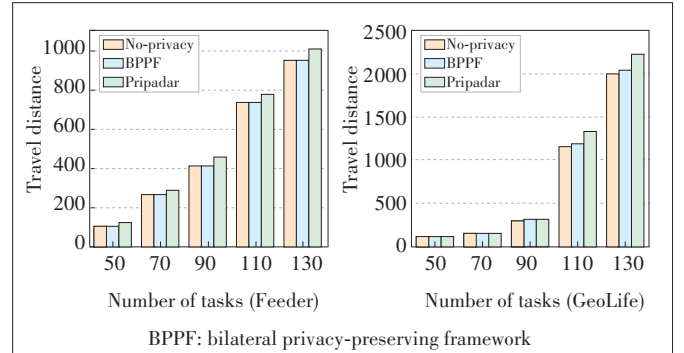
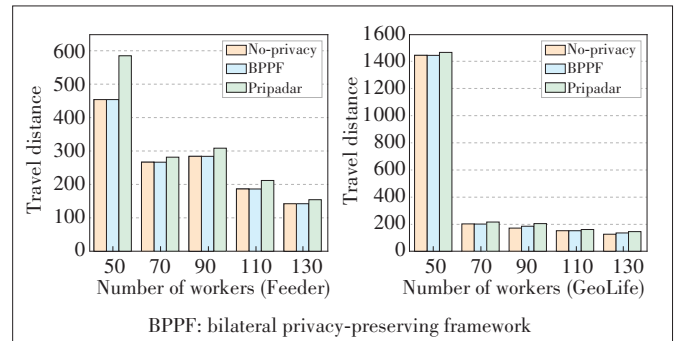
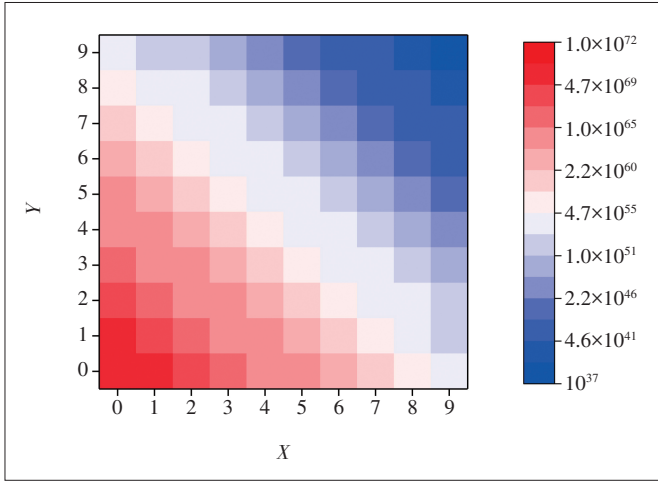


Figure 4. Travel distance vs. number of tasks



▲ Figure 5. Travel distance vs. number of workers



▲ Figure 6. Sum of diagonal elements of result matrix R

which also proves Theorem 1 mentioned in Section 4.

The sensing region of a task is represented by a matrix L_t , which is randomized by a confusing matrix R_r to generate the privacy preserving location matrix L_t^* , and the worker's location is also replaced by a privacy preserving location matrix L_w^* . The platform can not know any location information of tasks or workers through these two matrices L_t^* and L_w^* . The elements and location mapping rules of the actual location matrix are changed by matrix multiplication $L_t \times R_r$, and this change varies with the confusion matrix R_r , so the platform cannot find the information of the actual location matrix from the privacy protected location matrix. After the platform receives the privacy protection location mentioned above, the user's actual location is hidden in the grid. Even if the user's actual grid is obtained, it can protect the user's location privacy to a certain extent.

5.5 Computational Overhead

Finally, we discuss the computation cost of our method. Because our privacy-preserving method is based on matrix multiplication, the computation cost in BPPF depends on the dimension of the matrix, such as the granularity of the grid. The smaller the granularity, the larger the precision and the size of the matrix. For task requesters and workers, the computation cost is $M_{k \times k} \times M_{k \times k}$, which means one matrix multiplication operation. For the platform, in order to allocate tasks, the platform needs to calculate the sum of diagonal elements of two matrix multiplication result matrices, so the platform can only calculate diagonal elements, and the computation cost is $1/k \times (M_{k \times k} \times M_{k \times k}) + O(n^3)$, where $O(n^3)$ means the time complexity of the weighted bipartite graph matching algorithm. The time taken by a user to generate the privacy protection location matrix in the case of different size matrices is shown in Fig. 7. It is acceptable for users to spend such computing resources to ensure privacy. Finally, Fig. 7 shows the running time of the whole framework with a fixed number of tasks of 50.

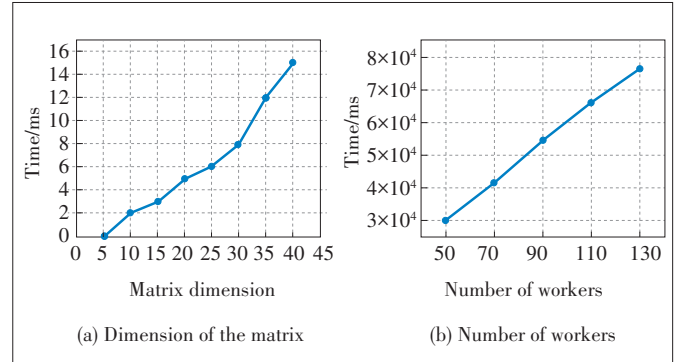


Figure 7. Time vs. dimension of the matrix and number of workers

6 Conclusions

In this paper, we study the problem of bilateral privacy protection in mobile sensing. We map the location of workers and tasks to the grid, use a novel location matrix generation method to represent the user's location information, and propose a location matrix obfuscation method based on matrix multiplication, which can preserve the relative distance information between tasks and workers while protecting their location privacy. Finally, extensive simulations based on real world datasets verify the performance of our method.

References

- [1] LUO T, HUANG J W, KANHERE S S, et al. Improving IoT data quality in mobile crowd sensing: a cross validation approach [J]. IEEE Internet of Things journal, 2019, 6(3): 5651 – 5664. DOI: 10.1109/JIOT.2019.2904704
- [2] GANTI R K, YE F, LEI H. Mobile crowdsensing: current state and future challenges [J]. IEEE communications magazine, 2011, 49(11): 32 – 39. DOI: 10.1109/MCOM.2011.6069707
- [3] LIU Y T, KONG L H, CHEN G H. Data-oriented mobile crowdsensing: A comprehensive survey [J]. IEEE communications surveys & tutorials, 2019, 21(3): 2849 – 2885. DOI: 10.1109/COMST.2019.2910855
- [4] WANG L Y, ZHANG D Q, YANG D Q, et al. Sparse mobile crowdsensing with differential and distortion location privacy [J]. IEEE transactions on information forensics and security, 2020, 15: 2735 – 2749. DOI: 10.1109/TIFS.2020.2975925
- [5] XIAO M J, GAO G J, WU J, et al. Privacy-preserving user recruitment protocol for mobile crowdsensing [J]. IEEE/ACM transactions on networking, 2020, 28(2): 519 – 532. DOI: 10.1109/TNET.2019.2962362
- [6] ZHANG Y H, LI M, YANG D J, et al. Tradeoff between location quality and privacy in crowdsensing: An optimization perspective [J]. IEEE Internet of Things journal, 2020, 7(4): 3535 – 3544. DOI: 10.1109/JIOT.2020.2972555
- [7] ZENG B, YAN X F, ZHANG X L, et al. BRAKE: bilateral privacy-preserving and accurate task assignment in fog-assisted mobile crowdsensing [J]. IEEE systems journal, 9278, (99): 1 – 12. DOI: 10.1109/JSYST.2020.3009278
- [8] GAO G J, XIAO M J, WU J, et al. DPDT: A differentially private crowd-sensed data trading mechanism [J]. IEEE Internet of Things journal, 2020, 7(1): 751 – 762. DOI: 10.1109/JIOT.2019.2944107
- [9] SONG S W, LIU Z D, LI Z J, et al. Coverage-oriented task assignment for mobile crowdsensing [J]. IEEE Internet of Things journal, 2020, 7(8): 7407 – 7418. DOI: 10.1109/JIOT.2020.2984826

- [10] WANG T, XIE X K, CAO X, et al. On efficient and scalable time-continuous spatial crowdsourcing: full version [EB/OL]. (2020-10-29) [2021-01-25]. <https://arxiv.org/abs/2010.15404>
- [11] AKTER S, YOON S. Location-aware task assignment and routing in mobile crowd sensing [C]//2020 International Conference on Information and Communication Technology Convergence (ICTC). Jeju, Korea (South): IEEE, 2020: 51 – 53. DOI: 10.1109/ICTC49870.2020.9289316
- [12] LIU Y, GUO B, WANG Y, et al. TaskMe: multi-task allocation in mobile crowd sensing [C]//Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. Heidelberg, Germany: ACM, 2016: 403 – 414. DOI: 10.1145/2971648.2971709
- [13] GUO B, LIU Y, WU W L, et al. ActiveCrowd: A framework for optimized multi-task allocation in mobile crowdsensing systems [J]. IEEE transactions on human-machine systems, 2017, 47(3): 392 – 403. DOI: 10.1109/THMS.2016.2599489
- [14] WANG L Y, YANG D Q, HAN X, et al. Mobile crowdsourcing task allocation with differential-and-distortion geobfuscation [J]. IEEE transactions on dependable and secure computing, 2021, 18(2): 967 – 981. DOI: 10.1109/TDSC.2019.2912886
- [15] WANG Z B, HU J H, LV R, et al. Personalized privacy-preserving task allocation for mobile crowdsensing [J]. IEEE transactions on mobile computing, 2019, 18(6): 1330 – 1341. DOI: 10.1109/TMC.2018.2861393
- [16] WANG X, LIU Z, TIAN X H, et al. Incentivizing crowdsensing with location-privacy preserving [J]. IEEE transactions on wireless communications, 2017, 16(10): 6940 – 6952. DOI: 10.1109/TWC.2017.2734758
- [17] NI J B, ZHANG K, XIA Q, et al. Enabling strong privacy preservation and accurate task allocation for mobile crowdsensing [J]. IEEE transactions on mobile computing, 2020, 19(6): 1317 – 1331. DOI: 10.1109/TMC.2019.2908638
- [18] YUAN D, LI Q, LI G L, et al. PriRadar: A privacy-preserving framework for spatial crowdsourcing [J]. IEEE transactions on information forensics and security, 2020, 15: 299 – 314. DOI: 10.1109/TIFS.2019.2913232
- [19] YAN K, LUO G C, ZHENG X, et al. A comprehensive location-privacy-awareness task selection mechanism in mobile crowd-sensing [J]. IEEE access, 2019, 7: 77541 – 77554. DOI: 10.1109/ACCESS.2019.2921274
- [20] GISDAKIS S, GIANNETSOS T, PAPADIMITRATOS P. Security, privacy, and incentive provision for mobile crowd sensing systems [J]. IEEE Internet of Things journal, 2016, 3(5): 839 – 853. DOI: 10.1109/IIOT.2016.2560768
- [21] ZHUO G Q, JIA Q, GUO L K, et al. Privacy-preserving verifiable data aggregation and analysis for cloud-assisted mobile crowdsourcing [C]//IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications. San Francisco, USA: IEEE, 2016: 1 – 9. DOI: 10.1109/INFOCOM.2016.7524547
- [22] ZHANG D S, ZHAO J J, ZHANG F, et al. Feeder: supporting last-mile transit with extreme-scale urban infrastructure data [C]//Proceedings of the 14th International Conference on Information Processing in Sensor Networks. Seattle, USA: IPSN, 2015: 226 – 237. DOI: 10.1145/2737095.2737121
- [23] ZHENG Y, ZHANG L Z, XIE X, et al. Mining interesting locations and travel sequences from GPS trajectories [C]//Proceedings of the 18th international conference on World Wide Web. Madrid, Spain: ACM, 2009: 791 – 800. DOI: 10.1145/1526709.1526816

Biographies

LIU Junyu received his bachelor's degree in computer science and technology from Jilin University, China in 2019. Currently, he is pursuing for the master's degree in computer science and technology at Jilin University. His current research interests include mobile crowdsensing and privacy preserving in mobile computing.

YANG Yongjian received his B.E. degree in automatization from Jilin University of Technology, China in 1983, M.E. degree in computer communication from Beijing University of Post and Telecommunications, China in 1991, and Ph.D. in software and theory of computer from Jilin University, China in 2005. He is currently a professor and a Ph.D. supervisor at Jilin University, Director of Key lab under the Ministry of Information Industry, and Standing Director of the Communication Academy. His research interests include network intelligence management, wireless mobile communication and services, and wireless mobile communication.

WANG En (wangen@jlu.edu.cn) received his B.E. degree in software engineering from Jilin University, China in 2011, and his M.E. degree and Ph.D. in computer science and technology from Jilin University in 2013 and 2016. He is currently an associate professor in the Department of Computer Science and Technology, Jilin University. His current research interests include the efficient utilization of network resources, scheduling and drop strategy in terms of buffer-management, energy-efficient communication between human-carried devices, and mobile crowdsensing.



Maximum-Profit Advertising Strategy Using Crowdsensing Trajectory Data

LOU Kaihao¹, YANG Yongjian¹, YANG Funing¹, ZHANG Xingliang²

(1. Jilin University, Changchun 130012, China;

2. China Mobile Group Jilin Co., Ltd., Changchun 130021, China)

Abstract: Out-door billboard advertising plays an important role in attracting potential customers. However, whether a customer can be attracted is influenced by many factors, such as the probability that he/she sees the billboard, the degree of his/her interest, and the detour distance for buying the product. Taking the above factors into account, we propose advertising strategies for selecting an effective set of billboards under the advertising budget to maximize commercial profit. By using the data collected by Mobile Crowdsensing (MCS), we extract potential customers' implicit information, such as their trajectories and preferences. We then study the billboard selection problem under two situations, where the advertiser may have only one or multiple products. When only one kind of product needs advertising, the billboard selection problem is formulated as the probabilistic set coverage problem. We propose two heuristic advertising strategies to greedily select advertising billboards, which achieves the expected maximum commercial profit with the lowest cost. When the advertiser has multiple products, we formulate the problem as searching for an optimal solution and adopt the simulated annealing algorithm to search for global optimum instead of local optimum. Extensive experiments based on three real-world data sets verify that our proposed advertising strategies can achieve the superior commercial profit compared with the state-of-the-art strategies.

DOI: 10.12142/ZTECOM.202102005

<https://kns.cnki.net/kcms/detail/34.1294.TN.20210629.1639.002.html>, published online June 30, 2021

Manuscript received: 2021-03-11

Keywords: billboard advertising; mobile Crowdsensing; probabilistic set coverage problem; simulated annealing; optimization problem

Citation (IEEE Format): K. H. Lou, Y. J. Yang, F. N. Yang, et al., "Maximum-profit advertising strategy using crowdsensing trajectory data," *ZTE Communications*, vol. 19, no. 2, pp. 29 – 43, Jun. 2021. doi: 10.12142/ZTECOM.202102005.

1 Introduction

Out-door billboards are one of the most effective tools for advertising. According to PQ Media^[1], global digital roadside billboard advertising industry grew by a large margin in 2017; specifically, digital roadside

billboard advertising sales have increased by 10% to a total amount of 3.2 billion dollars in the US. Compared with other advertising methods, the out-door billboard can easily make a deeper impression on potential customers, since it provides the strong visual impact, long placement duration and rich information content.

By advertising on out-door billboards, an advertiser can attract potential customers for his/her products. For some products or activities such as temporary promotion, potential customers may immediately decide whether to go to the shop to purchase products after seeing the advertisement. In this situation, once a potential customer is attracted by the advertisement on the billboard, he/she will purchase the relative prod-

This work is supported by Jilin Science and Technology Department Key Technology Project (20190304127YY), the National Natural Science Foundations of China (1772230, 61972450 and 62072209), Natural Science Foundations of Jilin Province (20190201022JC), National Science Key Lab Fund Project (61421010418), Innovation Capacity Building Project of Jilin Province Development and Reform Commission (2020C017-2), Changchun Science and Technology Development Project (18DY005), Key Laboratory of Defense Science and Technology Foundations (61421010418), and Jilin Province Young Talents Lifting Project (3D4196993421).

A conference version of the paper appeared in *Proceedings of SocialSec*^[2].

uct and the advertiser will obtain the commercial profit. However, due to the advertising budget constraint, an advertiser cannot advertise on all the available billboards. Hence, an advertiser should decide which billboards to do advertising to attract as many potential customers as possible, in order to maximize the commercial profit. Note that whether a billboard can attract a customer is determined by many factors, such as customers' mobility (whether they can see the billboard), customers' preferences (whether they are interested in the product), and the location to buy the product (the detour distance). More importantly, all the above information is privacy-sensitive, especially for the customers' trajectories and preferences, which greatly limits the billboard advertising research.

Most of the existing advertising strategies focus on what advertising content should be delivered and how to select locations for the static roadside billboards, but they do not jointly take potential customers' mobility, preferences and detour distance into consideration. In Ref. [3], NIGAM et al. decide the locations of billboards by using the data collected by radio frequency identification (RFID). In Refs. [4] and [5], the advertisers can select the billboard locations by using GPS and phone data. Besides, in Refs. [6] and [7], the advertising content on the billboard is determined by the potential customers' preferences or their detour distance. All these existing works do not jointly take potential customers' mobility, preferences and detour distance into consideration, which results in the advertiser failing to accurately quantify the commercial profit of the available billboards and thus prevents the corresponding strategies from achieving the maximal commercial profit.

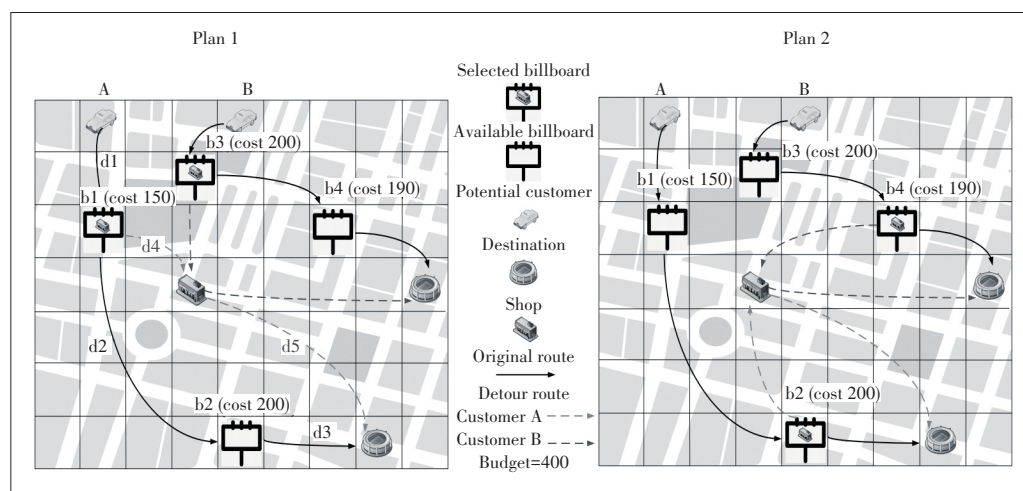
In this paper, we focus on a billboard advertising scenario, which is shown in Fig. 1. There are four available billboards placed at different locations. Two potential customers unconsciously move among the billboard locations. An advertiser wants to do advertising for a product with a limited budget (400 in Fig. 1) and that is not enough to place advertisements

on all billboards. Each billboard has an advertising cost (e.g., $b1=150$, etc.) and also a potential profit. The profit is measured quantitatively by the expected number of attracted customers. Obviously, whether a potential customer is attracted by the billboard is influenced by many factors, such as the probability that the customer can see the billboard, the degree of his/her interest in billboard advertising, and the detour distance that he/she goes to buy the product. We use the detour distance as an example to describe our problem (Fig. 1). Under a total budget of 400, the advertiser can at most cover the costs of two billboards. Two example plans show the different advertising strategies: Plan 1 advertises at billboards $b1$ and $b3$, while plan 2 advertises at billboards $b2$ and $b4$. If the purpose is to minimize the detour distance, obviously plan 1 is better than plan 2. But actually, only using detour distance is not comprehensive, as described above, and many factors should be taken into consideration when we measure a billboard's potential profit. For example, if customers A and B cannot pass by the billboards $b1$ and $b3$, but they can pass by the billboards $b2$ and $b4$, plan 2 is better for this situation. Hence, in order to maximize the commercial profit under the limited budget, this paper focuses on the problem of selecting billboards with the consideration of multiple factors, which can affect the probabilities of customers being attracted.

To solve the above billboard selection problem, we need to collect the potential customers' preferences, mobility patterns and their detour distance, which are privacy-sensitive. Hence, how to collect the potential customers' preferences, mobility patterns and their detour distance is the first challenge. Moreover, the advertiser needs to accurately quantify the expected commercial profit of each billboard to decide which billboards to do advertising, which is the second challenge. When the advertiser has multiple products to advertise, it is essential to make rational use of the budget and a reasonable distribution of advertisements, which is the third challenge.

In this paper, we normalize the influences of the above

three factors, formulate the billboard selection problem as a probabilistic set coverage problem, and propose some effective advertising strategies to address this problem. First, we adopt Mobile Crowdsensing (MCS)^[8-10] to collect the privacy-sensitive customer profiles^[6, 11] including their vehicular trajectories and preferences. For example, if a user has performed some sensing tasks for an MCS application, the application may record his/her GPS trajectories



▲ Figure 1. Problem description of billboard advertising, where the advertiser has only one kind of product

during his/her working time. Also, the user's historical information can be used to infer his/her preferences for the advertising^[11]. For example, if a user has completed many tasks near the food market, the food market may be considered as a preference of this user. Based on their vehicular trajectories and the location of the advertiser's shop, we can estimate the potential customers' detour distance for purchasing the product. With the information, we then study the advertising problem under two situations, where the advertiser may have only one or multiple products. For the first situation where the advertiser has only one kind of product to advertise, we propose two heuristic advertising strategies to greedily choose advertising billboards, which can maximize the total expected commercial profit for the advertiser. For the second situation where the advertiser has multiple products, we adopt the simulated annealing algorithm to search the global optimum instead of local optimum.

The main contributions of this paper are summarized as follows:

- We formulate this billboard advertising problem as a non-deterministic polynomial (NP)-hard problem to select appropriate billboards for the advertiser to achieve the maximum commercial profit with the constraint of budget. We design our advertising strategies with consideration of customers' mobility, preferences and detour distance.
- For the situation where the advertiser has only one kind of product to advertise, we propose two bounded heuristic advertising strategies, whose approximation ratios are $(1 - \frac{1}{e})$.
- For the other situation where the advertiser has multiple products to advertise, we propose an advertising strategy by using the simulated annealing algorithm to search the global optimum.
- We conduct extensive simulations based on three real-world trajectories: roma/taxi^[12], epfl^[13], and geolife^[14]. The results show that compared with other strategies, our advertising strategies can achieve superior commercial profit for the advertiser.

The remainder of this paper is organized as follows. We review the related work in Section 2. We describe the system models and formulate this billboard selection problem in Section 3. We describe the general technologies we used in Section 4. The detailed advertising strategies are proposed in Section 5. In Section 6, we conduct the simulations to determine the performances of our advertising strategies. We conclude this paper in Section 7.

2 Related Work

2.1 Advertising Strategy

There have been many works on advertising strategy. In Ref. [6], WANG et al. propose a utility-evaluation-based optimal searching approach to empower audience targeted bill-

board advertising by using vehicle trajectory data with consideration of audiences' interests. In Ref. [7], ZHENG et al. study a promising application in Vehicular Cyber-Physical Systems (VCPS) to attract potential customers for the shopkeeper by using Roadside Access Points (RAPs). In Ref. [3], NIGAM et al. present the design and implementation of an intelligent advertising system which is integrated in a network and can be used in many retail stores, shopping malls and shopping centers. In Ref. [4], LIU et al. propose a system which uses taxi trajectories to help select the locations of billboards. In Ref. [5], HUANG et al. propose a strategy to maximize the coverage of advertisements with consideration of individuals' interests and mobility patterns. In Ref. [15], KRISHNA et al. develop a new application for detecting significant billboards for adults and older people in street-laying areas. In Ref. [16], AN et al. propose an advertisement system for enhancing the efficiency of advertisement by using the Wi-Fi union mechanism. In Ref. [17], ZHANG et al. optimize the influence of outdoor advertising with the consideration of impression counts. They propose a tangent line based algorithm to select roadside billboards for maximizing the influence of outdoor advertising.

Different from the research works mentioned above where the authors do not jointly take potential customers' mobility, preferences and detour distance into consideration, in this paper, we focus on the problem of selecting billboards with the consideration of the above factors, in order to maximize the commercial profit for an advertiser.

2.2 Mobile Crowdsensing

There have also been some works focusing on mobile crowdsensing. In Ref. [11], KARALIPOULOS et al. draw on logistic-regression techniques from machine learning to learn users' individual preferences from past data in Mobile Crowdsensing. In Ref. [18], ARIYA SANJAYA et al. present an application program, which provides data for SOROT (Citizen Reporting MCS Application) platform by analyzing citizen participation, and speeds up the solution of urban problems. In Ref. [19], CHEUNG et al. propose an algorithm for calculating the optimal user decision-making under general conditions by using the dynamic programming method. Based on the game-theory approach, CAO et al. in Ref. [20] propose an incentive mechanism in order to encourage mobile devices to share their own resource to perform sensing tasks cooperatively. In Ref. [21], GONG et al. focus on the path planning and task assignment problem in Mobile Crowdsensing, so that the total task quality can be maximized with the constraints of user travel distance and budget. In Ref. [22], MARJANOVIĆ et al. propose an edge computing architecture, which is suitable for large-scale MCS services by putting the main MCS functions in the reference of MEC architecture. For truth discovery in mobile crowdsensing, ZHENG et al. in Ref. [23] propose two novel privacy-aware crowdsensing designs with truth discov-

ery so that the bandwidth and computation performance on individual users can be significantly improved. In Ref. [24], WANG et al. propose a two-stage solution to the heterogeneous multi-task assignment (HMTA) problem, which utilizes the implicit spatiotemporal correlation between heterogeneous tasks to effectively handle multiple concurrent tasks in shared resource pools. In Ref. [25], WANG et al. propose a new framework of participatory perceptual multi-task allocation, which coordinates the allocation of multiple tasks on the multi-task PS platform to maximize the overall effectiveness of the system.

These works we mention above focus on different areas of mobile crowdsensing, while we attempt to extract potential customers' implicit information, such as their trajectories and preferences, by using the MCS data, in order to select appropriate billboards for the advertiser.

3 System Model and Problem Formulation

3.1 System Model

Consider that there is an advertising system which is composed of a crowd of potential customers, denoted by the set $U = \{u_1, u_2, \dots, u_n\}$ and a set of available billboards $V = \{v_1, v_2, \dots, v_m\}$. The costs of available billboards are denoted by $C = \{c_1, c_2, \dots, c_m\}$. The different areas in the map can be represented as $L = \{l_1, l_2, \dots, l_h\}$. All billboards are located at different areas and each area has only one billboard. Moreover, the preference types are denoted by the set $A = \{a_1, a_2, \dots, a_j\}$. Hence, without the loss of generality, we denote the preferences of potential customer u_i as $A_{u_i} \subseteq A$. We suppose that each kind of products has an advertisement which can be denoted by $T = \{t_1, t_2, \dots, t_r\}$, and the attributes of product t_x can be denoted by $A_{t_x} \subseteq A$. Meanwhile, each billboard is available for only one advertisement.

In our scenario, at the beginning, each potential customer u_i starts moving from his/her initial location, and goes to his/her destination. Every time a potential customer sees an advertisement for a product, he/she will decide whether to buy the product. The detour distance is an important factor affecting the customer's decision, and then we use $d(u_i)$ to denote the detour distance for u_i , which represents how much more distance u_i needs to go than his/her original route for buying the product. If a potential customer has been attracted to buy the product, this customer cannot be attracted by the same advertisement again. In other words, each potential customer can be attracted by a product at most once. The attracted customers are denoted by U_{attr} .

3.2 Problem Description

With the limit of budget, which is denoted by B , we attempt to choose a set of billboards denoted by $S = \{s_1, s_2, \dots, s_k\}$ from V for advertising. When the advertiser has multiple products,

we also need to choose the advertisements for the selected billboards to maximize the commercial profit. In this paper, we suppose that if a billboard v_i is selected in S , then it will deliver advertisement content to those potential customers whose locations are in its area until the deadline. The deadline is how long the advertiser can use a billboard, and we suppose that all billboards have the same deadline. Our purpose is to find the best advertising strategy for the following billboard selection problem:

$$\begin{aligned} & \text{Maximize } F = \sum_{i=1}^{U_{attr}} \sum_{x=1}^r \phi f^{t_x} - B \\ & \text{s.t. } \sum_{j=1}^k c_{s_j} \leq B, \forall s_j \in S, S \subseteq V, \phi \in \{0, 1\}, \end{aligned} \quad (1)$$

where F is the total commercial profit for the advertiser from billboard advertising. The problem is that an advertiser should attract potential customers as many as possible with limited budget B . In this paper, we assume that if a customer is attracted after he/she sees the advertisement, then the advertiser will obtain a profit from the customer. If a potential customer is attracted by an advertisement t_x , the profit that the advertiser can get is denoted by f^{t_x} and $\phi = 1$. If the customer is not attracted, $\phi = 0$. In order to reduce the complexity of the calculation, we suppose that the customers attracted by the same advertisements can create the same profit for the advertiser, and the customers attracted by the different advertisements may create different profits for the advertiser. In other words, it can be denoted by $f^{t_x} = f^{t_y}, \forall t_x \in T, \forall t_y \in T$ or $f^{t_x} \neq f^{t_y}, \forall t_x \in T, \forall t_y \in T$. Our advertising strategies aim at finding the best set of billboards to deliver the advertisements, so that the commercial profit for the advertiser is maximized, with the constraint that the total costs of selected billboards should be less or equal to the budget.

3.3 NP-Hard Proof

Before solving the above optimization problem, we first attempt to prove that the billboard selection problem is NP-hard, which is shown as follows:

First of all, we formulate this problem in Eq. (1) as the probabilistic set coverage problem, which includes a collection of element sets $X = \{X_1, X_2, \dots, X_m\}$ with the corresponding costs $c = \{c_1, c_2, \dots, c_m\}$. X_i consists of a lot of elements, which is denoted by $O = \{O_1, O_2, \dots, O_n\}$. The associated possibilities that the elements can be covered are denoted by $p = \{p_1, p_2, \dots, p_n\}$ and the associated weights of elements are denoted by $W = \{w_1, w_2, \dots, w_n\}$. The objective is to select a subcollection of X under the budget constraint B to maximize the weights of covered elements.

Then, we reconsider the billboard selection problem in this paper. First of all, we consider the situation where the advertiser has only one type of product. Since the commercial profit depends on the number of attracted customers, we can regard

the potential customers as the element set O . The probabilities that the potential customers decide to buy the product when they see the advertisement can be regarded as p . The profit that the advertiser gets from each customer can be considered as W . We can regard the billboard set we need to choose as X , and the total costs of selected billboards can be regarded as c . We need to select the billboards to attract as many potential customers as possible, hence the billboard selection problem can be regarded as the probabilistic set coverage problem. Since the probabilistic set coverage problem is NP-hard, the billboard selection problem when the advertiser has only one type of product is NP-hard. Moreover, when the advertiser has multiple products to do advertising, the selection problem becomes more complicated, since we should not only consider how to select billboards but also the advertisement placed on the billboard. Hence, under this situation where the advertiser has multiple products, the billboard selection problem is also NP-hard.

4 General Technologies

4.1 Single-Product

In this section, we consider the situation where the advertiser needs to advertise for only one kind of product and potential customers need to decide whether to go to buy the product every time they see the advertisement. We use T to denote the advertisement of the product for ease of calculation and each attracted customer can create the same profit f for the advertiser in this section. We first discuss how to predict the potential customers' mobility patterns. Then, we quantify the influences of customers' preferences and detour distance on customers' attraction probabilities, respectively. Finally, we quantify the utilities of different billboards.

4.1.1 Mobility Prediction

First of all, we attempt to predict each potential customer's location so that we can select the appropriate billboards to improve the effectiveness of advertising. It is not difficult to map each customer's trajectories into a square area in a plane region, especially when the area is small^[26]. Thus, we can grid the area in the map like Fig. 1 and convert each customer's trace into a sequence of grids, in order to reduce the difficulty of calculation. After we grid the map, the billboards' locations can be converted into fixed grids. We assume that, if a potential customer enters a grid which has a selected billboard, he/she will see the advertisement and decide whether to buy the product. In this paper, we adopt the semi-markov model^[27-29] to predict the customers' mobility. One of the most important equations of semi-markov, $Z(\cdot)$ is defined by Eq. (2).

$$Z_u(l_i, l_j, X) = P(S_u^{n+1} = l_j, x_u^{n+1} - x_u^n \leq X | S_u^n = l_i, x_u^n \leq X, S_u^0, \dots, S_u^n; x_u^0, \dots, x_u^n) =$$

$$P(S_u^{n+1} = l_j, x_u^{n+1} - x_u^n \leq X | S_u^n = l_i), \quad (2)$$

where $Z_u(l_i, l_j, X)$ is the probability that the customer u will move from his/her current grid l_i to the grid l_j at or before time X when he/she moves next time. S_u^k represents the customer u 's k -th location during his/her moving and its corresponding arrival time is denoted as x_u^k . The grid that the customer will enter in the next time unit is related to his/her current grid, which can be obtained from the customer's historical trace records. Then, we can define another key equation $Q(\cdot)$, denoted by Eq. (3).

$$Q_u(l_i, l_j, X) = \begin{cases} \sum_{k=1}^h \sum_{x=1}^X (Z_u(l_i, l_k, x) - Z_u(l_i, l_k, x-1)) \cdot Q_u(l_k, l_j, X-x), & i \neq j \\ 1 - \sum_{k=1, k \neq i}^h Z_u(l_i, l_k, X) + \sum_{k=1, k \neq i}^h \sum_{x=1}^X (Z_u(l_i, l_k, x) - Z_u(l_i, l_k, x-1)) \cdot Q_u(l_k, l_i, X-x), & i = j \end{cases} \quad (3)$$

$Q(\cdot)$ denotes the probability of a potential customer u moving across l_j from l_i before time slot X . It is easy to find that the potential customers cannot move from one grid to another when $X = 0$, which is reasonable, so we can get $Q_u(l_i, l_i, 0) = 1$ and $Q_u(l_i, l_j, 0) = 0, (i \neq j)$. Next, we calculate the probability of a customer passing any grid l_j before deadline X as follows:

$$P^j(u) = 1 - \prod_{x=0}^X (1 - Q_u(l_i, l_j, x)). \quad (4)$$

4.1.2 Customer Preference Level

After considering customers' mobility, in order to determine the expected commercial profit of each billboard, we need to measure a potential customer u_i 's preference level for the product T , which is denoted as $P_{prefer}(u_i)$. First, we collect the customer u_i 's preferences A_{u_i} and the product T 's attributes A_T where $A_{u_i} \subseteq A$ and $A_T \subseteq A$. Then the preference level P_{prefer} can be calculated by the following equation:

$$P_{prefer}(u_i) = \frac{A_{u_i} \cap A_T}{A_{u_i}}. \quad (5)$$

Obviously, if the product's attributes A_T can match all the customer's preferences A_{u_i} , then $P_{prefer} = 1$, which means the potential customer u_i would be likely to buy the product by the factor of preferences.

4.1.3 Detour Distance

In the single-product scenario, the potential customer may change his/her trajectory if he/she is attracted by the advertisement. Hence, as mentioned above, another factor that the customer u_i will consider when he/she decides whether to buy a

product is the detour distance, which is denoted by $d(u_i)$. We first use Euclidean distance to measure potential customer's path length. As we can see from the Fig. 1, the original path for the customer A is $d1 + d2 + d3$. When the customer A sees the advertisement from the billboard b1, which is selected for advertising, he/she will decide whether to go to the shop for buying the product. If he/she is attracted to buy the product, the path to the shop is $d4$, and the path from the shop to his/her original destination is $d5$. The detour distance can be calculated as follows:

$$d(u_i) = \begin{cases} \min_{v_j \in V, v_j \in S} d4 + d5 - (d2 + d3), & \text{if } \exists v_j, v_j \in S \\ \infty, & \text{otherwise} \end{cases} \quad (6)$$

During the path, the customer may see a lot of billboards that have been chosen for advertising, so he/she will decide whether to buy the product after he/she sees a selected billboard. Hence, the detour distance should be calculated as the distance from the current billboard to the customer's destination. Then we need to measure how the detour distance affects the probability that the customer would go to the shop. The equation is shown as follows:

$$P_{detour}(u_i) = \begin{cases} 1 - \frac{d(u_i)}{D_{max}}, & \text{if } d(u_i) \leq D_{max} \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

where D_{max} is a predefined threshold and $P_{detour}(u_i)$ represents the detour distance level which affects the probability that the customer will be attracted to buy the product. In this paper, we set D_{max} as the maximum diagonal length in the selected area. It is not difficult to find that the less the detour distance is, the higher chance that the customer will go to the shop for buying the product, which is reasonable.

4.1.4 Billboard Utility

In this part, we use utility to represent the expected commercial profit of a billboard v_j , which is denoted by $F(v_j)$. $F(v_j)$ is the expected commercial profit that the billboard v_j can bring to the advertiser. First of all, we need to calculate the probability that the customer will be attracted to buy the product after he/she sees the advertisement, the equation of which is shown as follows:

$$P_{attract}^{v_j}(u_i) = \alpha P^{l_j}(u_i) \times \beta P_{prefer}(u_i) \times \gamma P_{detour}(u_i), \quad (8)$$

where l_j is the grid that the billboard v_j is located. α , β and γ are relative weights where $\alpha + \beta + \gamma = 1$. By now, the probability that the customer will be attracted to buy the product after he/she sees the advertisement is obtained. Besides, the probability that the customer will be attracted to buy the product can be affected by the different billboards that the customer

sees. Therefore, it is necessary to determine the influences of different billboards on the same potential customer, which can be denoted as follows:

$$P_{attract}(u_i) = 1 - \prod (1 - P_{attract}^{v_j}(u_i)), \quad \forall v_j \in S, \quad (9)$$

where $P_{attract}(u_i)$ is the probability that the customer u_i will be attracted to the shop after he/she sees the current billboard with consideration of different billboards' impacts. In other words, $P_{attract}(u_i)$ is the probability that the potential customer will decide to buy the product at least once when he/she sees the same advertisement many times. Then the utility of a billboard for a specific advertisement can be calculated as follows:

$$F(v_j) = [1 - \prod_{i=1}^n (1 - P_{attract}^{v_j}(u_i))] \times f, \quad v_j \in V, u_i \in U. \quad (10)$$

Then we can get the total utility of the billboard set, which is shown in Eq. (11):

$$F = f \times \sum P_{attract}(u_i) - B, \quad \forall u_i \in U. \quad (11)$$

4.2 Multi-Product

In this subsection, we consider the situation where the advertiser may have multiple products and potential customers do not need to go to the shop immediately to buy products. In other words, the potential customer will not change his/her trajectory if he/she is attracted by the advertisement. We suppose that each product has a corresponding advertisement $T = \{t_1, t_2, \dots, t_r\}$ and each advertisement of a product attracts a customer with different commercial profit which can be denoted as $\eta = \{f_1, f_2, \dots, f_r\}$. Each billboard can only be selected for one advertisement. For example, as shown in Fig. 2, there are four available billboards (b1 - b4) placed at different locations. Two potential customers are unconsciously moving among the billboard locations. The advertiser wants to do advertising for three different products (Type 1 - 3) while he/she has a limited budget (500 in Fig. 2), and that is not enough to place the advertisements on all billboards. Each billboard has an advertising cost (e.g., b1=150, etc.) and also an expected commercial profit. Two example plans show the different advertising strategies under the budget constraint. Plan 1 advertises at billboards b1, b2 and b4 for products 1, 2 and 3, while plan 2 advertises at billboards b1, b2 and b3 for products 3, 1 and 2. In order to maximize the profit within the limited budget, the advertiser needs to determine which plan is better. Next, we will discuss how to address this problem.

4.2.1 Mobility Prediction

First of all, we still consider how to predict customers' mobility patterns. Due to the reason that potential customers' mobility patterns wouldn't be affected by the number of products, the mobility prediction method we proposed earlier can

be applied for this situation.

4.2.2 Customer Preference Level

Next, we will discuss how to determine the customers' preference level in this part. Since the advertiser has multiple products which have different attributes, it is not difficult for us to find that we can reformulate Eq. (5) as follows:

$$P_{prefer}^{t_x}(u_i) = \frac{A_{u_i} \cap A_{t_x}}{A_{u_i}}, \quad (12)$$

where A_{u_i} denotes the preferences of a customer u_i and A_{t_x} denotes the attributes of product t_x . It is obvious that the more the product's attributes match the customer's preferences, the more likely that the customer will decide to buy the corresponding product.

4.2.3 Billboard Utility

In this situation, potential customers do not need to go to the shop immediately to buy products, hence, we can ignore the influence of detour distance on customers' decisions. As a result, the probability that the customer will be attracted can be reformulated as follows:

$$P_{attract}^{v_j, t_x}(u_i) = \alpha P^{l_j}(u_i) \times \beta P_{prefer}^{t_x}(u_i), \quad (13)$$

where l_j is the grid where the billboard v_j is located. α and β are relative weights where $\alpha + \beta = 1$. Since different products have different attributes, we consider there is no competition among different products and the probabilities of a potential customer buying different products are independent. In other words, a customer's purchase of one product does not affect the possibility of buying another different product. Hence, each customer can be attracted by different products and bring more commercial profit to the advertiser.

The utility of a billboard for a specific advertisement t_x can

be calculated as follows:

$$F^{t_x}(v_j) = [1 - \prod_{i=1}^n (1 - P_{attract}^{v_j, t_x}(u_i))] \times f^{t_x}, \quad (14)$$

$$v_j \in V, u_i \in U.$$

Then we can get the total utility of the billboard set, which is shown as follows:

$$F = \sum_{x=1}^r (f_{t_x} \times P_{attract}^{v_j, t_x}(u_i)) - B, \forall u_i \in U, \forall v_j \in S. \quad (15)$$

5 Advertising Strategies

5.1 Single-Product

In this section, we propose two heuristic advertising strategies to address the billboard selection problem for the situation where the advertiser has only one kind of product to advertise.

5.1.1 Same Cost for Each Billboard

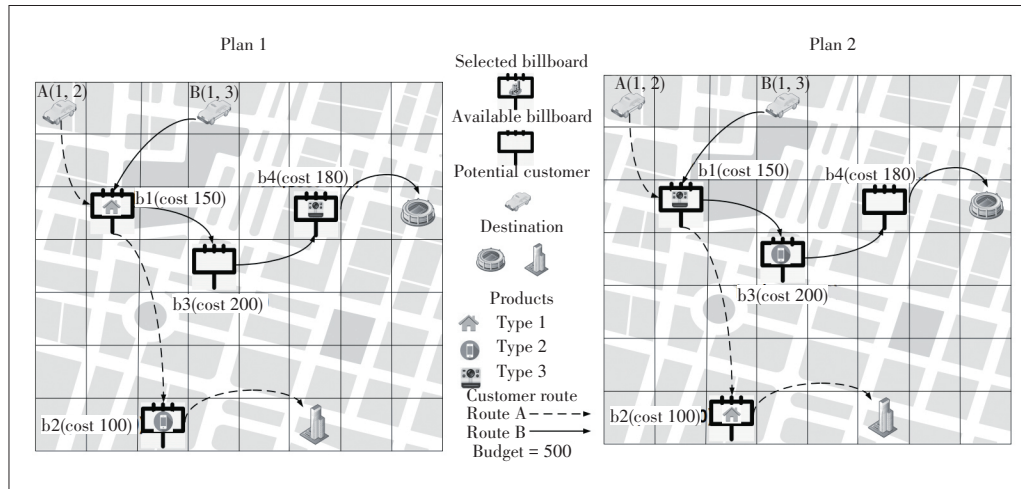
First of all, we consider the situation where all billboards have the same cost. In this situation, we can convert the budget restriction into the billboard's quantity restriction where we need to select a billboard set to maximize the profit for the advertiser with the constraint of billboard number k . The detailed greedy algorithm Advertising Strategy for Constant Cost (ASFCC) is shown in Algorithm 1.

Algorithm 1. Advertising Strategy for Constant Cost (ASFCC)

Input: number of billboard k , a set of billboards V

Output: the selected billboard set S

- 1: $S \leftarrow \emptyset$;
- 2: $F \leftarrow 0$;
- 3: **for** $i=1$ to k **do**
- 4: $v_h \leftarrow \arg \max_{v_h \in V \setminus S} F_{S \cup v_h}$
- 5: $S = S \cup v_h$; update F
- 6: $V = V \setminus v_h$
- 7: **return** the selected billboard set S .



▲ Figure 2. Problem description of billboard advertising, when the advertiser has multiple products

In Algorithm 1, we select the billboard which has the largest utility at the beginning of the algorithm. Then, we continue selecting the billboard, which can maximize the value of the total utility F to be the second billboard among the other billboards and add it into S . This process will be executed for k times until the number of billboards which have been selected meets the requirements or all the billboards have been selected.

The reason why we do not select the billboard with the current largest utility is that the local optimal solution is not necessarily the global optimal solution. For example, consider that there are three billboards a, b, c and two customers u_1, u_2 . Billboards a and b may attract customer u_1 with probabilities 1 and 0.8. Billboard c may attract customer u_2 with probability 0.5. Since each attracted customer can create the same profit for the advertiser, it is obvious that we should select the billboards a and c to achieve the maximum commercial profit.

By now, we have proposed a greedy advertising strategy to address the above NP-hard problem. According to Ref. [30], we can confirm that F is a submodular function, which can be summarized as follows: consider that there are two arbitrary node sets S_1 and S_2 , $S_1 \subset S_2$, and $\forall v_k \in V \setminus S_2$, the submodular property holds, i. e., $F_{S_1 \cup v_k} - F_{S_1} \geq F_{S_2 \cup v_k} - F_{S_2}$. The bound can also be derived from Ref. [30], which is $(1 - \frac{1}{e})$.

5.1.2 Different Costs for Each Billboard

Now, we attempt to propose another heuristic advertising strategy Advertising Strategy for Different Costs (ASFDC) for the situation where all billboards have different costs. As we can see from Algorithm 2, the exhaustive method is used to determine the billboard set S_1 which has the largest expected commercial profit where k is the quantity restriction. Then we execute the greedy process until the budget is lower than the lowest cost of available billboards or all the billboards have been selected to the S_2 . At last, we compare the utility of billboard set S_1 with the utility of billboard set S_2 to determine which is larger to be the final result.

Algorithm 2. Advertising Strategy for Different Costs (ASFDC)

Input: number of billboard k , a set of billboards V , cost set for each billboard C , total budget B

Output: the selected billboard set S

```

1:  $S_1 \leftarrow \arg \max \{F(S_{temp}) \mid |S_{temp}| < k, S_{temp} \subseteq V, \text{ and } c(S_{temp}) \leq B\}$ ;
2:  $S_2 \leftarrow \emptyset$ ;
3: for  $S_{temp} \subseteq V, |S_{temp}| = k$  and  $c(S_{temp}) \leq B$  do
4:   while  $\forall S_{temp} \neq \emptyset$  do
5:      $v_h \leftarrow \arg \max_{v_h \in V \setminus S_2} \frac{F(v_h)}{c(v_h)}$ 
6:     if  $c(S_{temp}) + c_{v_h} \leq B$  then
7:        $S_{temp} = S_{temp} \cup v_h$ 
8:     if  $F(S_{temp}) > F(S_2)$  then
9:        $S_2 \leftarrow S_{temp}$ 
10: if  $F(S_1) > F(S_2)$  then
11:   return the selected billboard set  $S_1$ .
12: else
13:   return the selected billboard set  $S_2$ .
```

By now, we have proposed another advertising strategy to address the billboard selection problem when all billboards

have different costs. According to Ref. [31], we can get $F(S_j) \geq (1 - \frac{1}{e})F(S_{opt}), k \geq 3$, which represents that when $k \geq 3$, the approximation ratio for this algorithm is $(1 - \frac{1}{e})$.

5.2 Multi-Product

We have described the scenario of the situation where the advertiser has multiple products to advertise, which is more difficult than the situation in Section 4. In order to address this problem, we attempt to adopt the simulated annealing algorithm. In this situation, we consider that all billboards have different costs and the same billboard would cost the same for different products.

First of all, we modify the algorithm ASFDC so that it can be applied in this situation, which is shown in Algorithm 3. The difference is that we need to calculate the utilities of each billboard for different products in each iteration, and then we select the billboard with the advertisement which can maximize the total expected commercial profit. This process will be repeated until we run out of the budget. Then we take the result obtained in the previous step as the initial solution of the simulated annealing algorithm Advertising Strategy for Multi-advertisement with Different Costs (ASFMD), which is shown in Algorithm 4.

Algorithm 3. Advertising Strategy for Different Advertisements with Different Costs (ASFDC)

Input: number of billboard k , a set of billboards V , cost set for each billboard C , total budget B , advertisement set T

Output: the selected billboard set S

```

1:  $S_1 \leftarrow \arg \max \{F(S_{temp}) \mid |S_{temp}| < k, S_{temp} \subseteq V, \text{ and } c(S_{temp}) \leq B, \forall t_x \in T\}$ ;
2:  $S_2 \leftarrow \emptyset$ ;
3: for  $S_{temp} \subseteq V, |S_{temp}| = k$  and  $c(S_{temp}) \leq B$  do
4:   while  $\forall S_{temp} \neq \emptyset$  do
5:      $v_h \leftarrow \arg \max_{v_h \in V \setminus S_2} \frac{F^{t_x}(v_h)}{c(v_h)} \quad \forall t_x \in T$ 
6:     if  $c(S_{temp}) + c_{v_h}^{t_x} \leq B$  then
7:        $S_{temp} = S_{temp} \cup v_h^{t_x}$ 
8:     if  $F(S_{temp}) > F(S_2)$  then
9:        $S_2 \leftarrow S_{temp}$ 
10: if  $F(S_1) > F(S_2)$  then
11:   return the selected billboard set  $S_1$ .
12: else
13:   return the selected billboard set  $S_2$ .
```

Algorithm 4. Advertising Strategy for Multi-Advertisement with Different Costs (ASFMD)

Input: a set of billboards V , cost set for each billboard C , total budget B , a set of potential users U , an initial set

$S, \theta, Temp, Temp_{max}$

Output: the final selected billboard set S^*

```

1:  $S^* = S$ 
2: while stopping condition not met do
3:   choose a billboard from  $S$  to delete
4:   choose a billboard from available billboard in  $V$  to add
5:   generate a new billboard set  $S'$ 
6:   if  $F(S') > F(S)$  then
7:      $S = S'$ 
8:   else
9:      $S = S'$  with a probability  $p = \exp(-(F(S) - F(S'))/Temp)$ 
10:  if  $F(S) > F(S^*)$  then
11:     $S^* = S, Temp_b = Temp$ 
12:   $Temp = \theta \times Temp$ 
13:  if  $Temp < 0.01$  then
14:     $Temp_b = 2 \times Temp_b$ 
15:   $Temp = \min\{Temp_b, Temp_{max}\}$ 
16: return the selected billboard set  $S^*$ 

```

In Algorithm 4, we first select a billboard to delete from the selected billboard set S and then select a billboard from the available billboards in V to generate a new solution S' (lines 3–5). Next, we compare the expected commercial profit of S' with that of S to determine which solution should be accepted in the next iteration (lines 6–9). Then, we compare the expected commercial profit of the current solution with the current best solution to update the current best solution (lines 10–11). The temperature $Temp$ is updated at each iteration (lines 11–15). Next, we will discuss the method of removal and insertion we used in Algorithm 4, as well as the stopping conditions.

5.2.1 Billboard Removal Method

We adopt two methods to decide which billboards to delete, which are described as follows:

- Random removal: we randomly select a billboard from the selected billboard set S , and remove it.
- Probability removal: we calculate the expected commercial profit of each selected billboard and remove one of them with probability. The higher the expected profit is, the lower the probability of removal would be.

5.2.2 Billboard Insertion Method

We also design two methods to determine which billboards to insert and the advertisements on the billboards. The detailed description is as follows:

- Probability Insertion: we calculate the expected commercial profit of each available billboard, and select one of them with probability to add to the selected billboard set S . The higher the expected profit is, the higher the probability of insertion would be.
- Expectation Maximization Insertion: we calculate the expected commercial profit of each available billboard, and select

one of them with the max expected profit to insert to the billboard set S .

For a better understanding, we provide an example: as shown in Table. 1, there are two available billboards and three different advertisements. The expected profit has been calculated and shown in the table. First of all, we consider the advertisement that brings the maximum expected profit for each billboard as the final advertisement for that corresponding billboard. Obviously, the final advertisement for billboard A should be advertisement 3, and the final advertisement for billboard B should be advertisement 1. If Probability Insertion is adopted, then we should normalize the expected profit. Thus, the probability for selecting billboard A is $4/7$, and the probability for selecting billboard B is $3/7$. On the other hand, if Expectation Maximization Insertion is adopted, then we will select the billboard with the maximum expected profit, i.e., the billboard A in this example. The process of removal is similar to that of insertion.

We use a multi-thread approach to improve the experimental performance which combines the above removal and insertion methods. As a result, we can create four threads with different combinations of removal and insertion.

5.2.3 Stopping Conditions

We determine two stopping conditions in our simulations, which are listed as follows:

- The maximum number of iterations is set to 1 000 000, and when the iteration number exceeds the limited number, the algorithm would stop.
- The maximum number of iterations without improving is set to 100 000, which means after 100 000 iterations, if the result is not improved, the algorithm would stop.

6 Performance Evaluation

6.1 Simulation Traces and Settings

In this paper, three real-world datasets, the roma/taxi trace set^[12], epfl trace set^[13] and geolife trace set^[14], are adopted to verify the performances of our proposed strategies. In the roma/taxi trace set, 320 taxi drivers that work in the center of Rome are included. The traces in this dataset represent the positions of those taxi drivers, which are collected every 7 seconds and sent to a central server. In the epfl trace set, about 500 taxis' GPS coordinates are included which are collected over 30 days in the San Francisco Bay Area. Each taxi is equipped with a GPS receiver and sends a location-update to

▼Table 1. An example for removing and inserting billboards

	Billboard A	Billboard B
Advertisement 1	200	300
Advertisement 2	300	200
Advertisement 3	400	100

a central server. The records are fine-grained so that we can accurately interpolate user positions between location-updates. In the geolife trace set, there are 17 621 trajectories whose total distance is about 1.2×10^6 km. The total duration of this dataset is about 48 000 h which are collected by different GPS loggers and phones.

First of all, we process the dataset by filtering out some abnormal users including those with discontinuous traces or remote locations. Next, we match these traces into a map area and convert it into a gridded map, which are processed by Baidu Map application programming interface (API). We randomly select a supermarket or a mall in the city as the advertiser's shop and we also select 35 billboards located in different areas to be the candidate billboards.

For the first situation where the advertiser has only one kind of product, we set the α , β and γ in Eq. (8) to $1/3$. The preferences of each customer are randomly generated to reduce the difficulty, but they can also be inferred from each customer's historical information, which is not the focus of this paper. The total number of preference types $|A| = 20$. The deadline is set from 500 to 600. The costs of billboards are set to 10 when the costs are constant and the costs of billboards are set from 10 to 20 when the costs are different. The budget is set from 50 to 170 when the costs are constant and it is set from 100 to 200 when the costs are different. The number of each customer's preferences is set from 5 to 15 in our simulations. Each attracted customer can bring a profit of 10 to the advertiser. We repeat our simulation 10 000 times, taking the average result as the final result.

For the situation where the advertiser has multiple products, we set the deadline in the simulation from 150 to 250. We set the α and β in Eq. (13) to $1/2$. The budget in this situation is set from 100 to 200 and the profit per attracted customer for each product is set from 10 to 30. We set $\theta = 0.999$, $Temp = 10\,000$ and $Temp_{max} = 10\,000$. The other experimental parameters are the same as those when the costs are different for the single-product scene. In this paper, we take the total commercial profit as the evaluation metric to measure the performances of different strategies.

6.2 Strategies for Comparison and Metric

For the first situation where the advertiser has only one kind of product, in order to determine the performances of our advertising strategies, we compare the ASFCC with ASFCC-Basic, Random and Capped Greedy (CG)^[7] when all billboards have the same cost. ASFCC-Basic would select the billboards which have the largest commercial profit, and Random would randomly select the billboards for advertising. The CG would select the billboards, which can maximize the total commercial profit of selected billboards without consideration of customers' preferences.

When all billboards have different costs, we compare ASFDC with ASFDC-Basic and Random, where ASFDC-Basic

would choose the billboards which have the largest commercial profit for advertising. Random would randomly select the billboards with the limit of budget.

For the other situation where the advertiser has multiple products, we compare the ASFMDCC and ASFDADC with ASFDADC-Basic and Random, where ASFDADC-Basic would choose the billboards with the largest commercial profit. The other strategy Random would randomly select the billboards and their advertisements.

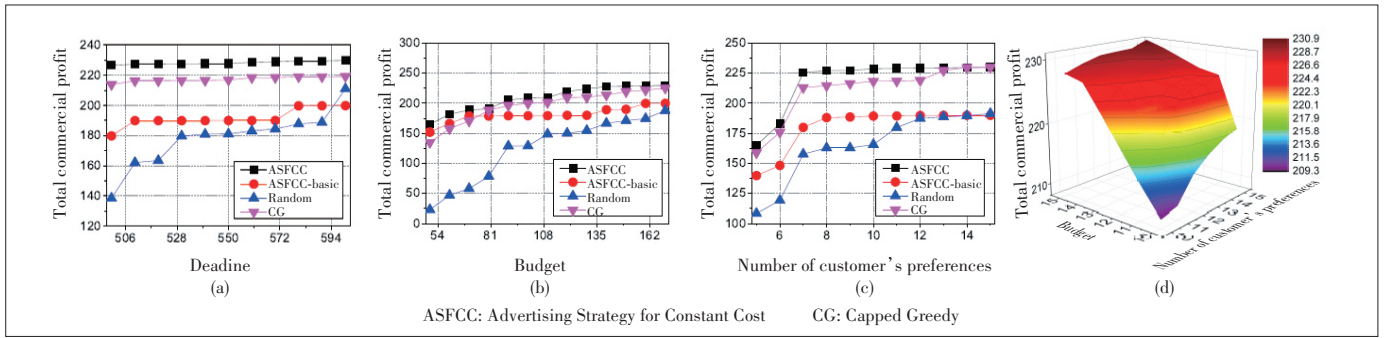
We use the commercial profit as the metric to measure the performances of different advertising strategies. When an advertising strategy performs better, it would have higher commercial profit, which is reasonable. In order to calculate the commercial profit, we need to judge whether a customer is attracted to purchase a product. When a potential customer sees an advertisement on a billboard, he/she has a chance to buy the product in the advertisement. After the deadline of the whole experiment, each potential customer has different purchase probabilities for each product. Through these probabilities, we can use a random number generator to repeatedly test whether a potential customer has bought products, and finally we can get the commercial profit by averaging the payment of potential customers.

6.3 Simulation Results for Single-Product with Constant Billboard Cost

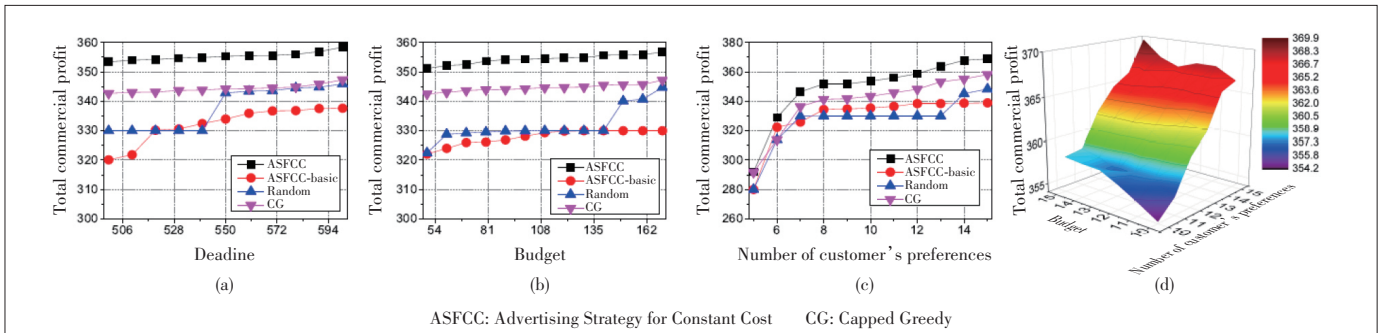
In this section, we aim to test the performance of the proposed strategy ASFCC, when all available billboards have the same cost. In this situation, the advertiser only needs to determine which k billboards to select for advertising. Hence, we compare our advertising strategy ASFCC with ASFCC-Basic, Random and CG on three datasets. The results of the simulation are shown in Figs. 3, 4 and 5.

First of all, we evaluate the performances of the above four strategies on the roma/taxi trace set. As illustrated in Fig. 3, we investigate the influence of the four variables on the total commercial profit of different strategies. Obviously, ASFCC can achieve the maximum commercial profit for the advertiser, while the performance of Random is the worst. We can find that the total commercial profit increases with the growth of deadline, which represents the duration of billboard advertising. It is reasonable for this phenomenon, because when the deadline increases, the selected billboards may have more chances to attract the potential customers so that the commercial profit may increase. We can also find that the total commercial profit shows an upward trend with the increase of budget. The reason is that with the budget increasing, the advertiser can select more billboards for advertising so that more potential customers may be attracted. By changing the number of customers' preferences, we can see that the performances of ASFCC and CG are very close but far better than those of ASFCC-Basic and Random.

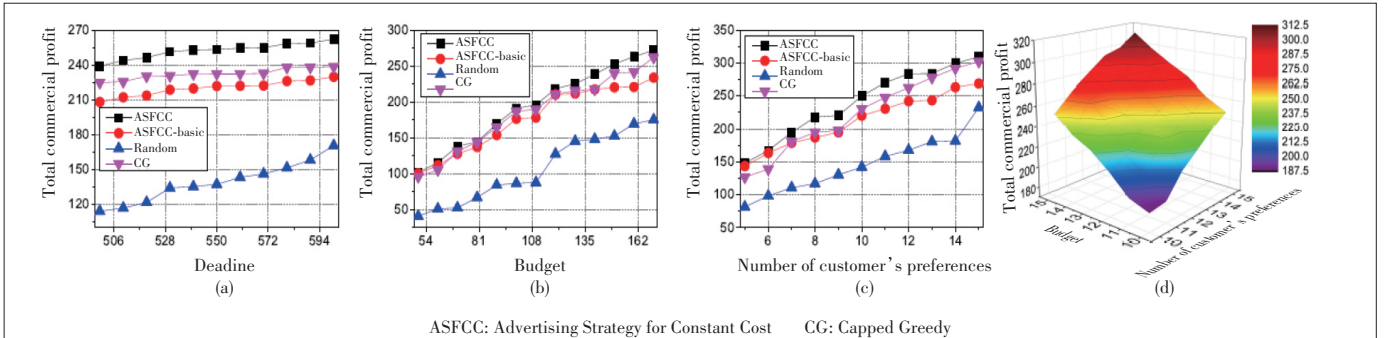
Next, we compare the performances of different advertising



▲ Figure 3. Performances on the roma/taxi trace set, when all billboards have the same cost



▲ Figure 4. Performances on the epfl trace set, when all billboards have the same cost



▲ Figure 5. Performances on the geolife trace set, when all billboards have the same cost

strategies on the epfl trace set, which are shown in Fig. 4. As we can see from Fig. 4a, the performance of ASFCC is far better than that of other strategies. In particular, the strategy Random outperforms the strategy ASFCC-Basic when the deadline is set to 550 – 600, which can prove the limitation of the local optimum of ASFCC-Basic. The similar phenomenon can also be seen in Fig. 4b, where the performance of Random is better than that of ASFCC-Basic. We also test the influence of budget and the number of customers' preferences on ASFCC's performance at the same time and the results are shown in Fig. 4d. It is obvious that the performance of ASFCC shows an upward trend when two variables increase, which is also consistent with the performances in Figs. 4b and 4c.

In Fig. 5, we show the performances of different strategies on geolife traces. We rank the performances of different strategies as follows: ASFCC > CG > ASFCC-Basic > Random. It is reasonable because the instability of the Random leads to poor

performance. At the same time, CG and ASFCC-Basic have their limitations, resulting in the final results not as good as ASFCC. Based on the experimental results of the three datasets, we can find that our strategy ASFCC can always achieve the optimal results under different conditions.

Finally, we conduct the simulations to verify the correctness of the approximation ratio for ASFCC. As shown in Table 2, the results of ASFCC when the deadline is from 500 to 550 are obviously larger than $(1 - \frac{1}{e})$ Optimal, which are consistent with our theoretical analysis.

6.4 Simulation Results for Single-Product with Different Billboard Costs

In this part, we conduct the simulations to compare the performance of ASFDC with other two strategies when all billboards have different costs. The results are shown in Figs. 6, 7 and 8.

First of all, we compare the performances of different advertising strategies on the roma/taxi trace set, and the results are shown in Fig. 6. By changing the deadline, we can find that the strategy ASFDC outperforms the other two strategies. Because ASFDC selects the billboards with the consideration of potential customers' preferences, detour distance and the probabilities of seeing the billboards, thus the billboards selected by ASFDC can achieve better commercial profit. Note that it is reasonable that the commercial profit may increase when the budget grows, because the advertiser may select more billboards for advertising so that more potential customers can be attracted. The results in Fig. 6b can match our analysis.

Next, we conduct the simulations on the epfl trace set and the results are shown in Fig. 7. Obviously, we can find that the ASFDC performs much better than ASFDC-Basic and Random. Because ASFDC-Basic selects the billboards with maximum commercial profit which is a local optimum, the billboards selected by ASFDC-Basic may be seen by a small group of potential customers so that the commercial profit cannot be maximized. We also test the performance difference be-

tween ASFDC and Random when the costs of billboards are under different distributions (uniform distribution, poisson distribution and normal distribution), which are shown in Fig. 7d. The difference values in Fig. 7d are calculated by subtracting the profit of strategy Random from that of strategy ASFDC. The larger the difference value is, the greater the performance gap between these two strategies would be. As we can see from Fig. 7d, it is clear that the difference value decreases as the budget increases. When the costs of billboards are under uniform distribution, the difference between ASFDC and Random is the greatest, and when the costs of billboards are under normal distribution, the difference is the smallest. However, this difference is small among the three distributions when the budget is same. Hence, our proposed strategy ASFDC can be adopted when the costs of billboards are under these three distributions.

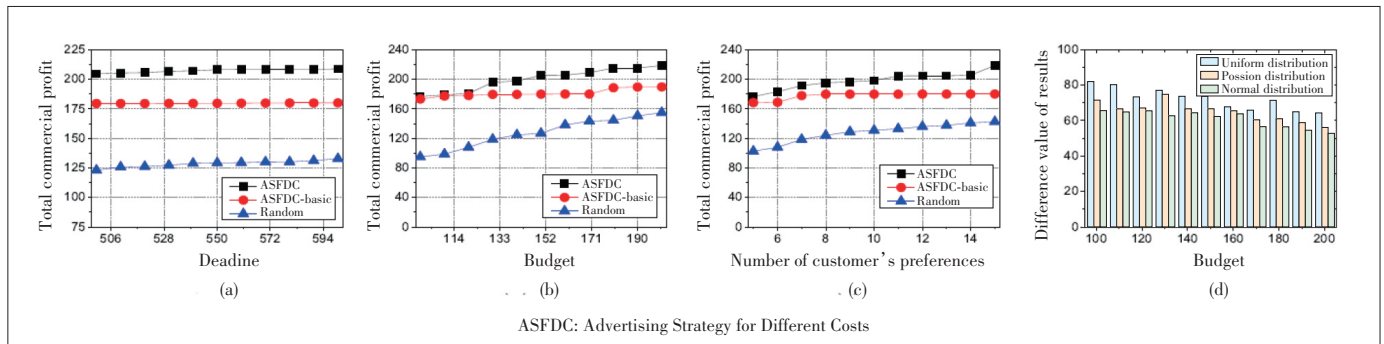
We then evaluate the performances of three strategies: ASFDC, ASFDC-Basic and Random on the geolife traces, which are shown in Fig. 8. We can find that similar phenomenon as that in Figs. 6 and 7 also appears in Fig. 8, where the performance of ASFDC is much better than other two strategies. The reason for this phenomenon is similar to that in Figs. 6 and 7, so it is omitted here.

Finally, we conduct simulations to verify the correctness of the approximation ratio for ASFDC on epfl trace set where the deadline is set from 500 to 550, and the results are shown in Table 3. Compared with the results of $(1 - \frac{1}{e})$ optimal, we can easily see that the results of ASFDC are larger, which matches

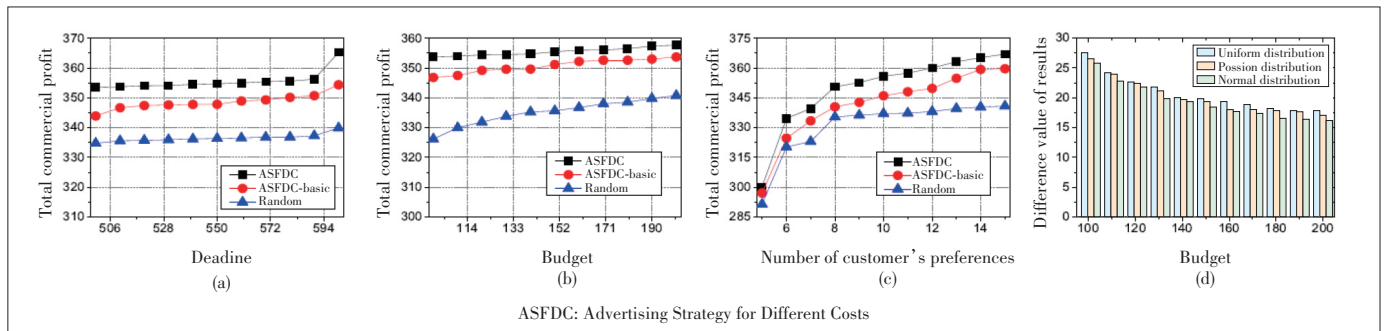
▼ Table 2. Simulation results on epfl, when all billboards have the same cost

Algorithm	Deadline					
	500	510	520	530	540	550
Optimal	37.06	37.27	37.71	37.85	37.87	37.91
$(1 - \frac{1}{e})$ Optimal	23.42	23.56	23.84	23.92	23.94	23.96
ASFCC	35.34	35.39	35.42	35.48	35.52	35.54

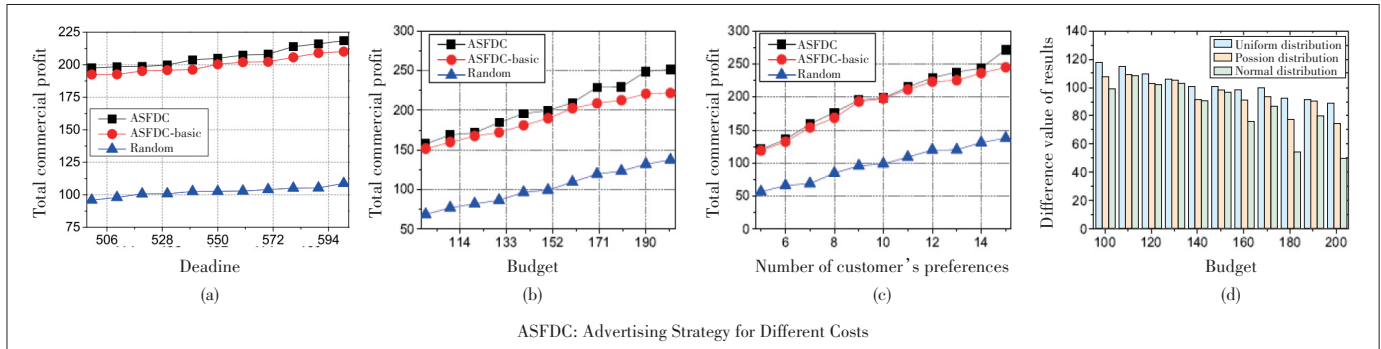
ASFCC: Advertising Strategy for Constant Cost



▲ Figure 6. Performances on the roma/taxi trace set, when all billboards have different costs



▲ Figure 7. Performances on the epfl trace set, when all billboards have different costs



▲ Figure 8. Performances on the geolife trace set, when all billboards have different costs

▼ Table 3. Simulation results on epfl, when all billboards have different costs

Algorithm	Deadline					
	500	510	520	530	540	550
Optimal	37.38	37.41	37.59	37.62	38.02	38.71
$(1 - \frac{1}{e})$ Optimal	23.63	23.65	23.76	23.78	24.03	24.47
ASFCC	35.37	35.40	35.44	35.45	35.47	35.54

ASFCC: Advertising Strategy for Constant Cost

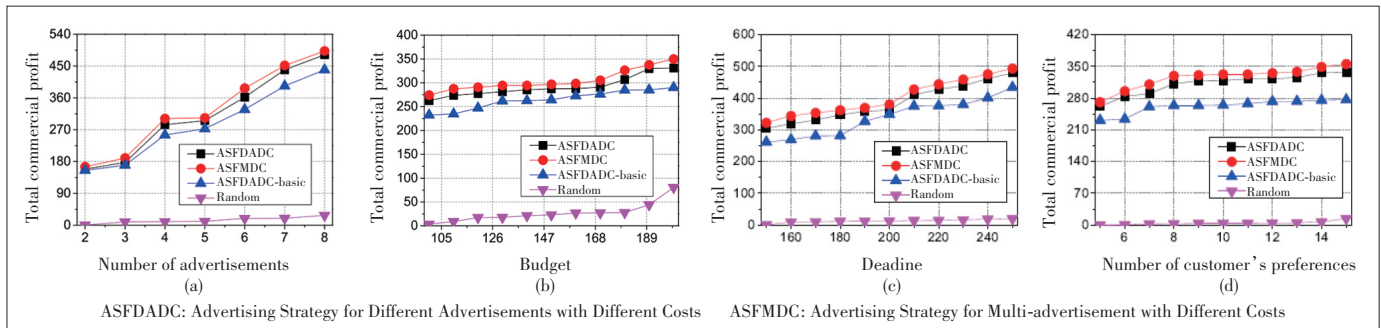
the theoretical analysis.

6.5 Simulation Results for Multi-Product

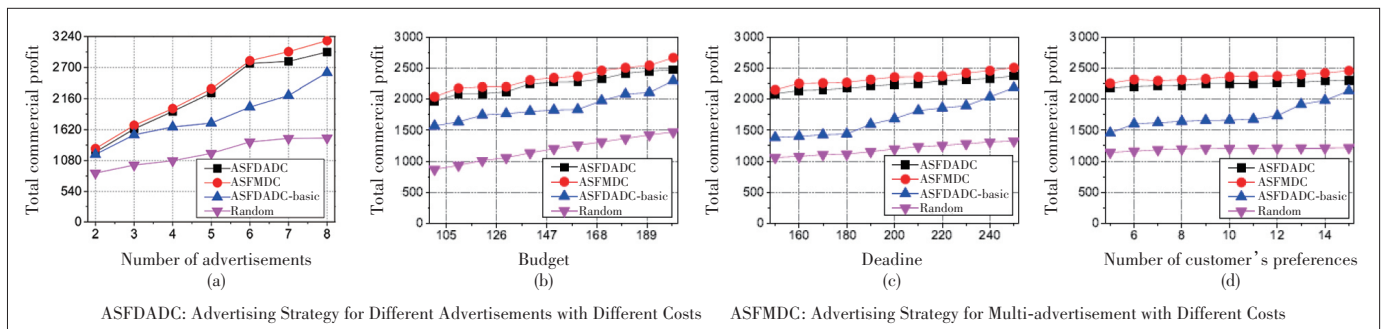
In this part, we conduct the simulations to compare the performance of ASFMDC with other three advertising strategies: ASFDADC, ASFDADC-Basic and Random. The detailed results are shown in Figs. 9, 10 and 11.

Firstly, as shown in Fig. 9, we conduct the simulations on

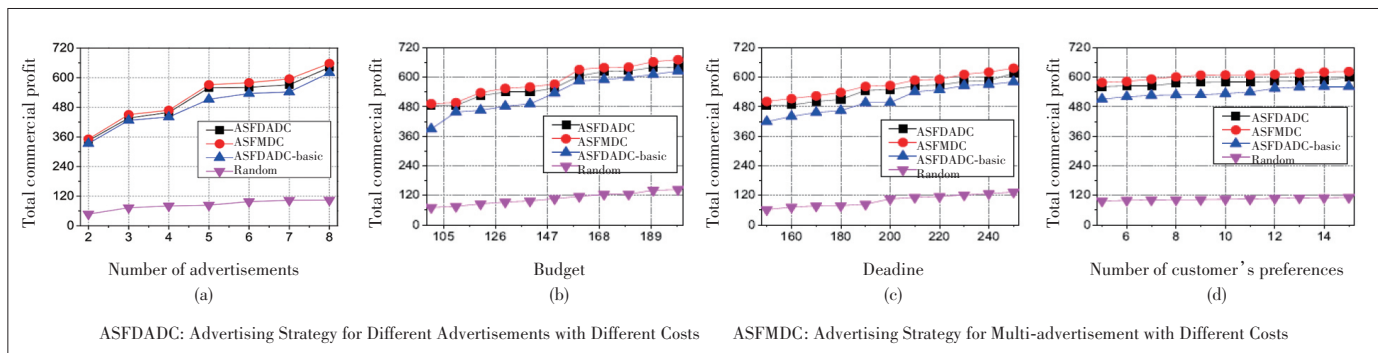
the roma/taxi trace set to compare the performances of different strategies. We can rank the performances of different strategies as follows: ASFMDC > ASFDADC > ASFDADC-Basic > Random, which can prove the effectiveness of our proposed strategy ASFMDC. We can also find that the result of ASFMDC is improved by about 5% - 10% compared with ASFDADC, where ASFMDC is based on ASFDADC. The reason is that ASFDADC selects the billboards which can maximize the total expected commercial profit, while ASFMDC conducts a search in different directions based on ASFDADC, selecting the optimal neighbor solution as the final result. Thus, our strategy can improve 5% to 10% compared with ASFDADC, which is reasonable. The performance of Random is much worse than that of the previous two experiments. Because the advertiser needs to select billboards and determines their corresponding advertisements, Random introduces a lot of uncertainties, which makes the result much worse than the other



▲ Figure 9. Performances on the roma/taxi trace set, when the advertiser has multiple products



▲ Figure 10. Performances on the epfl trace set, when the advertiser has multiple products



▲Figure 11. Performances on the geolife trace set, when the advertiser has multiple products

strategies.

Then, we show the results of the simulations on the epfl trace set in Fig. 10. ASFMDC can achieve the best commercial profit for the advertiser and the performance of ASFDADC is close to ASFMDC. However, the performances of the above two strategies are far better than ASFDADC-Basic, which is different from the phenomenon in Fig. 9. Because the billboards selected by ASFDADC-Basic may achieve a local optimum, when the global optimal solution is very different from the local optimal solution, the result of ASFDADC-Basic would be very bad. In addition, we can find that changing the number of advertisements has a greater impact on the results. This is because a potential customer can be attracted by different products, when the number of advertisements increases, potential customers can be attracted to buy new products and the commercial profit is improved significantly.

Finally, Fig. 11 shows the performances of different strategies on the geolife traces. We can find that the phenomenon in Fig. 11 is similar to that in Fig. 9, where ASFMDC performs better than other three strategies. The performances of ASFDADC and ASFDADC-Basic are close to ASFMDC and much better than that of Random. It is not difficult to find that the results of different strategies increase with the growth of the four variables in Fig. 11, which is reasonable. Because the growth of budget and deadline would increase the number of the selected billboards and the probabilities that selected billboards would be seen by the customers. When the number of products and customers' preferences grow, the potential customers who see the advertisements would be more likely to be attracted. From the above figures, we can prove that our advertising strategy ASFMDC can bring more commercial profit to the advertiser, compared with other common strategies.

7 Conclusions

In this paper, a billboard selection problem is formulated in order to maximize the commercial profit for the advertiser under the limited budget. In order to address this problem, first of all, we adopt MCS to collect potential customers' traces and preferences. Then, we use the semi-markov model to pre-

dict the customers' mobility patterns. Next, we quantify the utility of each billboard with the consideration of customers' preferences, detour distance and the probabilities of seeing the billboards. Two heuristic advertising strategies are proposed in this paper to determine which billboards to select for the situation where the advertiser has only one type of product. Then, we adopt the simulated annealing algorithm to address this problem when the advertiser has multiple products. We conduct the extensive simulations based on the widely-used real-world trajectories: roma/taxi, epfl, and geolife. The results show that our advertising strategies can bring the best commercial profit for the advertiser compared with other advertising strategies.

References

- [1] QUINN P. Global digital OOH media revenues pacing up 13% in 2017, US DOOH advertising expands 10%: PQ media [EB/OL]. (2017-09-06) [2020-01-01]. <https://digitalsignagepulse.com/news/global-digital-oooh-media-revenues-pacing-up-13-in-2017-us-doooh-advertising-expands>
- [2] LOU K H, LI S Q, YANG F N, et al. Advertising strategy for maximizing profit using Crowdsensing trajectory data [C]// Security and Privacy in Social Networks and Big Data. Singapore: Springer Singapore, 2020: 395 - 406. DOI: 10.1007/978-981-15-9031-3_35
- [3] NIGAM S, ASTHANA S, GUPTA P. IoT based intelligent billboard using data mining [C]//International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH). Greater Noida, India: IEEE, 2016: 107 - 110
- [4] LIU D Y, WENG D, LI Y H, et al. SmartadP: visual analytics of large-scale taxi trajectories for selecting billboard locations [J]. IEEE transactions on visualization and computer graphics, 2017, 23(1): 1 - 10. DOI: 10.1109/TVCG.2016.2598432
- [5] HUANG M, FANG Z X, XIONG S L, et al. Interest-driven outdoor advertising display location selection using mobile phone data [J]. IEEE access, 2019, 7: 30878 - 30889. DOI: 10.1109/ACCESS.2019.2903277
- [6] WANG L, YU Z W, YANG D Q, et al. Efficiently targeted billboard advertising using crowdsensing vehicle trajectory data [J]. IEEE transactions on industrial informatics, 2020, 16(2): 1058 - 1066. DOI: 10.1109/TII.2019.2891258
- [7] ZHENG H Y, WU J. Placement optimization for advertisement dissemination in smart City [J]. IEEE transactions on network science and engineering, 2020, 7 (1): 239 - 252. DOI: 10.1109/TNSE.2018.2805768
- [8] GANTI R K, YE F, LEI H. Mobile crowdsensing: current state and future chal-

- lenges [J]. IEEE communications magazine, 2011, 49(11): 32 – 39. DOI: 10.1109/MCOM.2011.6069707
- [9] LIU J W, SHEN H Y, ZHANG X. A survey of mobile crowdsensing techniques: a critical component for the Internet of Things [C]//25th International Conference on Computer Communication and Networks (ICCCN). Waikoloa, USA: IEEE, 2016: 1 – 6. DOI: 10.1109/ICCCN.2016.7568484
- [10] CAPPONI A, FIANDRINO C, KANTARCI B, et al. A survey on mobile crowdsensing systems: challenges, solutions, and opportunities [J]. IEEE communications surveys & tutorials, 2019, 21(3): 2419 – 2465. DOI: 10.1109/COMST.2019.2914030
- [11] KARALIOPOULOS M, KOUTSOPOULOS I, TITSIAS M. First learn then earn: optimizing mobile crowdsensing campaigns through data-driven user profiling [C]//Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc '16). New York, USA: ACM, 2016: 271 – 280
- [12] BRACCIALE L, BONOLA M, LORETI P, et al. CRAWDAD dataset roma/taxi [EB/OL]. (2014-07-17) [2020-01-01]. <https://crawdad.org/roma/taxi/20140717>
- [13] PIORKOWSKI M, SARAFIJANOVIC - DJUKIC N, GROSSGLAUSER M. CRAWDAD dataset epfl/mobility [EB/OL]. (2009-02-24)[2020-01-01]. <https://crawdad.org/epfl/mobility/20090224>
- [14] ZHENG Y, ZHANG L, XIE X, et al. Mining interesting locations and travel sequences from GPS trajectories [C]//Proceedings 13 of the 18th International Conference on World Wide Web (WWW '09). New York, USA: ACM, 2009: 791 – 800
- [15] KRISHNA O, AIZAWA K. Billboard saliency detection in street videos for adults and elderly [C]//25th IEEE International Conference on Image Processing (ICIP). Athens, Greece: IEEE, 2018: 2326 – 2330
- [16] AN T T, CHANG C P, LI Y H, et al. Fog computing architecture-based Wi-Fi union mechanism for Internet advertising system [C]//International Conference on Applied System Innovation (ICASI). Sapporo, Japan: IEEE, 2017: 1024 – 1027. DOI: 10.1109/ICASI.2017.7988110
- [17] ZHANG Y P, LI Y C, BAO Z F, et al. Optimizing impression counts for outdoor advertising [C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: ACM, 2019: 1205 – 1215. DOI: 10.1145/3292500.3330829
- [18] ARIYA SANJAYA I M, SUPANGKAT S H, SEMBIRING J. Citizen reporting through mobile crowdsensing: a smart city case of bekasi [C]//International Conference on ICT for Smart Society (ICISS). Semarang, Indonesia: IEEE, 2018: 1 – 4. DOI: 10.1109/ICTSS.2018.8549976
- [19] CHEUNG M H, HOU F, HUANG J W. Delay-sensitive mobile crowdsensing: Algorithm design and economics [J]. IEEE transactions on mobile computing, 2018, 17(12): 2761 – 2774. DOI: 10.1109/TMC.2018.2815694
- [20] CAO B, XIA S C, HAN J W, et al. A distributed game methodology for crowdsensing in uncertain wireless scenario [J]. IEEE transactions on mobile computing, 2020, 19(1): 15 – 28. DOI: 10.1109/TMC.2019.2892953
- [21] GONG W, ZHANG B X, LI C. Location-based online task assignment and path planning for mobile crowdsensing [J]. IEEE transactions on vehicular technology, 2019, 68(2): 1772 – 1783. DOI: 10.1109/TVT.2018.2884318
- [22] MARJANOVIĆ M, ANTONIĆ A, ŽARKO I P. Edge computing architecture for mobile crowdsensing [J]. IEEE access, 2018, 6: 10662 – 10674. DOI: 10.1109/ACCESS.2018.2799707
- [23] ZHENG Y F, DUAN H Y, YUAN X L, et al. Privacy-aware and efficient mobile crowdsensing with truth discovery [J]. IEEE transactions on dependable and secure computing, 2020, 17(1): 121 – 133. DOI: 10.1109/TDSC.2017.2753245
- [24] WANG L, YU Z W, ZHANG D Q, et al. Heterogeneous multi-task assignment in mobile crowdsensing using spatiotemporal correlation [J]. IEEE transactions on mobile computing, 2019, 18(1): 84 – 97. DOI: 10.1109/TMC.2018.2827375
- [25] WANG J T, WANG Y S, ZHANG D Q, et al. PSAllocator: multi-task allocation for participatory sensing with sensing capability constraints [C]//Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. New York, USA: ACM, 2017: 1139 – 1151. DOI: 10.1145/2998181.2998193
- [26] LIN M, HSU W -J, LEE Z Q. Predictability of individuals' mobility with high-resolution positioning data [C]//Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12). New York, USA: ACM, 2012: 381 – 390
- [27] WANG E, YANG Y J, WU J, et al. An efficient prediction-based user recruitment for mobile crowdsensing [J]. IEEE transactions on mobile computing, 2018, 17(1): 16 – 28. DOI: 10.1109/TMC.2017.2702613
- [28] YUAN Q, CARDEI I, WU J. An efficient prediction-based routing in disruption-tolerant networks [J]. IEEE transactions on parallel and distributed systems, 2012, 23(1): 19 – 31. DOI: 10.1109/TPDS.2011.140
- [29] YANG Y J, LIU W B, WANG E, et al. A prediction-based user selection framework for heterogeneous mobile Crowdsensing [J]. IEEE transactions on mobile computing, 2019, 18(11): 2460 – 2473. DOI: 10.1109/TMC.2018.2879098
- [30] YANG Y J, XU Y B, WANG E, et al. Exploring influence maximization in on-line and offline double-layer propagation scheme [J]. Information sciences, 2018, 450: 182 – 199. DOI: 10.1016/j.ins.2018.03.048
- [31] KHULLER S, MOSS A, NAOR J S. The budgeted maximum coverage problem [J]. Information processing letters, 1999, 70(1): 39 – 45. DOI: 10.1016/s0020-0190(99)00031-9

Biographies

LOU Kaihao (loukh20@mails.jlu.edu.cn) received the B.E. degree in software engineering from Jilin University, China in 2017, M.S. degree in computer science and technology from Jilin University in 2020. He is currently pursuing the Ph.D. degree in computer science and technology at Jilin University. His current research focuses on mobile crowdsensing and multi-agent reinforcement learning.

YANG Yongjian received his B.E. degree in automatization from Jilin University of Technology, China in 1983, M.E. degree in computer communication from Beijing University of Post and Telecommunications, China 1991, and Ph. D. in software and theory of computer from Jilin University, China in 2005. He is currently a professor and a Ph.D. supervisor at Jilin University, the Vice Dean of the Software College of Jilin University, Director of Key lab under the Ministry of Information Industry, Standing Director of the Communication Academy, and a member of the Computer Science Academy of Jilin Province. His research interests include network intelligence management, wireless mobile communication and services, and wireless mobile communications.

YANG Funing received her B.E. degree in software engineering from Jilin University, China in 2010 and master's degree from the school of computer science, Beijing University of Posts and Telecommunications, China in 2013. She is currently a teacher at Jilin University, China. Her current research interests include management, mobile crowdsensing, integration and mining of massive traffic data.

ZHANG Xingliang received his B.E. degree in software engineering from Jilin University, China in 2010. He is currently an engineer at the Network Management Center of China Mobile Group Jilin Co., Ltd. His current research interests include mobile communication data analysis and big data analysis of 5G.

Speed Estimation Using Commercial Wi-Fi Device in Smart Home



TIAN Zengshan, YE Chenglin, ZHANG Gongzhui, HE Wei, JIN Yue
(Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: With the development of Internet of Things (IoT), the speed estimation technology has attracted significant attention in the field of indoor security, intelligent home and personalized service. Due to the indoor multipath propagation, the speed information is implicit in the motion-induced reflected signal. Thus, the wireless signal can be leveraged to measure the speed of moving target. Among existing speed estimation approaches, users need to either carry a specialized device or walk in a predefined route. Wi-Fi based approaches provide an alternative solution in a device-free way. In this paper, we propose a direction independent indoor speed estimation system in terms of Electromagnetic (EM) wave statistical theory. Based on the statistical characteristics of EM waves, we establish the deterministic relationship between the Autocorrelation Function (ACF) of Channel State Information (CSI) and the speed of a moving target. Extensive experiments show that the system achieves a median error of 0.18 m/s for device-free single target walking speed estimation.

Keywords: CSI; speed estimation; electromagnetic wave; direction-independent; autocorrelation function

DOI: 10.12142/ZTECOM.202102006

<https://kns.cnki.net/kcms/detail/34.1294.TN.20210518.0912.002.html>, published online May 18, 2021

Manuscript received: 2021-03-31

Citation (IEEE Format): Z. S. Tian, C. L. Ye, G. Z. Zhang, et al., "Speed estimation using commercial Wi-Fi device in smart home," *ZTE Communications*, vol. 19, no. 2, pp. 44 – 52, Jun. 2021. doi: 10.12142/ZTECOM.202102006.

1 Introduction

Speed estimation systems are appropriate for many emerging smart applications (e.g., human identification and home security). In recent years, significant efforts have been made to explore the indoor device-free speed estimation with fine-grained Channel State Information (CSI). Compared with traditional speed estimation techniques, such as vision^[1], floor sensors^[2] and wearable sensors^[3], the follow-

ing difficulties need to be overcome. Firstly, vision-based schemes are easily to be sheltered by obstacles, which can only work in Line-of-Sight (LoS) environment, and its performance will reduce under dim light or dark conditions, while a potential privacy issue occurs as well. Secondly, the wearable sensor-based approaches require user's positive cooperation, which reduces the user experience and causes inconvenience. However, Wi-Fi based schemes are not affected by obstacles and light conditions, which not only provide a solution with larger coverage and better privacy protection but also can realize indoor device-free speed estimation.

Existing Wi-Fi based systems achieve speed estimation by extracting the motion-induced reflection path using CSI. WiFiU^[4], measures the changing rate of reflected path length

This work was supported in part by the Science and Technology Research Project of the Chongqing Natural Science Foundation Project under Grant No. CSTC2020jcyj-msxmX0842 and the National Natural Science Foundation of China under Grant Nos. 61771083 and 61771209.

to extract gait pattern, but users are usually required to walk along a predetermined route. On the other hand, based on 2D-Frenel zone model, WiDIGR^[5] further eliminates the influence of moving direction but multiple transceiver links are used, which limits the application of speed estimation system. WiWho^[6] distinguishes characteristics of different people to achieve walking gait extraction in the training process where per-person gait signatures are built. These systems either require a time-consuming training process or require users to walk in a predefined route.

In this paper, we propose a Wi-Fi based direction-independent speed estimation system, which avoids redundant training process and other machine learning algorithms like Ref. [7]. Fig. 1 shows an application scenario of the system. Firstly, based on the statistical theory of electromagnetic field, we analyze the relationship between the electromagnetic (EM) wave and human motion theoretically. Then, a speed estimation model based on the autocorrelation function (ACF) of electric field power is derived from the statistical characteristic of angular spectrum. Next, since the information of electromagnetic field is difficult to measure, we further adapt ACF of CSI power to characterize speed information. Finally, an Automatic Multi-scale Peak Detection (AMPD) algorithm is proposed to extract the moving speed from ACF. Fig. 2 shows the system framework.

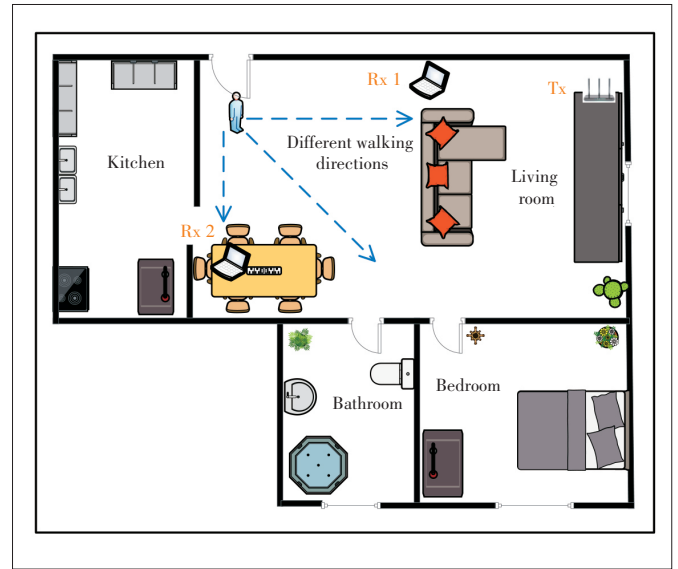
The main contributions of this paper are summarized as follows.

1) Based on the statistical theory of electromagnetic field, we analyze the influence of moving objects on ACF of EM wave, which provides a theoretical basis for the model establishment.

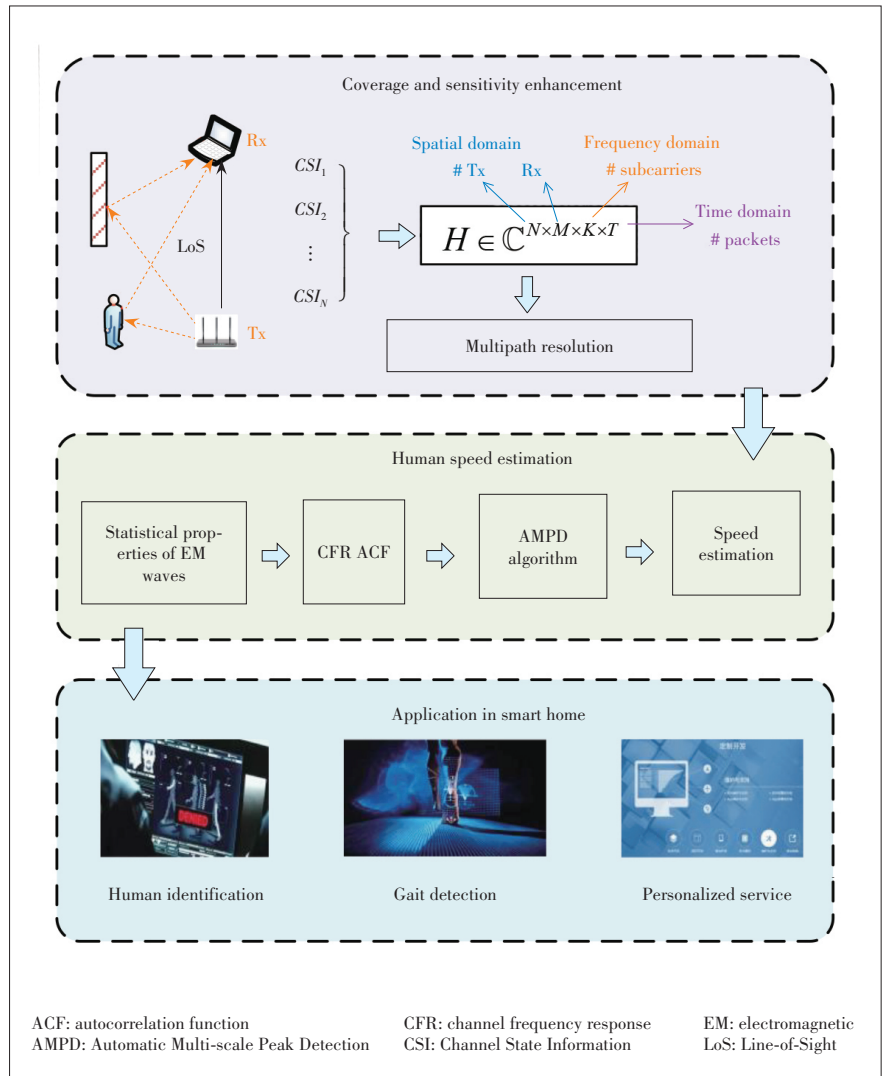
2) We replace the ACF of EM wave with channel power response ACF to eliminate the interference of moving direction. An improved peak detection method is further proposed to extract true speed from ACF.

3) We conduct extensive experiments on commodity Wi-Fi devices in a typical indoor environment with one pair of transceivers. Experimental results show that our proposed speed estimation system achieves a mean absolute error of 0.18 m/s for device-free human walking speed estimation, which is enough in smart home environments.

The rest of this paper is organized as follows. Section 2 introduces the statistical theory of EM waves in indoor environment. Section 3 constructs a speed estimation



▲ Figure 1. Application scenario of the system



▲ Figure 2. System overview

model by studying the ACF of CSI power. The analysis of experimental results are shown in Section 4. Section 5 concludes this paper.

2 Preliminary

2.1 Equivalent Model of Indoor Multipath Propagation

Since the EM waves can be absorbed and scattered by walls, doors, windows, moving objects, etc., radio propagation inside buildings is very difficult to analyze. However, in indoor buildings, EM waves can be approximated as plane waves with obvious electric field statistical characteristics^[8], such as uniform distribution of direction of arrival, polarization and phase. Therefore, the multipath propagation of the wireless signal in indoor environment can be equivalent to the radiation of electric field.

Fig. 3 shows the equivalent model of indoor multipath propagation. In the indoor environment, the signal arrives at the receiver through different paths, including one LoS path, several static reflected paths, and other dynamic reflected paths. In the reverberation chamber, in order to study the influence of target motion on the EM wave, the human can be regarded as an infinite number of scatterers, which can reflect the incident EM wave in all directions. In practice, the transceivers are equipped with an omni-directional antenna, and according to the electric field superposition principle, the received electric field can be decomposed into the sum of the electric fields contributed by all scatterers.

$$\vec{E}_{Rx}(t, f) = \vec{E}_s(f) + \sum_{j \in \Omega_d} \vec{E}_j(t, f), \quad (1)$$

where $\vec{E}_s(f)$ and $\sum_{j \in \Omega_d} \vec{E}_j(t, f)$ are the sum of the electric fields contributed by the static and dynamic scatterers, respectively. $\Omega_d(t)$ denotes the set of dynamic (moving) scatterers.

As the received electric field is a vector, it is very difficult to measure and analyze its characteristics at the receiver. Since the power of the electric field is quantitatively equivalent to the power of the channel frequency response (CFR) of commercial Wi-Fi devices^[9], the power of CFR can be expressed as:

$$G(t, f) = |H(t, f)|^2 = \|\vec{E}_{Rx}(t, f)\|^2, \quad (2)$$

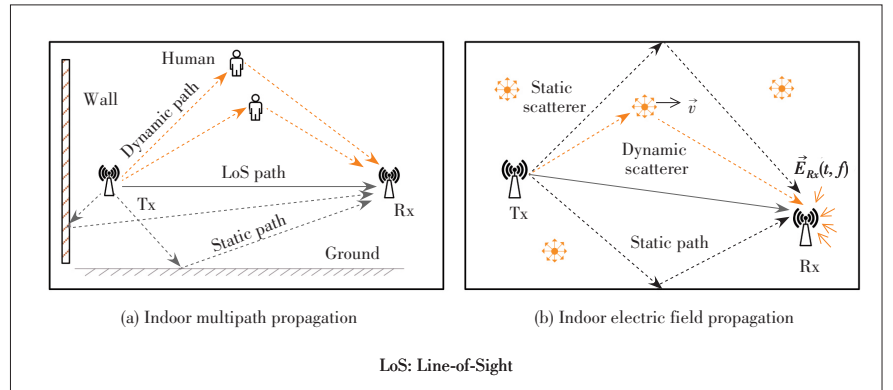
where $H(t, f)$ is the CFR. By measuring the change in CFR power instead of the phase of CFR, we can safely ignore the phase

noises introduced by Carrier Frequency Offset (CFO)^[10].

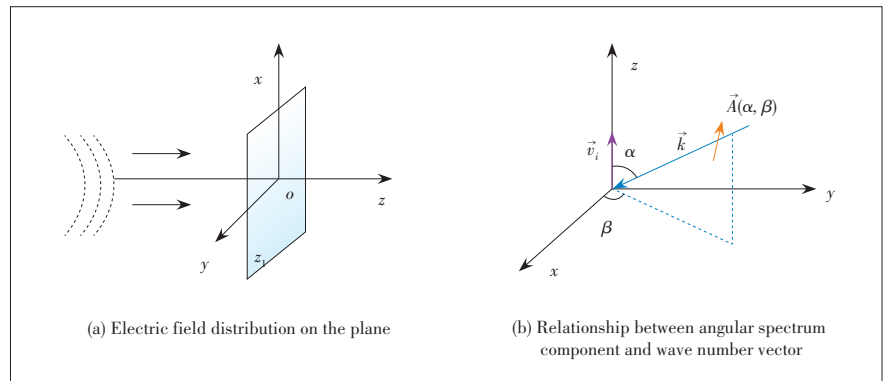
2.2 Statistical Theory of Angular Spectrum

In a relatively short time, the signal will experience the same channel fading in the transmission process according to the channel reciprocity theory. In order to explore the transmission process of EM wave in space, we first analyze the distribution of electric field in a two-dimensional plane, and then extend it to three-dimensional space. The influence of the moving target on the electric field of the receiver can be characterized as the transmission process of the EM wave angular spectrum. Because the complex amplitude of the electric field and the EM wave angular spectrum are Fourier transforms of each other, we further explore the statistical characteristics of the EM field angle spectrum to establish the speed estimation model.

Since the EM waves are spherical waves, we first study the distribution of EM waves on the plane using the surface integral theory, and deduce the complex amplitude of plane electric field. The motion of human will affect the propagation of plane wave angular spectrum. The complex amplitude of electric field is usually represented by the propagation of angular spectrum. Therefore, the propagation of angular spectrum in space is analyzed firstly. As shown in Fig. 4a, a series of EM waves are projected on the $z_1 = (x, o, y)$ plane along the axis. According to the superposition principle of



▲ Figure 3. Equivalent model of indoor multipath propagation



▲ Figure 4. Propagation of angular spectrum

electric field, the complex amplitude of electric field can be regarded as a linear superposition of infinite plane components, written as

$$U_z(x, y, z_1) = \iint_{\infty} A_z(u, v, z_1) \exp[j2\pi(ux + vy)] dx dy, \quad (3)$$

where $U_z(x, y, z_1)$ is the complex amplitude of electric field, and $A_z(u, v, z_1)$ is called the angular spectrum of $U_z(x, y, z_1)$. u and v are the spatial frequencies of the plane wave components respectively, where $u = \cos\alpha/\lambda$ and $v = \cos\beta/\lambda$. $\cos\alpha$ and $\cos\beta$ are called directional cosines of plane waves, where α and β are called the direction angles, which represents the direction of signal transmission. Then, we can get

$$A_z(u, v, z_1) = \iint_{\infty} U_z(x, y, z_1) \exp(-j2\pi(ux + vy)) dx dy. \quad (4)$$

According to the superposition principle of waves, any complex wave can be expressed as a linear combination of plane wave and spherical wave. They both satisfy the wave equation because they are the basic solutions of wave equation^[9]. Taking $U_z(x, y, z_1)$ into Gibbs-Helmholtz equation

$$(\nabla^2 + k^2) \cdot U_z(x, y, z_1) = 0, \quad (5)$$

where the symbol ∇ is Hamiltonian operator and $k = 2\pi f/c$ is wave number. f is the frequency of the wave, and c is the speed of light. The solution of the above equation can be obtained as

$$A_z(u, v, z_1) = A_0(u, v) \exp(jkz\sqrt{1 - \cos^2\alpha - \cos^2\beta}), \quad (6)$$

where $\exp(jkz\sqrt{1 - \cos^2\alpha - \cos^2\beta})$ is called phase delay factor. It can be seen from Eq. (6) that $A_0(u, v)$ is a particular solution of Eq. (5) and it is independent of z , which means the amplitude of angular spectrum is irrelevant of the moving distance. Next, we discuss the changes of angular spectrum in different propagation directions.

If $\cos^2\alpha + \cos^2\beta < 1$, the following conclusions can be obtained.

1) When the plane wave propagates along a certain distance z , only a certain phase shift will be introduced and the amplitude is a constant.

2) The distance between the different components of the plane EM wave to the receiver is related to its propagation direction and the resulting phase shift is also related to the propagation direction.

3) In the propagation of angular spectrum, its phase changes with the direction angle, while its amplitude will not. The spatial frequency of angular spectrum is inversely proportional to its phase delay.

If $\cos^2\alpha + \cos^2\beta = 1$, we have

$$A_z(u, v, z_1) = A_0(u, v). \quad (7)$$

Eq. (7) shows that when the propagation direction of plane wave component is perpendicular to the z axis, and there is no energy propagation along the z axis, which means the component without contribution to the angular spectrum along the z axis.

If $\cos^2\alpha + \cos^2\beta > 1$, we have

$$A_z(u, v, z_1) = A_0(u, v) \exp(-d \cdot \mu), \quad (8)$$

$$\mu = k\sqrt{1 - \cos^2\alpha - \cos^2\beta}. \quad (9)$$

Eqs. (8) and (9) show that the component of the angular spectrum decays exponentially with the increase of the propagation distance d , and will decay to zero in the distance of several wavelengths.

In practice, the received EM wave will radiate in all directions. According to Eq. (7), no energy will propagate along the z axis. Thus, we can establish a space rectangular coordinate system with the z axis as the speed direction.

As shown in Fig. 4b, $\vec{A}(\alpha, \beta)$ represents the angular spectrum of the plane EM wave, which is used to characterize the complex amplitude of the electric field. \vec{k} represents the wave number vector in free space, where its amplitude k represents the wave number, and its phase represents the propagation direction of the wave. The relationship between the wave number vector and the direction cosine is as follows

$$\vec{k} = -k(\vec{x} \cos\beta \sin\alpha + \vec{y} \sin\beta \sin\alpha + \vec{z} \cos\alpha). \quad (10)$$

According to the principle of vector field decomposition, the angular spectrum can be decomposed into two parts

$$\vec{A}(\alpha, \beta) = A_\alpha(\alpha, \beta) \hat{\alpha} + A_\beta(\alpha, \beta) \hat{\beta}, \quad (11)$$

where $A_\alpha(\alpha, \beta)$ and $A_\beta(\alpha, \beta)$ are two complex scalars, and $\hat{\alpha}$ and $\hat{\beta}$ are unit vectors orthogonal to \vec{k} . The complex form of $A_\alpha(\alpha, \beta)$ and $A_\beta(\alpha, \beta)$ can be written as

$$\begin{cases} A_\alpha(\alpha, \beta) = A_{\alpha r}(\alpha, \beta) + jA_{\alpha i}(\alpha, \beta) \\ A_\beta(\alpha, \beta) = A_{\beta r}(\alpha, \beta) + jA_{\beta i}(\alpha, \beta) \end{cases} \quad (12)$$

In a short period of time t , assuming the speed of a single reflector is \vec{v}_i and its displacement is $\vec{v} = \vec{v}_i t$, the electric field complex amplitude of a single reflector can be expressed as

$$\vec{U}(t, f) = \int_0^{2\pi} \int_0^\pi \vec{A}(\alpha, \beta) \exp(-j\vec{k} \cdot \vec{r}) \sin\alpha d\alpha d\beta. \quad (13)$$

In the reverberation chamber, EM wave can be seen as the plane wave. Because the plane wave satisfies Maxwell equa-

tions, the complex amplitude of electric field expressed in Eq. (13) also satisfies Maxwell equations. For a spherical wave, Eq. (13) is a complete and strict expression of plane wave expansion. However, for an aspheric wave, the expansion of plane wave can start from a spherical wave and be extended analytically based on a certain sphere^[8].

Angular spectrum can be deterministic or stochastic. Due to the uniform distribution of the direction of arrival, polarization and phase of plane EM wave, the angular spectrum of EM wave can be regarded as a random variable. In indoor multipath environment, the statistics of received electric field are generated by the multipath and reflection of wireless signal in the process of propagation. The statistical characteristics of angular spectrum can represent the statistical characteristics of electric field in a complex indoor environment.

Because the angular spectrum can be regarded as a series of rays with random phase, its orthogonal components will obey Gaussian distribution according to the central limit theorem^[9]. Due to multipath propagation, the angular spectrum component with orthogonal phase is uncorrelated and its expectation is zero.

$$E[A_\alpha(\alpha, \beta)] = E[A_\beta(\alpha, \beta)] = 0. \quad (14)$$

Due to the real part and the imaginary part of angular spectrum is a constant, its mathematical expectation can be expressed as

$$E[A_{ar}(\alpha_1, \beta_1)A_{ar}(\alpha_2, \beta_2)] = E[A_{ai}(\alpha_1, \beta_1)A_{ai}(\alpha_2, \beta_2)] = C\delta(\alpha_1 - \alpha_2, \beta_1 - \beta_2), \quad (15)$$

where C is a constant. Based on Eqs. (14) and (15), two important relationships can be derived.

$$E[A_\alpha(\alpha_1, \beta_1)A_\beta^*(\alpha_2, \beta_2)] = 0, \quad (16)$$

$$E[A_\alpha(\alpha_1, \beta_1)A_\alpha^*(\alpha_2, \beta_2)] = E[A_\beta(\alpha_1, \beta_1)A_\beta^*(\alpha_2, \beta_2)] = 2C\delta(\alpha_1 - \alpha_2, \beta_1 - \beta_2), \quad (17)$$

where $(\cdot)^*$ is conjugate operation. These two relationships will be leveraged to calculate the ACF of signal power.

3 Speed Estimation Model

3.1 Channel Frequency Response Autocorrelation Function

The electric field at the receiver can be regarded as the superposition of a large number of plane waves with uniformly distributed arrival direction, antenna polarization and phase. Therefore, the angular spectrum can be considered as a random variable with following assumptions:

1) For any (α, β) , both $A_\alpha(\alpha, \beta)$ and $A_\beta(\alpha, \beta)$ are circularly symmetric Gaussian random variables with the same variance

and they are statistical independent^[11].

2) For each dynamic scatterer, the angular spectrums from different directions are not correlated, statistically.

3) For two moments t_1, t_2 and different dynamic scatterers $i_1, i_2 \in \Omega_d$, $\vec{U}(t_1, f)$ and $\vec{U}(t_2, f)$ are statistically independent.

The rationality of Assumption 1 lies in the fact that the angular spectrum is superimposed of many rays bouncing with random phases and thus can assume that each orthogonal component of $\vec{A}(\alpha, \beta)$ tends to be Gaussian under the central limit theorem.

For Assumption 2, because the angular spectrum components corresponding to different directions will result in multiple uncorrelated scattering paths, thus, these angular spectrum components can be assumed to be uncorrelated. Meanwhile, Assumption 3 results from the fact that the CFR is statistically uncorrelated if the transmission distance difference larger than half a wavelength, and the electric fields contributed by different scatterers can thus be assumed to be uncorrelated^[12].

Next, we analyze the CSI power autocorrelation function. Since the expectation of the angular spectrum is zero, the expectation of the complex amplitude of the received electric field is also zero.

$$E[\vec{U}(t, f)] = \int_0^{2\pi} \int_0^\pi E[\vec{A}(\alpha, \beta)] \exp(j\vec{k} \cdot \vec{r}) \sin\alpha d\alpha d\beta = 0. \quad (18)$$

The mean square value of the electric field is directly proportional to the energy density of the electric field, so it is very important to learn the statistical characteristics of the electric field. According to Eq. (18), the expectation of electric field mean square value is

$$E[|\vec{U}(t, f)|^2] = \iint_{4\pi} E[\vec{A}(\Omega_1)\vec{A}^*(\Omega_2)] \exp[j(\vec{k}_1 - \vec{k}_2) \cdot \vec{r}] d\Omega_1 d\Omega_2 = 16\pi C \equiv U_0^2, \quad (19)$$

where $\iint_{4\pi} \triangleq \int_0^{2\pi} \int_0^\pi$, $\Omega = (\alpha, \beta)$, and $d\Omega = \sin\alpha d\alpha d\beta$. Therefore, the mean square value of electric field is a constant and is independent of the reflector position. By deriving the mean square value of each orthogonal component in electric field space, we can get

$$E[|\vec{U}_x|^2] = E[|\vec{U}_y|^2] = E[|\vec{U}_z|^2] = \frac{U_0^2}{3}. \quad (20)$$

Eq. (20) shows that each component of the electric field in an ideal space is the same, which provides theoretical bases for the following study. Based on the above assumptions, the ACF of the electric field can be defined. The received electric field of the reflector can be regarded as a stationary random

process with respect to time T . The Pearson ACF of the received electric field at different times can be written as

$$\rho_{\vec{U}}(\tau, f) = \frac{E[\vec{U}(0, f), \vec{U}(\tau, f)]}{\sqrt{E[|\vec{U}(0, f)|^2] \cdot E[|\vec{U}(\tau, f)|^2]}}, \quad (21)$$

where $E[X, Y] \triangleq E[X \cdot Y^*]$. According to Eq. (19), the denominator of $\rho_{\vec{U}}(\tau, f)$ is U_0^2 . From Eqs. (19) and (21), the ACF can be further deduced as

$$\rho_{\vec{U}}(\tau, f) = \frac{\sin(kv\tau)}{kv\tau}. \quad (22)$$

Eq. (22) shows that the essence of the electric field ACF is a sinusoidal attenuation caused by the motion of the scatterer, and it will decay to zero in a few short wavelengths. The importance of the formula is that the speed information of a single reflector can be derived from the ACF of the received electric field, and the direction of the reflector has no effect on ACF.

3.2 Speed Estimation Algorithm

From Eq. (22), it is very simple to estimate the speed of a single reflector out of the ACF. However, it is difficult to measure the electric field and analyze its ACF. As mentioned above, the power of the electric field can be equivalent to the power of CSI.

$$G(t, f) \triangleq |H(t, f)|^2 = \|\vec{U}(t, f)\|^2. \quad (23)$$

In order to extract speed information from CSI power, the ACF of CSI power is adapted. According to Ref. [9], the ACF of CSI power can be expressed as

$$\rho_c(\tau, f) = \gamma_c(\tau, f) / \gamma_c(0, f), \quad (24)$$

$$\gamma_c(\tau, f) \triangleq \text{cov}(G(t, f), G(t - \tau, f)) = \frac{1}{T} \sum_{t=\tau+1}^T (G(t - \tau, f) - \bar{G}(f))(G(t, f) - \bar{G}(f)), \quad (25)$$

$$\bar{G}(f) \triangleq \frac{1}{T} \sum_{t=1}^T G(t, f), \quad (26)$$

where $\gamma_c(\tau, f)$ represents the auto-covariance function, T is the number of samples, and $\bar{G}(f)$ is the sample mean.

For a Wi-Fi system with bandwidth of 40 MHz and carrier frequency of 5.805 GHz, the difference in wave number k of each subcarrier can be neglected. According to $k = 2\pi f/c$, the maximum EM wave number can be calculated as $k_{\max} = 122.36$, and the minimum is $k_{\min} = 120.68$. Thus, we can get $\forall f, \rho(\tau, f) \approx \rho(\tau)$. Meanwhile, in a short time interval, $\hat{\rho}_c(\tau)$ can be approximated as:

$$\hat{\rho}_c(\tau) \approx S_c \sum_{u \in \{x, y, z\}} (W_1 \hat{\rho}_{U_u}^2(\tau) + W_2 \hat{\rho}_{U_u}(\tau)), \quad (27)$$

where S_c is the scale factor, and W_1 and W_2 represent the weight of $\hat{\rho}_{U_u}^2(\tau)$ and $\hat{\rho}_{U_u}(\tau)$.

$\hat{\rho}_c(\tau)$ can be obtained from the power information of CSI and the speed information is in the right part of the equation. In order to estimate the speed information of the reflector from $\hat{\rho}_c(\tau)$, it is necessary to establish the internal relationship between $\hat{\rho}_c(\tau)$ and each term on the right side of the equation. According to the properties of sampling function, the first closed solution of $\hat{\rho}_U(\tau) = 0$ is exactly the first closed solution of quadratic differential. The symbol Δ is used to represent the differential of CSI power ACF, then we can get

$$\Delta^{(2)} \hat{\rho}_c(\tau) \approx S_c \sum_{u \in \{x, y, z\}} (W_1 \Delta^{(2)} \hat{\rho}_{U_u}^2(\tau) + W_2 \Delta^{(2)} \hat{\rho}_{U_u}(\tau)). \quad (28)$$

From Eq. (28), and the first peak of $\Delta \hat{\rho}_c(\tau)$ is exactly equal to the first solution of equation $\hat{\rho}_U(\tau) = 0$. Therefore, the key of speed estimation is to identify the first local peak of $\Delta \hat{\rho}_c(\tau)$. The speed of moving reflector is calculated as

$$\hat{v} = r/\hat{\tau}, \quad (29)$$

where r is the solution of equation $\Delta^{(2)} \hat{\rho}_U^2(d, f) = 0$. Since the equation has no closed solution, the second smallest solution $r = 0.54\lambda$ is taken as the solution of the equation. The specific algorithm implementation is shown in Algorithm 1.

Algorithm 1. Speed estimation algorithm

Input: Continuous CSI sequence of length T before time t .

$$H(s, f), s = t - \frac{T-1}{F_s}, \dots, t - \frac{1}{F_s}, t.$$

Output: The speed after filtering is $v(t)$.

1: The power response is calculated by CSI amplitude sequence: $G(s, f) = |H(s, f)|^2$

2: for power response

3: Calculation of power response autocorrelation function:

$$\rho_c(\tau, f) = \frac{1}{T} \sum_{s=s_0}^t (G(s - \tau, f) - \bar{G}(f))(G(s, f) - \bar{G}(f))$$

4: where $s_0 = t - (T-1)/F_s + \tau$, $\bar{G}(f) \triangleq \frac{1}{T} \sum_{t=1}^T G(t, f)$,

$\tau \in (0, 0.2)$

5: if $t > T$

6: break

7: end if

8: end for

9: Calculate the average ACF of all carriers: $\rho_c(\tau) =$

$$\frac{1}{F} \sum_{f \in F} \rho_c(\tau, f)$$

10: Calculate the first order differential of ACF: $\Delta \rho_c(\tau) =$

$$\rho_c(\tau) - \rho_c(\tau - 1/F_s)$$

11: For $\Delta\rho_c(\tau)$, the location of the first peak point τ is calculated by peak detection algorithm.

12: The target speed at time t is $v(t) = 0.54\lambda/\tau$.

13: For $v(t)$, using Kalman filter to remove outliers.

3.3 Automatic Multi-Scale Peak Detection Algorithm

The key of speed estimation is to identify the first local peak point of ACF. The traditional peak detection method cannot be adapted to the waves with multiple local peak values, which will lead to the misjudgment of the first peak point in the local scope. Therefore, we design an automatic multi-scale peak detection algorithm to solve this problem, which can maintain a high estimation accuracy when there are multiple local peak values for time-varying waveforms.

The basic principle of AMPD algorithm is to use the local maximum of the signal to detect the peak of all signals. Let $X = [x_1, x_2, \dots, x_N]$ represent the ACF of each sampling time in a moving window. Firstly, we calculate the Local Maxima Scatogram (LMS) the linear detrending of the signal. Then, the local maximum of the signal X is determined using a moving window whose length w_φ satisfies $\{w_\varphi = 2\varphi | \varphi = 1, 2, \dots, L\}$, where φ is the φ -th scale and $L = \lfloor N/2 \rfloor - 1$, and $\lceil x \rceil$ is the ceiling function that gives the smallest integer not less than x . This is realized for every scale k and for $i = \varphi + 2, \dots, N - \varphi + 1$, according to

$$m_{\varphi,i} = \begin{cases} 0 & x_{i-1} > x_{i-\varphi-1} \text{ and } x_{i-1} > x_{i+\varphi-1} \\ r + \alpha_m & \text{other} \end{cases}, \quad (30)$$

where r is a uniformly distributed random number in the range $[0, 1]$, α_m is a constant factor and is usually set as 1. When $i = 1, \dots, \varphi + 1$ or $i = N - \varphi + 2, \dots, N$, the value $r + \alpha$ can be assigned to $m_{\varphi,i}$. These operations of Eq. (30) result in the matrix:

$$\mathbf{M} = \begin{pmatrix} m_{1,1} & m_{1,2} & \dots & m_{1,N} \\ m_{2,1} & m_{2,2} & \dots & m_{2,N} \\ m_{L,1} & m_{L,1} & \dots & m_{L,N} \end{pmatrix} = (m_{\varphi,i}). \quad (31)$$

The value of window length w_k is contained in the φ -th row of the matrix. All elements in matrix \mathbf{M} are in the range of $[0, \alpha_m + 1]$, and matrix \mathbf{M} is the LMS of signal X .

The second step of the algorithm calculates the sum of each row of the LMS matrix:

$$s_\varphi = \sum_{i=1}^N m_{\varphi,i}, \quad \varphi \in \{1, 2, \dots, L\}. \quad (32)$$

The vector $\vec{S} = [s_1, s_2, \dots, s_L]$ contains the information about the scale-dependent distribution. The global minimum of \vec{S} is μ , i.e., $\mu = \arg \min(s_\varphi)$, which represents the scale with the largest local maxima. The value μ will be used in the third step of the algorithm to reshape the LMS matrix by removing all elements $m_{\varphi,i}$ satisfy $\varphi > \mu$, leading to a new matrix $\mathbf{M}_r =$

$(m_{\varphi,i})$ of size $\mu \times N$.

Finally, the AMPD algorithm detects the peak value by calculating the column standard deviation σ_i of the matrix \mathbf{M}_r :

$$\sigma_i = \frac{1}{\mu - 1} \left[\sum_{\varphi=1}^{\mu} \left(m_{\varphi,i} - \frac{1}{\lambda} \sum_{\varphi=1}^{\mu} m_{\varphi,i} \right)^2 \right]^{\frac{1}{2}}, \quad (33)$$

$$i \in \{1, 2, \dots, N\}.$$

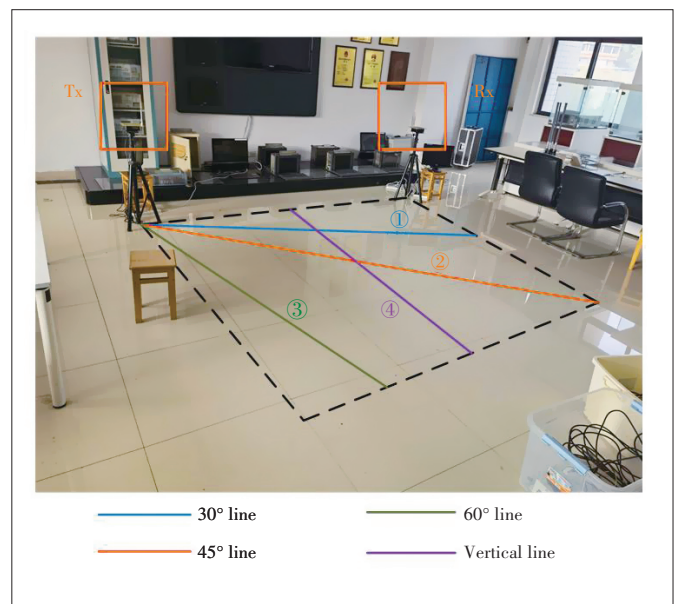
When the indices i satisfy $\sigma_i = 0$, we store them in a vector $\vec{p} = [p_1, p_2, \dots, p_N]$, where p refers to the indices of the detected peaks. Therefore, the sampling point in the matrix that satisfies its column standard deviation $\sigma_i = 0$ can be considered as the peak point.

4 Experimental Results

4.1 Experimental Environment

We collect 50 sets of data from two volunteers to validate the performance of our proposed system. The walking route is shown in Fig. 5.

The transmitter and receiver consist of two mini-PCs with one Intel 5300 NIC. The PCs are installed with 64-bit operating system of Ubuntu10.04. Specifically, one PC equipped with an external antenna works as transmitter, while the other with three works as receiver. The original CSI data is obtained through the open source tool Linux 802.11n CSI tool. The detailed installation process is shown in Ref. [5]. Our system works on Channel 149 at 5.745 GHz band and the packet transmission frequency is 1 000 Hz. At the same time, the transmitter is set to the injection mode and the receiver is set to the monitor mode.

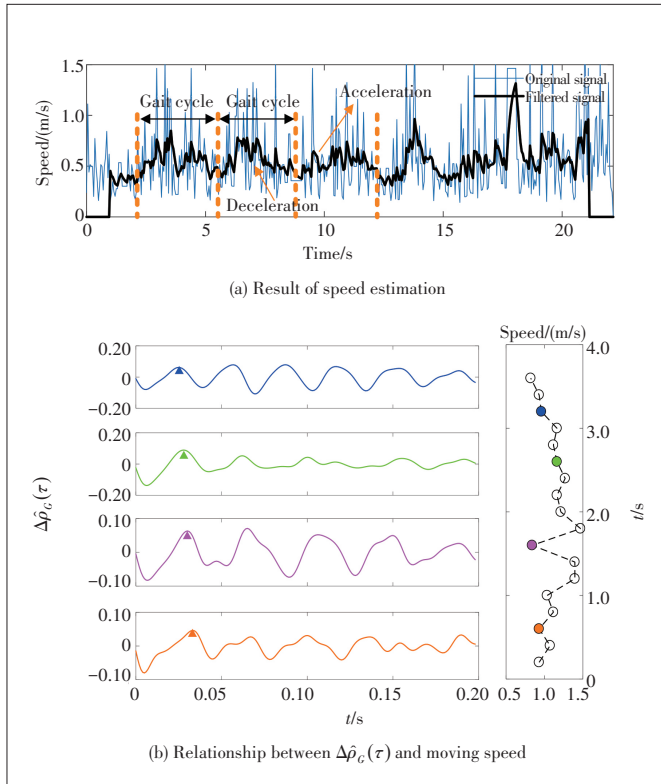


▲ Figure 5. Illustration of walking direction

The data collection is described as follows. Firstly, for single target speed estimation, we collected CSI data of two different volunteers walking on different routes with a consistent speed of 1 m/s respectively to analyze the influence of movement direction, and 50 samples are collected for each walking direction. Next, in order to further explore the influence of the moving targets number on the estimated results, we collect data of two, three and four volunteers walking on different routes, where one target walks along predefined routes and others walk randomly in the area as miscellaneous targets. In the experiment, we obtain ground truth via accelerometer-based solution. Specifically, a smartphone with accelerometer is installed on each tester's ankle to capture true speed measurement.

4.2 Analysis of Results

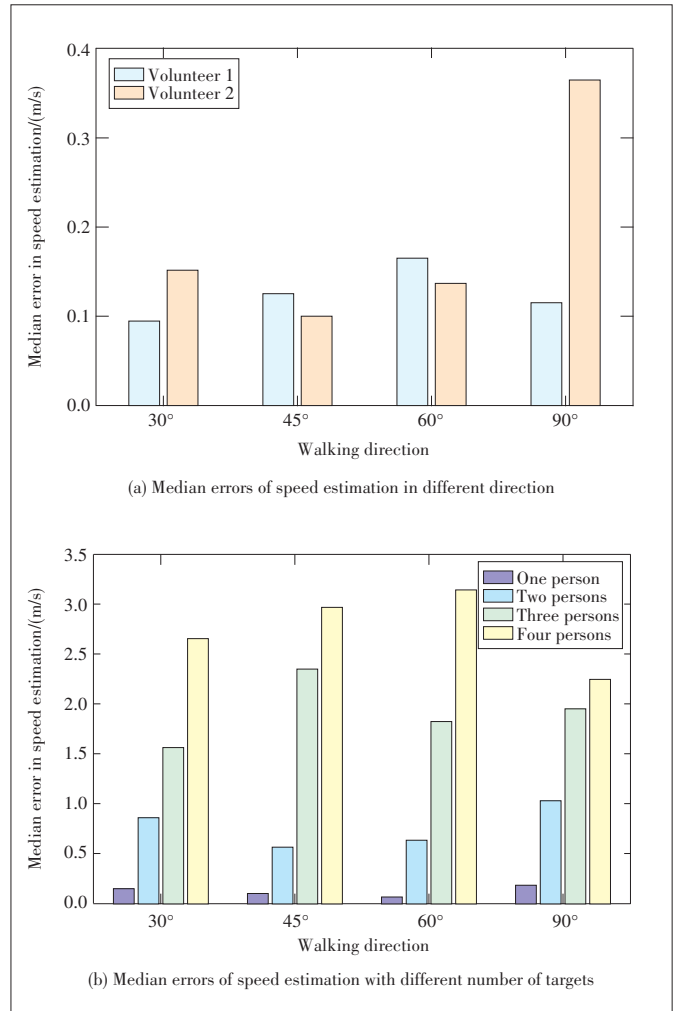
Since the theoretical assumption is feasible in a short time, in the step of the speed estimation algorithm, the maximum time interval τ is set as 0.2 s. The ACF of a sample is calculated every 0.05 s. As known for all, human walking is a periodic motion. Fig. 6a is an estimation result of walking speed, which shows a clear periodic acceleration and deceleration motion pattern. Although there are a large number of outliers in the original speed estimation, a Kalman filter can be applied to obtain a smoothed speed curve. In order to avoid the influence of external factors on the results in the process of data collection, the first second and last second data are set to zero. Mean-



▲ Figure 6. Speed estimation results

while, Fig. 6b shows the relationship between $\Delta\hat{\rho}_c(\tau)$ and moving speed. We choose four different time to calculate $\Delta\hat{\rho}_c(\tau)$ and the corresponding moving speed with different colors. We can conclude that although the ACFs are very different, the locations of the first peak of $\Delta\hat{\rho}_c(\tau)$ are highly consistent as long as the ACFs are calculated under similar walking speed. For each different sampling time in speed estimation, it can be calculated that the $\Delta\hat{\rho}_c(\tau)$ within 0.2 s, then we can get the first peak point in $\Delta\hat{\rho}_c(\tau)$ where $\tau_1 = 0.030$ s (orange), $\tau_2 = 0.033$ s (purple), $\tau_3 = 0.024$ s (green), $\tau_4 = 0.029$ s (blue), and the corresponding speed values are $v_1 = 0.93$ m/s, $v_2 = 0.84$ m/s, $v_3 = 1.16$ m/s, $v_4 = 0.96$ m/s.

Then, we verify the direction-independence of our speed estimation system. Fig. 7a shows the median error of speed estimation in different direction. From the figure, it can be seen that our proposed system achieves a mean absolute error of 0.18 m/s for device-free human walking speed estimation. Moreover, our proposed system achieves a consistent error



▲ Figure 7. Median errors of speed estimation in different direction

among different walking directions with the minimum error as 0.3 m/s and maximum error as 0.35 m/s (Volunteer 2 in 90°), which is enough in smart home environments.

Finally, we verify the influence of the number of moving targets in the environment. Fig. 7b shows the median error of speed estimation when there are multiple targets walking in the area. From the figure, it can be seen that the median error of speed estimation reaches 0.77 m/s for two targets, 1.92 m/s for three targets and 2.75 m/s for four targets. This is because the speed of moving target is calculated with the first local peak point of ACF, and the target with larger speed will contribute more to the ACF components. Thus, the system will take the target with larger speed as the final estimation result when there are multiple targets walking in the environment, which will lead to a significant reduction in the estimation accuracy.

5 Conclusions

In this paper, we propose a direction-independent indoor speed estimation system based on commercial Wi-Fi devices, which can be used in many fields, such as indoor security and intelligent identification. The system is designed to estimate the speed of a single moving object in the environment. If there exist multiple moving objects within the coverage of system, it would capture the highest speed among the objects. Firstly, we analyze the relationship between the electric field power and the CSI power, and then introduce the related concepts of optical angular spectrum to derive the ACF of the EM wave angular spectrum. Secondly, based on the speed model of EM wave, we derive directional processing in existing systems and summarize the robustness of direction independence. Finally, the AMPD peak detection algorithm is used to extract speed information from ACF. A large number of experiments are carried out in a typical indoor environment. With low cost, strong robustness and good real time performance, our system provides a new idea for speed estimation with wireless sense applications especially in smart home.

References

- [1] CHAO H, HE Y, ZHANG J, et al. GaitSet: Regarding gait as a set for cross-view gait recognition [C]//AAAI Conference on Artificial Intelligence. Hawaii, USA: AAAI, 2019: 8126 – 8133. DOI: 10.1609/AAAI.V33I01.33018126
- [2] MA H, LIAO W. Human gait modeling and analysis using a semi-Markov process with ground reaction forces [C]//International Conference of the IEEE Engineering in Medicine and Biology Society. Jeju Island, Korea: IEEE, 2017: 597 – 607. DOI: 10.1109/ICRA.2014.6906616
- [3] MUHAMMAD M, RENE M. Smartphone-based gait recognition: From authentication to imitation [J]. IEEE transactions on mobile computing. 2017, 16 (11): 1 – 1. DOI: 10.1109/TMC.2017.2686855
- [4] WANG W, LIU A X, MUHAMMAD S. Gait recognition using Wi-Fi signals [C]//ACM International Joint Conference on Pervasive and Ubiquitous Computing. Heidelberg, Germany: ACM, 2016: 363 – 373. DOI: 10.1145/2971648.2971670
- [5] ZHANG L, WANG C, MA M, et al. WiDGR: direction-independent gait recognition system using commercial Wi-Fi devices [J]. IEEE Internet of Things journal, 2020, 7(2), 1178 – 1191. DOI: 10.1109/JIOT.2019.2953488
- [6] ZENG Y, PARTH H, PRASANT M. WiWho: Wi-Fi-based person identification in smart spaces [C]//15th ACM/IEEE International Conference on Information Processing in Sensor Networks. Vienna, Austria: ACM/IEEE, 2016: 1 – 12. DOI: 10.5555/2959355.2959359
- [7] WANG W, LIU A X, MUHAMMAD S, et al. Understanding and modeling of Wi-Fi signal based human activity recognition [C]//21st Annual International Conference on Mobile Computing and Networking. Paris, France: ACM, 2015: 65 – 76. DOI: 10.1145/2789168.2790093
- [8] PETROV B M, DARIA T. Electromagnetic waves in rotating spherical cavities. E-Field [C]//Radiation and Scattering of Electromagnetic Waves. Divnomorskoe, Russia: IEEE, 2019: 12 – 15. DOI: 10.1109/RSEMW.2019.8792713
- [9] ZHANG F, CHEN C, WANG B, et al. WiSpeed: a statistical electromagnetic approach for device-free indoor speed estimation [J]. IEEE Internet of Things journal, 2018, 5(3): 2163 – 2177. DOI: 10.1109/JIOT.2018.2826227
- [10] WANG W, LIU A X, MUHAMMAD S, et al. Device-free human activity recognition using commercial Wi-Fi devices [J]. IEEE journal on selected areas in communications, 2017, 35(5): 1118 – 1131. DOI: 10.1109/JSAC.2017.2679658
- [11] KING C. Fundamentals of wireless communications [C]//67th Annual Conference for Protective Relay Engineers. College Station, USA: IEEE, 2014. DOI: 10.1109/REPCon.2013.6681855
- [12] CHEN C, CHEN Y, HAN Y, et al. Achieving centimeter-accuracy indoor localization on Wi-Fi platforms: A multi-antenna approach [J]. IEEE Internet of Things journal, 2017, 4(1): 111 – 121. DOI: 10.1109/JIOT.2016.2628701

Biographies

TIAN Zengshan is currently a professor of Chongqing University of Posts and Telecommunications, China. His research focuses on localization in a cellular network, personal communication, precise localization and attitude measurement with GPS, data compression, and deep learning.

YE Chenglin (s190101153@stu.cqupt.edu.cn) is currently pursuing the M. S. degree at Chongqing University of Posts and Telecommunications, China. His research interests include wireless sensing and localization.

ZHANG Gongzhui is currently pursuing the M. S. degree at Chongqing University of Posts and Telecommunications, China. His research interests include speed estimation and identity recognition.

HE Wei is an associate professor of Chongqing University of Posts and Telecommunications, China. His current research interests include wireless location, signal processing, mobile communication technology and communication software engineering.

JIN Yue is currently pursuing the Ph. D. degree at Chongqing University of Posts and Telecommunications, China. Her research interests include indoor intrusion detection and device-free tracking.



Analysis of Industrial Internet of Things and Digital Twins

TAN Jie^{1,2}, SHA Xiubin^{1,2}, DAI Bo^{1,2},
LU Ting^{1,2}

(1. ZTE Corporation, Shenzhen 518057, China;
2. State Key Laboratory of Mobile Network and
Mobile Multimedia, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202102007

[http://kns.cnki.net/kcms/detail/34.1294.
TN.20210525.1908.004.html](http://kns.cnki.net/kcms/detail/34.1294.TN.20210525.1908.004.html), published online
May 26, 2021

Manuscript received: 2021-02-01

Abstract: The industrial Internet of Things (IIoT) is an important engine for manufacturing enterprises to provide intelligent products and services. With the development of IIoT, more and more attention has been paid to the application of ultra-reliable and low latency communications (URLLC) in the 5G system. The data analysis model represented by digital twins is the core of IIoT development in the manufacturing industry. In this paper, the efforts of 3GPP are introduced for the development of URLLC in reducing delay and enhancing reliability, as well as the research on little jitter and high transmission efficiency. The enhanced key technologies required in the IIoT are also analyzed. Finally, digital twins are analyzed according to the actual IIoT situation.

Keywords: digital twins; industrial Internet of Things (IIoT); standards

Citation (IEEE Format): J. Tan, X. B. Sha, B. Dai, et al., "Analysis of industrial Internet of Things and digital twins," *ZTE Communications*, vol. 19, no. 2, pp. 53 – 60, Jun. 2021. doi: 10.12142/ZTECOM.202102007.

1 Introduction

In recent years, the fourth industrial revolution accelerated by the industrial Internet of things (IIoT) has raised a global upsurge^[1-2]. A cognitive IIoT system helps to establish the information relationship between the real world and the virtual space, which includes the perceptual layer (by perceptual control technology), transmission layer (by network communication technology), and application layer (by information processing technology)^[3]. The boom in IIoT cannot be achieved without technical support, including digital twins, edge computing, time-sensitive networks (TSN), and passive optical networks (PON).

With the application of IIoT, digital twins are endowed with new vitality. The concept of digital twins can be traced back to Dr. Michael GRIEVES in 2002. However, the unified concept of digital twinning has not been reached in the subsequent development, because different users have given different conceptual descriptions of digital twins based on different angles and needs. Digital twins mainly include the following technical features: digital representation, virtual-reality interconnection and data-driven. The IIoT extends the value and life cycle of digital twins, highlighting the advantages and capabilities of

digital twins in terms of models, data, and services. The application and iterative optimization of IIoT are becoming the incubator of digital twins^[4].

Based on the basic state of a physical entity, a digital twin enables a highly realistic analysis of the established model and collected data in a dynamic and real-time way, which is used for the monitoring, prediction and optimization of the physical entity. The digital twin creates a virtual model with a high degree of realism for a physical object, and simulates, analyzes and forecasts its behavior, which paves the way for the integration of information technology and manufacturing. In addition, as an edge-side technology, digital twins can effectively connect the perceptual layer and the transmission layer. Therefore, the industrial Internet platform is the incubator of digital twins, and the digital twin is important for industrial Internet platforms.

All kinds of data collection and exchange of physical entities may be realized in the IIoT. The advantages of the IIoT, such as resource aggregation, dynamic configuration, and supply and demand docking, will facilitate the integration and utilization of all kinds of resources. For example, the industrial Internet platform is used to associate the digital twin in the edge infrastructure in the downward direction, and transfer

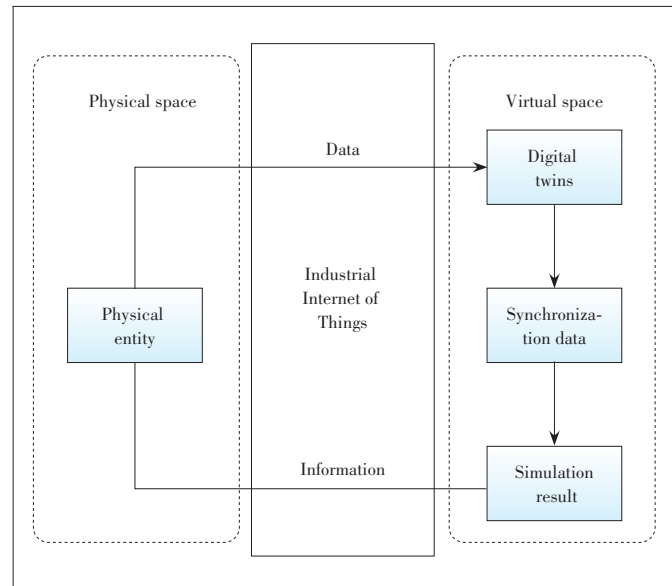
©International Telecommunication Union, 2020: This paper is based on "wireless technology and protocol for IIoT and digital twins" presented at "Industry-Driven Digital Transformation" ITU Kaleidoscope Academic Conference by ITU in 2020 and published in the proceedings of this conference.

and store the data in the cloud in the upward direction. Moreover, users can set up digital twins through platform services according to their own needs. It can be said that the industrial Internet platform has activated the life of digital twins. As shown in Fig. 1, a physical entity in the physical space and a digital twin in the virtual space are connected through the IIoT. Among them, the miniaturization of IIoT equipment makes the creation of digital twins possible and sensor systems are used to realize data sharing between virtual and physical objects. Furthermore, the emerging 5G technologies can provide a faster connection speed between virtual and real objects, and improve operation efficiency and reliability by reducing the response time.

The transmission layer in a cognitive IIoT framework mainly includes the short-distance wireless communication network, low-power wide area network, and industrial Ethernet. As we all know, the cellular 5G, Long Term Evolution Category Machine 1/Machine 2 (LTE CAT M1/M2), Long Range Radio Wide-Area Network (LoRaWAN), and Narrowband-Internet of Things (NB-IoT) are representative technologies in the Internet of Things (IoT), while the IIoT relies heavily on the availability of wireless connections^[5-6]. Considering that the features of classical field buses are incompatible with Internet features and their performance is not sufficient to transmit Internet packets, they cannot be directly included in the IIoT system. In particular, these classical networks do not support IIoT-based IPv6. However, they can be interconnected through gateway devices^[7-8]. It is quite challenging to introduce industrial networks into an IIoT system, because their applications often have stringent quality of service (QoS) requirements, which may be difficult to meet, such as configuration, robustness, reliability, latency, determinism, energy efficiency, battery lifetimes, and security^[9-11]. Refs. [12] and [13] report the suitability and achievable performance figures of industrial networks. As mentioned in Ref. [14], the low-power wide area network (LPWAN) is a new type of wireless network in IoT, which can be applied to indoor industrial monitoring^[15], intrusion detection^[16], remote monitoring and smart cities^[17] with the help of robust communications, wide coverage ranges and low power consumption.

IIoT will completely change the manufacturing industry by faster transmission speed, more efficient transmission and access to more data; examples are smart manufacturing, smart agriculture, smart cities, smart home, smart health care, smart transportation, etc. IIoT promotes the strong demand for more data acquisition, communications, real-time analysis and data-driven decision making in various industrial vertical fields.

With the rise of IIoT and the advent of Industry 4.0, the application of ultra-reliable and low latency communications (URLLC) technology has attracted more and more attention^[18]. URLLC is considered as a typical application scenario in 5G wireless communications^[19] and is generally regarded as the technical basis of new applications such as industrial automa-



▲ Figure 1. Relationship between digital twins and the industrial Internet of Things

tion, autonomous driving, IoT and tactile Internet. Therefore, the ultra-reliable communication and ultra-low delay required by URLLC have always been difficult in academia and industry. Refs. [20 - 23] have proposed some schemes to ensure the QoS requirements of URLLC. Multipath diversity was proposed in Ref. [20] to improve transmission reliability. A cross-layer optimization design was proposed in Ref. [21], in which a variety of factors affecting packet loss are considered in wireless access networks. Considering the tradeoff between QoS requirements and system throughput, the source coding rules for tactile data compression were studied in Ref. [22]. Besides, Ref. [23] proposed scheduling free uplink transmission to avoid scheduling delay.

Automation in different vertical domains has been developing rapidly. However, limited to the radio technology development, capital expenses (CAPEX) and operating expenses (OPEX), the communication technology applied in the vertical domains is mainly confined to the local area network, or even a network with wired connection. Although the cellular radio communication technology such as NB-IoT, Enhanced Machine-Type Communication (eMTC), and Long Range Radio (LoRa) can serve certain IoT use cases, it cannot satisfy all the requirements of any IoT use cases. For example, the requirements of the use cases in Refs. [24] and [25] cannot be served perfectly by the legacy cellular IoT system (Table 1). Generally, the use cases that cannot be satisfied by the legacy cellular IoT system always have rigorous requirements for low latency, high reliability, little jitter and/or frequent small data transmission. IIoT mainly focuses on providing wireless communications for these use cases.

The rest of this paper is organized as follows. Section 2 reviews the development of low latency and high reliability of

▼Table 1. Communication service performance requirements

Use case	Communication service availability: target value	Transfer interval: target value	Jitter
Motion control	99.999% to 99.99999%	500 μ s	<50% of E-to-E latency
100 Mbit/s wired-to-wireless link replacement	99.9999% to 99.999999%	≤ 1 ms	
Mobile robots	> 99.9999%	1 ms to 50 ms	< Transfer interval
Mobile control panels: remote control of assembly robots, milling machines, etc.	99.9999% to 99.999999%	4 ms to 8 ms	< 50% of interval
Mobile operation panel: motion control	99.999999%	1 ms	
Robotic aided surgery	> 99.999999%	1 ms	
Robotic aided diagnosis	> 99.999%	1 ms	

URLLC in the field of IIoT, as well as the research process of little jitter and high transmission efficiency. Section 3 discusses the possible future research directions and objectives in the field of IIoT. Section 4 discusses the cases and advantages of the application of IIoT for digital twins. Section 5 concludes the paper and shows the cooperation of IIoT technology and digital twin technology.

2 Development of URLLC for IIoT Technology

5G URLLC has two basic features: high reliability and low latency. With high reliability, its block error rate (BLER) reaches 10^{-5} or even 10^{-6} , and in terms of delay, it can implement 1 ms or even 0.5 ms air interface transmission delay. As one of the three application scenarios of 5G system, URLLC is widely used in various industries, such as Augmented Reality (AR)/Virtual Reality (VR) in the entertainment industry, industrial control system, transportation system, management of smart grid and smart home, and interactive telemedicine diagnosis^[26]. This paper will introduce the development of 3GPP Release15 (Rel-15) and Release16 (Rel-16) for URLLC in reducing delay and enhancing reliability, as well as the research on little jitter and high transmission efficiency.

2.1 Low Latency of URLLC

The 5G URLLC technology achieves a user plane delay of 0.5 ms in both the uplink and downlink between the gNB and the terminal. The delay refers to the time it takes to successfully transmit application layer IP packets/messages, specifically from the sender's 5G wireless protocol layer entry point, to the receiver's 5G wireless protocol layer exit point. Among them, the delay exists in both the uplink and downlink directions. The main technologies used by the 5G URLLC to implement low latency include: 1) introducing a smaller time resource unit, such as a mini slot; 2) the no-scheduling permission mechanism used for uplink access, with which the terminal can directly access the channel; 3) supporting an asynchronous process to save uplink time synchronization overhead; 4) adopting fast hybrid automatic repeat request (HARQ), fast dynamic scheduling, etc.

While 3GPP Rel-15 shows the research progress of URLLC delay, 3GPP further enhances on URLLC in the Rel-16 phase and proposes an improved delay reduction scheme^[27].

2.1.1 3GPP Rel-15

The study of URLLC delay supports for a more flexible frame structure. The 5G new radio (NR) supports the carrier spacing of 15 kHz in the LTE system. It also supports more spacing schemes, including 30 kHz, 60 kHz, 120 kHz, and 240 kHz. The higher the carrier spacing, the lower the delay performance. In addition, 5G NR supports frame structure adjustment. A slot is the minimum scheduling period. Compared with the LTE in which a fixed subframe includes 2 slots, the NR can flexibly switch between 1, 2 and 4 slots and configure uplink/downlink ratios, thus reducing the air interface transmission time of each slot.

The study of URLLC delay supports for more flexible scheduling units. The LTE includes a slot consisting of 14 symbols. However, the NR supports mini-slots. Mini-slots support the length of 2 symbols, 3 symbols, and 4 symbols, and a shorter slot can reduce the feedback delay.

The study of URLLC delay supports for flexible PDCCH configuration. The search space consists of a group of candidate physical downlink control channels (PDCCH), and the search space can be configured with parameters such as search type, period, slot offset, number of slots, CORESET, and downlink control information (DCI) format. By configuring a reasonable monitoring period and offset of the PDCCH in a slot, the PDCCH monitoring opportunity can be achieved densely. The slot has multiple PDCCH monitoring moments, which can meet the requirements for the burst service scenarios of the URLLC and meet the requirements for low latency.

The study of URLLC delay supports for URLLC high-priority transmission. To meet the URLLC service requirement of high priority, 5G/NR proposes that a URLLC service can preempt enhanced mobile broadband (eMBB) service resources to reduce the delay.

The study of URLLC delay introduces the function of mobile edge computing (MEC). In a 5G network, the user-plane function (UPF) can be deployed on the user side. The edge computing server and the UPF are co-located. The UPF recog-

nizes that the destination address of the service flow is local, so it distributes the service to the local edge computing server for service processing, which reduces the redundant transmission path of the service and the delay.

2.1.2 3GPP Rel-16

The study of URLLC delay supports for grant-free configuration. In the scheduling based on grant configuration, user equipment (UE) needs to obtain resources through the scheduling request. In order to reduce the delay, the resource can be pre-allocated to UE according to service characteristics^[28].

The study of URLLC delay supports for intra-UE priority and multiplex mechanism. In Rel-15, eMBB dynamic grant takes precedence over URLLC configured grant (CG). In order to ensure the delay of a URLLC service, Rel-16 proposes a selection scheme based on logical channel prioritization (LCP) to transmit URLLC services with a higher priority.

The study of URLLC delay supports for time sensitive network (TSN) and 5G convergence. The following mechanisms are adopted to realize TSN: 1) supporting semi-persistent scheduling (SPS) with a shorter period; 2) supporting the configuration of multiple SPS and CG for a bandwidth part (BWP) of UE; 3) supporting TSN services that do not match the CG/SPS period.

2.2 High Reliability of URLLC

At present, the reliability index of 5G URLLC is 99.999% for a 32-byte packet at the user plane with a delay of 1 ms. If the delay allows, 5G URLLC can also use the retransmission mechanism to further improve the success rate. In terms of improving system reliability, 5G URLLC adopts the following technologies: 1) adopting a more robust multi-antenna transmit diversity mechanism; 2) adopting robust coding and modulation in order to reduce the bit error rate; 3) adopting super robust channel state estimation. The reliability of URLLC has been enhanced in Rel-15 and Rel-16 of 3GPP.

2.2.1 3GPP Rel-15

The study of URLLC reliability supports for PDCP duplication mechanism. The sender replicates the data at the PDCP layer and then sends the two duplications to two independent logical channels for transmission, so as to achieve the frequency diversity gain and improve reliability.

The study of URLLC reliability supports for optimizing MCS/CQI tables. The modulation and coding scheme (MCS) and channel quality indication (CQI) of the LTE cannot meet the requirements of NR for system reliability and transmission rate. Therefore, NR adds two lower bit rates in the CQI table, and the corresponding gNB adds two MCS low-frequency options. A lower bit rate can be chosen between the UE and the gNB to ensure reliability.

The study of URLLC reliability supports for less load DCI design. By reducing the DCI overhead and improving the aggregation level, the PDCCH encoding rate is reduced. The de-

coding error rate is reduced and the reliability is improved.

2.2.2 3GPP Rel-16

The study of URLLC reliability supports for multi-TRP transmission mode. Rel-16 proposes that transmission blocks can be transmitted repeatedly based on space division, frequency division, intra-slot division and inter-slot division based on Rel-15. In order to improve the diversity gain, it also supports the combination of the above modes and the dynamic handover between different modes (including combined modes).

The study of URLLC reliability supports for PDCP duplication enhancement mechanism. Rel-15 supports two-branch PDCP duplication, in order to achieve higher reliability. Rel-16 supports up to four-branch PDCP duplication. This mechanism can be implemented through carrier aggregation (CA) duplication, dual connectivity (DC) duplication and the combination of CA duplication and DC duplication.

The study of URLLC reliability supports for redundant transmission scheme. NG-RAN duplicates uplink packets and sends them to the UPF via two redundant link (N3 interface) channels, where each N3 channel is associated with a PDU session, and two independent N3 channels are established to transmit data. The gNB, SMF and UPF will provide different routes for the two links^[29].

2.3 Little Jitter of URLLC

Requirements of time accuracy are typically specified with two values: the characteristic time and jitter. Characteristic time is the target value of the time parameter, e.g. end-to-end latency. The jitter is the variation of a (characteristic) time parameter and the maximum deviation of a time parameter relative to a reference or target value.

As depicted in Ref. [25], power distribution poses the jitter requirements and the traffic pattern is deterministic as well. In such a case, the maximum value of the characteristic time parameter needs to be known. Sometimes, a minimum value may also be given, and should not be undershot. A minimum value is only used in particular use cases, for instance, when putting labels at a specific location on moving objects. In Rel-15, a common understanding in radio access network (RAN) is that a delay-sensitive URLLC service with periodic traffic can be accommodated by the semi-persistent CG. That means the periodicity of the traffic should be a prerequisite in RAN to meet the data size and jitter requirements. An example is the variation of the end-to-end latency. If not stated otherwise, jitter specifies the symmetric value range around the target value (target value \pm jitter/2). If the actual time value is outside this interval, the transmission will not be successful. Ref. [25] shows an example of transmissions with jitter. It should be noted that the end-to-end latency may scatter even for successful transmissions.

As an important feature of TSN, the jitter requirement is to

provide a deterministic service with bounded delay. Typical characteristic parameters to which jitter values are ascribed are transfer interval, end-to-end latency, and update time. Further, the buffering mechanism is held and forwarded to eliminate jitter in the TSN. Device-side TSN translator (DS-TT) and network-side TSN translator (NW-TT) support a hold and forward mechanism to schedule traffic as defined in IEEE 802.1Qbv^[30], if 5GS is to participate transparently as a bridge in a TSN network. The hold and forward buffering mechanism allows packet delay budget (PDB) based 5GS QoS to be used for time-sensitive communication (TSC) traffic since packets need only arrive at NW-TT or DS-TT egress prior to their scheduled transmission time. The way that TSN translator supports the hold and forward mechanism depends on the implementation.

In addition, time synchronization precision is defined between a synchronization master and a synchronization device. The detailed objectives for NR TSC-related enhancements include specifying accurate reference timing delivery from gNB to the UE using broadcast and unicasting radio resource control (RRC) signaling for the synchronization requirements defined in Ref. [31]. To meet the high-precision time synchronization requirements of the TSN, a high-precision reference time transmission mechanism is introduced to NR. Broadcast messages (SIB9) or dedicated RRC messages (DLInformation-Transfer messages) with the high-precision time can be sent. The time granularity is enhanced from 10 ms to 10 ns. According to the simulation result of radio access network work group 1 (RAN 1), radio access network work group 2 (RAN 2) assumes that delay compensation is required in the scenario where the service range is greater than 200 m for the user with the subcarrier interval of 15 kHz. However, in Rel-16, RAN 1 only provides transmission delay compensation for the base station and UE in the Time Division Duplex (TDD) and Frequency Division Duplex (FDD) scenarios according to half of the timing advance, that is, $N_{TA} \times Tc/2$. In addition, although RAN 1 discusses a lot about when and how to implement the transmission delay supplement, this topic has not finished in the Rel-16 phase.

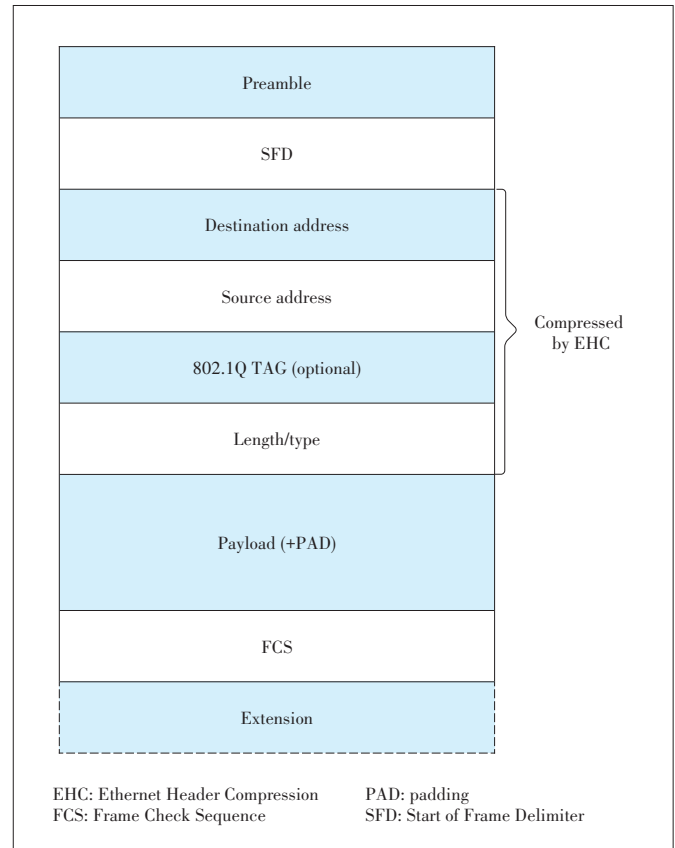
2.4 High Transmission Efficiency of URLLC

In the field of IIoT, the small packets of TSN are transmitted frequently in the ordinary communication network. TSN is also introduced into the 5G system, which has the characteristics of small packets with frequent transmission, low latency and high reliability. In this case, reducing the packet overhead can effectively improve the effective utilization of system bandwidth. Therefore, for TSN packets, header compression can be used to further reduce the size of data packets, thus saving the wireless resources used by a single packet and improve the utilization of wireless resources. The data stream transmitted by TSN is mainly an Ethernet data packet, so the Ethernet header compression (EHC) is introduced to reduce

the overhead caused by Ethernet header transmission.

EHC may be particularly beneficial when the payload size of an Ethernet frame is small relative to the overall size of the frame, which is typical in an Ethernet-based IIoT network. The EHC protocol compresses the Ethernet header as shown in Fig. 2. The fields that are compressed by the EHC protocol are Destination Address, Source Address, 802.1Q Tag, and Length/Type. The fields Preamble, Start of Frame Delimiter (SFD), and Frame Check Sequence (FCS) are not transmitted in a 3GPP system, and thus not considered in an EHC protocol. There may be more than one 802.1Q Tag field in the Ethernet header, and all are compressed by the EHC protocol. The padding is not compressed by the EHC protocol. The EHC compressor and the EHC decompressor store original header field information as an “EHC context”. Each EHC context is identified by a unique identifier, called Context ID (CID). For an Ethernet packet stream, the EHC compressor establishes the EHC context and associates it with the CID. Then, the EHC compressor transmits the “Full Header (FH)” packet to the EHC decompressor including the associated CID. The EHC compressor keeps transmitting the FH packets until the EHC feedback is received from the EHC decompressor.

The source Medium Access Control (MAC) address, destination MAC address and type fields of the Ethernet header



▲ Figure 2. Ethernet packet format

frame are all static and can be compressed, which is also the conclusion of the 3GPP RAN 2. Header compression is essentially the use of CID instead of Ethernet headers for transmission in the communication network.

The compression process includes: 1) The configuration of the CID is completed by the PDCP data PDU; 2) the compressor sends a full packet containing the full header and the CID; 3) the decompressor establishes the relationship between the header and the CID according to the received FH packet; 4) after the decompressor successfully establishes the context relationship, it sends the feedback message with the CID, informing the compressor that it can send the compression package; 5) after receiving the feedback information carrying CID, the compressor starts to send the compression packet corresponding to the CID. The header field in the compression packet is replaced by the CID, where the feedback information is sent by PDU control PDCP.

3 Future Research Directions of IIoT Technology

As a breakthrough in the integration of mobile communications and vertical industries, 5G is expected to bring about great changes to the whole society through URLLC services such as automatic driving, factory automation and smart grid. The research work for the 3GPP Rel-16 has been completed. At this stage, the research in related high-level technologies has laid a solid foundation for the integration of 5G and IIoT. With the development of IIoT, its services require more strict latency and are more deterministic, for example, up to 99.999999%. To this end, 3GPP has launched its research on standard technology in Rel-17 to further enhance the low latency and reliability, so that 5G can meet the various needs of the development of IIoT.

Considering the directions of the reliable private network of local service, extensible wireless connections of future platforms and multiple functions of new cases, a shorter frame structure will be selected in accordance with service development requirements to reduce the air interface delay of services. In addition, deep integration with the TSN proposed by the industry will also be considered to guarantee low transmission delay of services. On the road of the integration of 5G and industrial Internet TSN, related enhancements will be carried out in the following aspects: 1) Based on multi-carrier deployment, the reliability is further improved by introducing a PDCP layer and higher-level data replication transmission technology; 2) the feedback scheme for the physical layer will be further enhanced; 3) the UE service priority and uplink UCI will be enhanced; 4) with the NR-Unlicensed (NR-U), the 5G NR will support the licensed frequency, shared frequency domain and license-free spectrum; 5) URLLC will implement low latency, high reliability and multi-TRP cooperation; 6) mobility will be enhanced; 7) wired bus will be replaced by

wireless bus; 8) requirements of network and equipment positioning, positioning in IIoT, and intelligent factory/vehicle-to-everything (V2X) centimeter positioning will be met; 9) with the penetration of artificial intelligence (AI) technology, new deterministic requirements and key standard technology of AI in IIoT applications will be explored. These will be the key research directions in the future.

4 Digital Twins in IIoT

With the continuous development of the manufacturing industry, digital twins have become the focus of every digital enterprise, although they have not yet become the mainstream technology. The core of digital twins is model and data, but the creation of virtual models and data analysis require professional knowledge. For those who do not have relevant knowledge, it is a long way to go to build and use digital twins. IIoT can just solve the above problems, through the platform to achieve data analysis outsourcing, model sharing and other services. For example, the IIoT can be used to associate the edge-side infrastructure downward with the digital twin, and to transfer and store data upward in the cloud. Any users can establish digital twins through the IIoT services according to their own needs. It can be said that the industrial Internet platform activates the life of digital twins.

IIoT is a key link in the process of enterprise digital transformation, which accelerates the integration of various elements of information technology (IT) and operation technology (OT). Data is the most important binder in the integration process. In order to make the IT and OT integrate better, the hidden asset of data should be handled first. In addition, the IIoT is trying to break the boundaries of enterprises, trying to fill the gaps between IT and OT, and creating a new ecology of software definition, data-driven and mode innovation. The digital twin just provides the interface of data and technology for the development of such integration.

As we know, when each object (such as cars, airplanes, factories and people) in the real world has a digital twin, the space-time relationship between the digital twins becomes more valuable than a single digital twin. When the interactions between objects are optimized at the same level of a system, compared with the partial or independent optimization of the system, the efficiency is greatly improved. But in order to realize the optimization of the whole system, communication becomes a crucial factor. As described in Ref. [25], most of the communication technologies currently used in industry are still wired. However, with the advent of Industry 4.0 and 5G, this may change fundamentally, since only wireless connectivity can provide the degree of flexibility, mobility, versatility and ergonomics that are required for the factories of the future. Therefore, with the support of IIoT, the digital twin technology has been further promoted in different application fields of a "future factory", such as factory automation, pro-

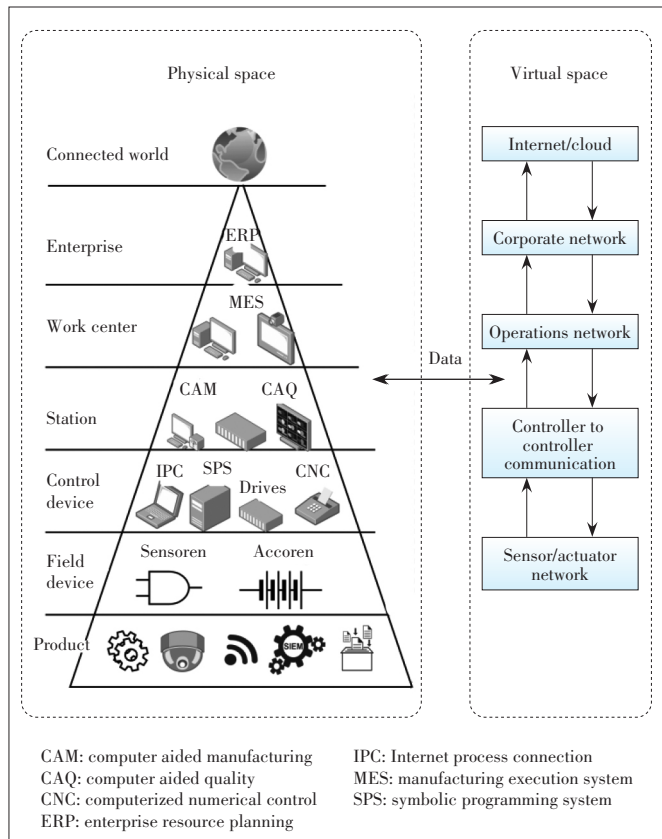
cess automation, hazardous material information system (HMIS) and production, logistics and warehouse, monitoring and maintenance. Fig. 3 contains enterprise resource planning (ERP), manufacturing execution system (MES), computer aided quality (CAQ), computer aided manufacturing (CAM), Internet process connection (IPC), symbolic programming system (SPS), computerized numerical control (CNC), and so on. Fig. 3 also shows the objects in the real production process, such as robots, cameras, mechanical arms, workbenches and mechanical tools. The information status of these objects in wired and wireless networks is uploaded to the Internet/cloud through sensors. After a high real-time data exchange, the state of a real object is simulated by a digital twin, and then the digital twin obtains the corresponding fault diagnosis results, evaluation and prediction results, behavior control of machinery, and other information. The high real-time network communication and the corresponding simulation information are used to control the production process. The high real-time data exchange between the real object and digital twin is the basis and prerequisite for the application of the digital twin technology. In addition, as an effective way to solve the interaction theory and implementation method of the physical world and the information world in the future, digital twins are gradually deepening in practice. This also means that the demand for network performance in its development is constant-

ly improving, and it also promotes the improvement of IIoT related project research and standards. Thus, it is necessary to further study the low latency and high reliability of the IIoT.

Therefore, digital twins can display, predict and analyze the interaction between a digital model and the physical world. The design based on digital twins is based on virtual mapping of existing physical products. A large amount of data is studied to obtain valuable knowledge for product innovation. Designers only need to publish the requirements to the industrial Internet platform, so that platform managers can precisely match the data services needed by designers, as well as the model and algorithm services for data processing. Digital twins are effectively applied to product design through services to reduce the modification caused by the inconsistency between expected behavior and design behavior. It greatly shortens the design cycle and reduces the design cost.

5 Conclusions

The development of IIoT technology makes the application scenarios of digital twin technology more extensive, fully demonstrating the advantages and capabilities of digital twins in terms of models, data, and services. As a new digital technology solution and human-computer interaction interface, it will effectively promote and deepen the further development of business models. On the basis of reviewing the development of URLLC and analyzing the future research direction in the field of IIoT, this paper discusses the application of IIoT in digital twin. The development of digital twin technology is closely related to the continuous evolution of URLLC technology in IIoT. In other words, the development requirement of digital twins also promotes the improvement of IIoT standards and technologies. Compared with the previous cellular mobile communication technology, the 5G URLLC technology has greatly improved in terms of delay, reliability, little jitter and high transmission efficiency. However, with the development of IIoT, the research of delay and reliability technology needs to be further enhanced to meet the needs of various services.



▲ Figure 3. An example use case of industrial IoT via digital twins

References

- [1] SISINNI E, SAIFULLAH A, HAN S, et al. Industrial Internet of Things: challenges, opportunities, and directions [J]. IEEE transactions on industrial informatics, 2018, 14(11): 4724 – 4734. DOI: 10.1109/TII.2018.2852491
- [2] TAYEB S, LATIFI S, KIM Y. A survey on IIoT communication and computation frameworks: an industrial perspective [C]/IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC). Las Vegas, USA: IEEE, 2017: 1 – 6. DOI: 10.1109/CCWC.2017.7868354
- [3] CHEN B T, WAN J F, LAN Y T, et al. Improving cognitive ability of edge intelligent IIoT through machine learning [J]. IEEE network, 2019, 33(5): 61 – 67. DOI: 10.1109/MNET.001.1800505
- [4] SAVAZZI S, RAMPA V, SPAGNOLINI U. Wireless cloud networks for the fac-

- tory of things: connectivity modeling and layout design [J]. IEEE Internet of Things journal, 2014, 1(2): 180 – 195. DOI: 10.1109/JIOT.2014.2313459
- [5] MUMTAZ S, ALSOHAILY A, PANG Z B, et al. Massive Internet of Things for industrial applications: addressing wireless IIoT connectivity challenges and ecosystem fragmentation [J]. IEEE industrial electronics magazine, 2017, 11(1): 28 – 33. DOI: 10.1109/MIE.2016.2618724
- [6] PERERA C, LIU C H, JAYAWARDENA S. The emerging Internet of Things marketplace from an industrial perspective: a survey [J]. IEEE transactions on emerging topics in computing, 2015, 3(4): 585 – 598. DOI: 10.1109/TETC.2015.2390034
- [7] BELLAGENTE P, FERRARI P, FLAMMINI A, et al. Enabling PROFINET devices to work in IoT: characterization and requirements [C]/IEEE International Instrumentation and Measurement Technology Conference Proceedings. Taiwan, China: IEEE, 2016: 1 – 6. DOI: 10.1109/I2MTC.2016.7520417
- [8] SAUTER T, LOBASHOV M. How to access factory floor information using Internet technologies and gateways [J]. IEEE transactions on industrial informatics, 2011, 7(4): 699 – 712. DOI: 10.1109/TII.2011.2166788
- [9] BARTOLOMEU P, ALAM M, FERREIRA J, et al. Supporting deterministic wireless communications in industrial IoT [J]. IEEE transactions on industrial informatics, 2018, 14(9): 4045 – 4054. DOI: 10.1109/TII.2018.2825998
- [10] QIU T, ZHANG Y S, QIAO D J, et al. A robust time synchronization scheme for industrial Internet of Things [J]. IEEE transactions on industrial informatics, 2018, 14(8): 3570 – 3580. DOI: 10.1109/TII.2017.2738842
- [11] KOUTSIAMANIS R A, PAPADOPOULOS G Z, FAFOUTIS X, et al. From best effort to deterministic packet delivery for wireless industrial IoT networks [J]. IEEE transactions on industrial informatics, 2018, 14(10): 4468 – 4480. DOI: 10.1109/TII.2018.2856884
- [12] SAEZ M, MATURANA F P, BARTON K, et al. Realtime manufacturing machine and system performance monitoring using Internet of Things [J]. IEEE transactions on automation science and engineering, 2018, 15(4): 1735 – 1748. DOI: 10.1109/TASE.2017.2784826
- [13] FUCHS S, SCHMIDT H P, WITTE S. Test and online monitoring of realtime Ethernet with mixed physical layer for Industry 4.0 [C]/IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA). Berlin, Germany: IEEE, 2016: 1 – 4. DOI: 10.1109/ETFA.2016.7733518
- [14] RAZA U, KULKARNI P, SOORIYABANDARA M. Low power wide area networks: an overview [J]. IEEE communications surveys & tutorials, 2017, 19(2): 855 – 873. DOI: 10.1109/COMST.2017.2652320
- [15] IQBAL Z, KIM K, LEE H N. A cooperative wireless sensor network for indoor industrial monitoring [J]. IEEE transactions on industrial informatics, 2017, 13(2): 482 – 491. DOI: 10.1109/TII.2016.2613504
- [16] SHIN S, KWON T, JO G Y, et al. An experimental study of hierarchical intrusion detection for wireless industrial sensor networks [J]. IEEE transactions on industrial informatics, 2010, 6(4): 744 – 757. DOI: 10.1109/TII.2010.2051556
- [17] MAGRIN D, CENTENARO M, VANGELISTA L. Performance evaluation of LoRa networks in a smart city scenario [C]/IEEE International Conference on Communications (ICC). Paris, France: IEEE, 2017: 1 – 7. DOI: 10.1109/ICC.2017.7996384
- [18] YANG S X, GAO Y H, ZHANG X, et al. 5G NR multiplexing eMBB and URLLC [J]. Telecom engineering techniques and standardization, 2018, 31(8): 23 – 28
- [19] 3GPP. Study on scenarios and requirements for next generation access technologies (release 14): 3GPP. TR 38.913 [S]. 2016
- [20] NIELSEN J J, LIU R K, POPOVSKI P. Ultrareliable low latency communication using interface diversity [J]. IEEE transactions on communications, 2018, 66(3): 1322 – 1334. DOI: 10.1109/TCOMM.2017.2771478
- [21] SHE C Y, YANG C Y, QUEK T Q S. Crosslayer optimization for ultrareliable and lowlatency radio access networks [J]. IEEE transactions on wireless communications, 2018, 17(1): 127 – 141. DOI: 10.1109/TWC.2017.2762684
- [22] TAKEYA M, KAWAMURA Y, KATSURA S. Data reduction design based on deltasigma modulator in quantized scalingbilateral control for realizing of haptic broadcasting [J]. IEEE transactions on industrial electronics, 2016, 63(3): 1962 – 1971. DOI: 10.1109/TIE.2015.2512233
- [23] SINGH B, TIRKKONEN O, LI Z X, et al. Contention based access for ultrareliable low latency uplink transmissions [J]. IEEE wireless communications letters, 2018, 7(2): 182 – 185. DOI: 10.1109/LWC.2017.2763594
- [24] 3GPP. Service requirements for cyberphysical control applications in vertical domains: 3GPP TS 22.104 [S]. 2019
- [25] 3GPP. Study on communication for automation in vertical domains: 3GPP TS 22.804 [S]. 2018
- [26] XU S, XIN J. Study on 5G URLLC enhancement technology for ultrareliable and low latency communications [J]. Mobile communications, 2019, 43(9): 62 – 67
- [27] 3GPP TSG RAN WG1. Evaluation of URLLC factory automation scenario at 30 GHz: Meeting#96 R11903448 [R]. Sophia Antipolis, France: 3GPP, 2019
- [28] 3GPP. Physical channels and modulation (Release 16): 3GPP TS 38.211 [S]. 2020
- [29] 3GPP TSGRAN. Revised SID: Study on NR Industrial Internet of Things (IIoT): Meeting#81 RP182090 [R]. Sophia Antipolis, France: 3GPP, 2018
- [30] IEEE. IEEE standard for local and metropolitan area networks bridges and bridged networks amendment 25: enhancements for scheduled traffic: 802.1Qbv [S]. 2015
- [31] 3GPP. Radio resource control (RRC) protocol specification (Release 16): 3GPP TS 38.331 [S]. 2020

Biographies

TAN Jie received his M.S. from Nanjing University of Posts and Telecommunications, China. He has been engaged in 3GPP standard pre-research for NR and IIoT TSN technology at ZTE Corporation since 2019.

SHA Xiubin (sha.xiubin@zte.com.cn) received his M.S. from Southwest Jiaotong University, China in 2001. He works with ZTE Corporation and has engaged in the radio resource management algorithm design and simulation for CDMA, TD-SCDMA and WCDMA products from 2001 to 2015 and in the 3GPP standard pre-research for NB-IoT, eMTC and IIoT TSN technology from 2015 to now.

DAI Bo received his master's degree from Harbin Institute of technology, China. He has been mainly engaged in the research of 4G, 5G and 6G technologies and participated in the standardization work of LTE, NB-IOT, LTE-MTC, NR IIoT, NR redcap UE, etc.

LU Ting received her Ph.D. from South China University of Technology, China in 2003. She has engaged in telecommunication product development for about 5 years and international telecommunication standards for about 13 years. She has comprehensive experience of 3G/4G/5G wireless systems and technologies, especially NB-IoT, LTE/eMTC, IIoT TSN, and NTN.



Higher Speed Passive Optical Networks for Low Latency Services

Abstract: Latency sensitive services have attracted much attention lately and imposed stringent requirements on the access network design. Passive optical networks (PONs) provide a potential long-term solution for the underlying transport network supporting these services. This paper discusses latency limitations in PON and recent progress in PON standardization to improve latency. Experimental results of a low latency PON system are presented as a proof of concept.

Keywords: passive optical networks; time-division multiple access; wavelength-division multiple access; low latency

ZHANG Weiliang, YUAN Liquan

(Wireline Product Planning Department, ZTE Corporation, Shanghai 201203, China)

DOI: 10.12142/ZTECOM.202102008

<http://kns.cnki.net/kcms/detail/34.1294.TN.20210510.1514.002.html>, published online May 10, 2021

Manuscript received: 2021-03-16

Citation (IEEE Format): W. L. Zhang and L. Q. Yuan, "Higher speed passive optical networks for low latency services," *ZTE Communications*, vol. 19, no. 2, pp. 61 – 66, Jun. 2021. doi: 10.12142/ZTECOM.202102008.

1 Introduction

With the continued growth of new applications and services over fixed and wireless communication networks, requirements for low latency, high bandwidth, timing and synchronization have taken central stage in new network architecture design. In 2015, the ITU-R laid out its vision on the framework and objectives of the development of International Mobile Telecommunications (IMT) for 2020 and beyond^[1]. In the IMT-2020 Recommendation, diverse services for ultra-reliable low-latency communications (URLLC), as well as for enhanced mobile broadband (eMBB) and massive machine type communications (mMTC), are envisioned.

These envisioned services translate to stringent requirements for the underlying transport network layer. Many fixed network technologies are being considered for the 5G transport infrastructure, e. g., point-to-point fibers, active wavelength division multiplexing (WDM), and passive optical networks (PONs).

Among these options, the PON stands out as a highly suitable choice. Due to its efficient fiber infrastructure and bandwidth efficiency, the time-division-multiplexed (TDM) PON has been successfully deployed worldwide to over 626 million subscribers as of December 2019. Many of the PON and 4G

devices share the same access office. In addition, PON and 5G transport networks share similar network topology. It is therefore attractive to make use of the abundant fiber resources in the PON infrastructure.

Many industry standards development organizations (SDOs) are working on new standardization projects to address the increasing demands of low-latency services^[2]. The Full Service Access Network (FSAN) group^[3] and the ITU-T Study Group (SG) 15 Question 2 (Q2), which focuses on optical access networks standardization, have conducted several projects to study PONs for 5G mobile x-haul transport, which will be discussed later in the paper.

This paper is structured as follows. We will start with examples of low latency services in Section 2. An overview of passive optical networks and their latency properties are discussed in Section 3, which is followed by recent progress in PON standards to support low-latency services in Section 4. Finally in Section 5, we describe in more detail a latency reduction method for TDM PON and present experimental results as a proof of concept.

2 Low Latency Services

Low latency services are characterized as services of which

the latency requirement between service end points is much lower than the latency needed in traditional services, e.g., internet browsing, files/data download, and IPTV. In these traditional services, there is often no clear boundary of time limitation. Some of the most prominent examples of low latency services include 5G x-haul transport for supporting URLLC, virtual reality (VR)/augmented reality (AR) video services, industry applications for factory networks, and robotic control^[4].

A 5G x-haul transport network is the underlying transport layer providing fronthaul connectivity between the 5G remote unit (RU) and the distributed unit (DU), midhaul connectivity between DU and the centralized unit (CU), and backhaul connectivity between CU and the 5G core. More details of its latency requirements will be described in Section 4.

VR/AR video services provide immersive viewing experience for end users. They require the data traffic between the server and clients be transported in very short duration to meet the latency requirements of the motion-to-photon (MTP) and motion-to-audio. The maximum end-to-end latency is 20 ms, of which a much lower value is allotted for the access network segment, e.g., the PON.

Industry applications are much more complex than 5G x-haul and AR/VR video services. The requirements of timing and latency for the control messages are much more critical than normal services in some cases. For example, for industrial robotic control, there is a need to synchronize the robots with each other, which could limit the latency to less than 10 ms in specific scenarios discussed by the ETSI F5G group. Currently, the ETSI F5G group is studying network design on how to use PON technology for industry applications.

3 Overview of Passive Optical Networks

In this section, we provide an overview of three types of PONs and discuss their latency properties: TDM PON, WDM PON, and time and wavelength division multiplexed (TWDM) PON. TDM and TWDM PONs are mainly used for residential services such as Fiber to the Home, while WDM PONs are for business services due to the higher cost. Therefore, these systems have quite different latency requirements.

3.1 TDM PON

In a TDM PON, as shown in Fig. 1, signals from the optical line terminal (OLT) are broadcasted downstream to all the optical network units (ONUs) in a TDM fashion. In the upstream direction, each ONU transmits its signal in a time slot assigned by the OLT through the dynamic bandwidth allocation (DBA) process.

As specified in the ITU-T G.989.3 Recommendation^[5], the DBA engine consists of a bandwidth assignment component and a bandwidth map (BWmap) generation component. The bandwidth assignment component computes the assigned bandwidths for every DBA cycle. The assigned bandwidths are

then supplied to the BWmap generator to generate a BWmap once every physical layer (PHY) frame (125 μ s).

In a typical TDM PON, the DBA process could result in an upstream latency in the order of several milliseconds because data transmission may take several DBA cycles to complete. When using a conventional TDM PON for wireless x-haul transport, the upstream data from DU and/or RU must wait in the ONU until the completion of DBA.

Another latency causing process for a TDM PON is the need of a quiet window during the activation and registration of new ONUs onto an operational TDM PON. No ONU can transmit upstream data during this period, which can last over a few 100 μ s. Such a latency value is incompatible with low latency mobile x-haul applications.

3.2 WDM PON

WDM PON is a logical point-to-point system, as shown in Fig. 2. Signals from the OLT, each transmitted over a different wavelength channel, are combined in a wavelength multiplexer before transmitting to the end user. In the optical distribution network (ODN), a wavelength splitter routes the individual wavelengths to different ONUs. In the case of WDM PON for mobile x-haul services, each of the ONU can be connected to an RU supporting one of the three sectors of an antenna at the cell site. The latency is purely limited by the transmission distance and processing delay. Therefore, no special latency improving mechanism is needed.

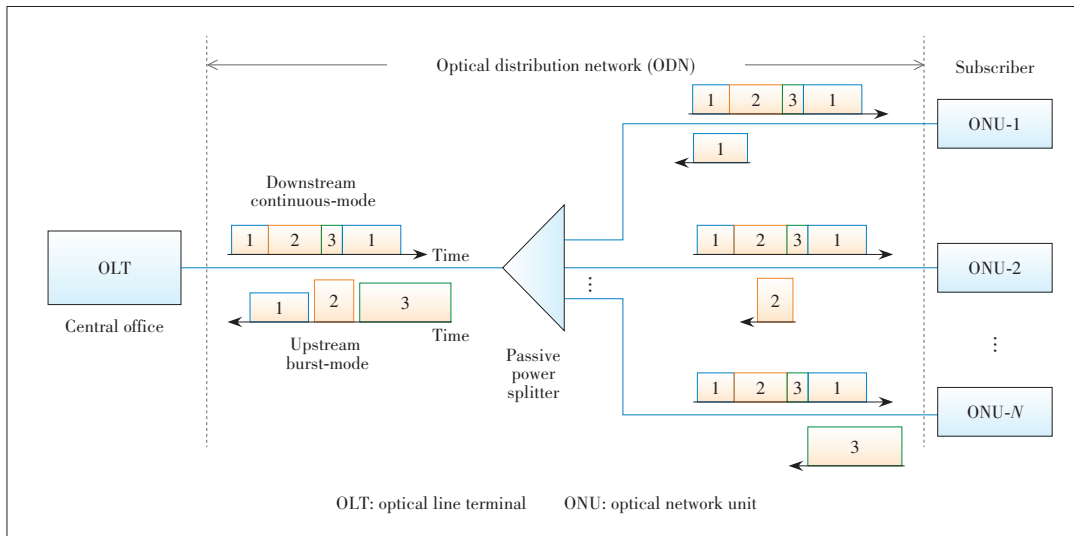
3.3 TWDM PON

TWDM PON is a combination of TDM and WDM PONs^[6]. Its latency limitation and methods for improvement thus follow the description in Section 3.1.

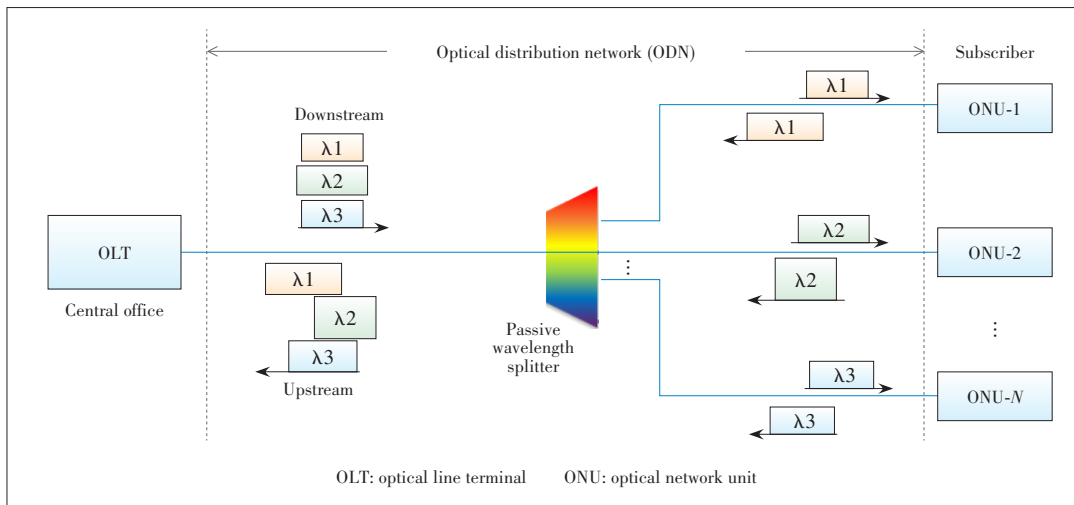
4 PON Standards Supporting Low Latency Services

As mentioned earlier, the ITU-T Q2/SG15 group is leading the efforts on standardizing optical access networks including PON for low latency services. The group began in June 2017 the “5G Wireless Fronthaul Requirements in a PON Context” project to analyze specifications from 5G standards, PON system requirements, and practically realizable PON architectures. Results of the study were agreed in October 2018 and published in the supplementary document G. Sup66^[7]. In addition, the Q2 group has completed the standard for single fiber bidirectional point-to-point optical access system covering line rates of 10 Gbit/s, 25 Gbit/s, and 50 Gbit/s in the G.9806 Recommendation to support 5G x-haul and business services^[8–9].

In this section, we provide a high-level summary of the findings in Supplement G.Sup66, with a focus on the latency aspect^[7, 10]. We will first describe the requirements from the wireless transport network perspective. We then discuss modifica-



▲ Figure 1. Schematic of a typical time-division-multiplexed (TDM) passive optical network (PON)



▲ Figure 2. Schematic of a typical wavelength-division multiplexing (WDM) passive optical network (PON)

tions to existing PON standards for practical PON implementations to meet the transport latency requirements.

4.1 Wireless Transport Network Latency Requirement

In a centralized radio access network (C-RAN) functional split architecture, the industry has converged on two interfaces: the F1 interface (midhaul/backhaul) for the high layer split Option 2, and the Fx interface (fronthaul) for the low layer functional split Option 6 or 7 (Fig. 3). In both cases, the transport capacity varies with the actual aggregated user traffic on the air interface. This is an essential feature which allows for applying more bandwidth and cost efficient x-haul networks.

The transport at the F1 interface is very similar to backhaul transport. The required end-to-end latency is in the order of tens of milliseconds for eMBB and in the 1 ms range for URLLC services, which leaves a sub-millisecond range for the transport layer.

At the Option 6 (media access control (MAC)-PHY split), Option 7 (Intra PHY split) and Option 8 (PHY-RF split), the acceptable transport latency is in the range of a few 100 μ s, similar to LTE, for eMBB and non-realtime mMTC services. Such low latency tolerance plays a critical role in the fronthaul network design. The Common Public Radio Interface (CPRI) Cooperation has defined transport network classes for categorizing the tolerances accordingly, ranging from 25 μ s to 500 μ s one way^[11].

4.2 Modifications to PON Standards

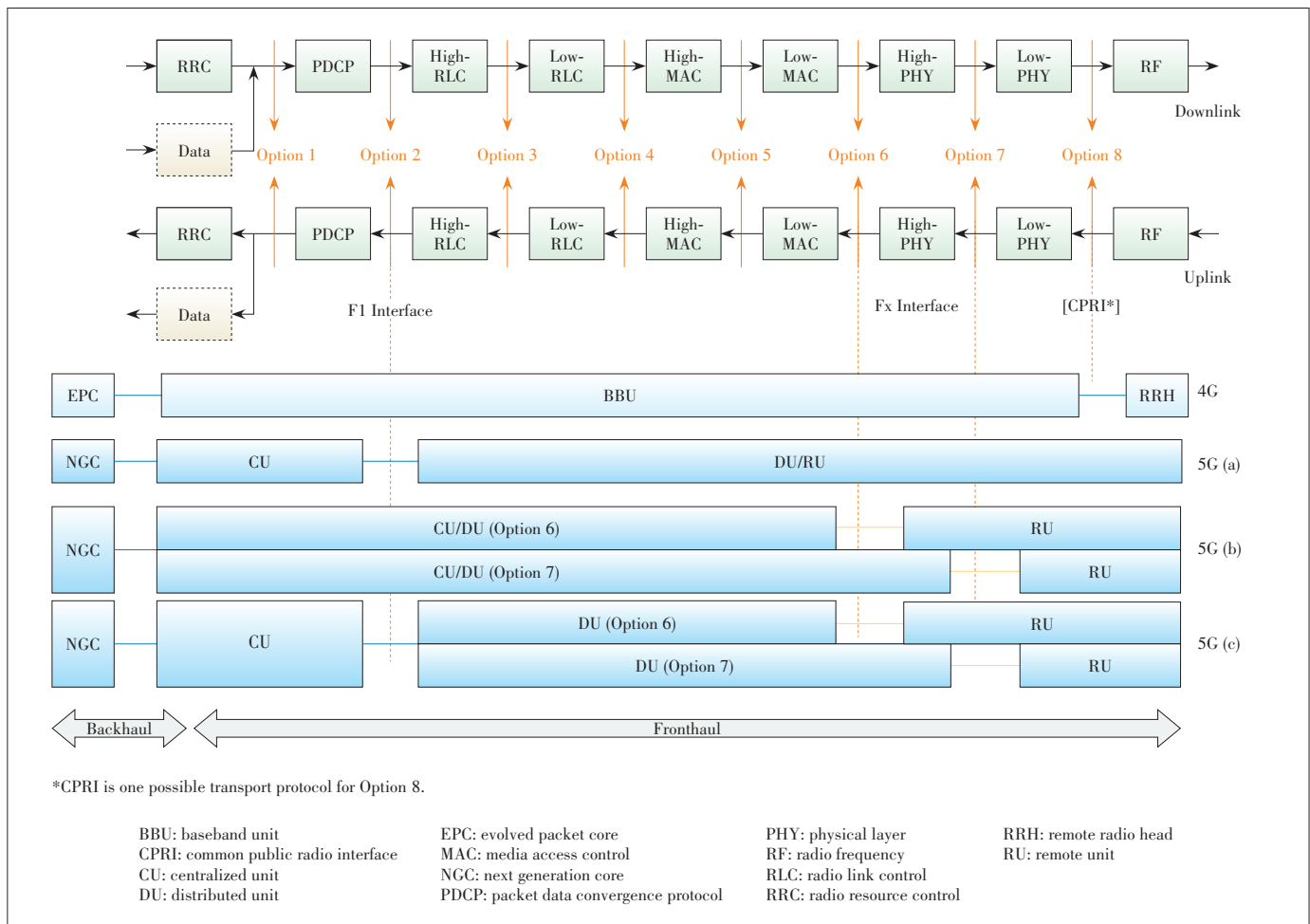
For a practical PON system to meet the transport latency requirements described above, methods to reduce the processing delays must be implemented. Here we consider two typical PON systems: TDM-PON and WDM-PON.

For the TDM PON,

as discussed in Section 3.1, there are two latency inducing factors, namely the DBA process and the quiet window during ONU activation.

To mitigate the DBA-induced latency, a straightforward method is to differentiate service classes, where the mobile traffic is assigned the highest priority with fixed bandwidth allocation. However, this leads to low bandwidth efficiency as any unused portion of the bandwidth cannot be reallocated.

Another method is the cooperative (CO) DBA, in which information exchange is introduced between the mobile scheduler (CU/DU) and the PON scheduler (DBA) in the OLT. This method allows the OLT to determine upstream bandwidth allocations in advance and then allocate the bandwidth at the expected arrival time of the upstream mobile traffic based on the actual traffic volume. This method is currently being studied in the ITU-T G.Sup.CODBA Supplementary project^[12] in collaboration with the O-RAN group. In addition, traffic descrip-



▲ Figure 3. Mapping of CU/DU/RU functions according to the split points: 5G(a) is high layer split (F1); 5G(b) is low layer split (Fx); 5G(c) is cascaded split (Reprint of Fig. 6-5 in G.Supp66^[7])

tors in traditional DBA need to be extended to support low latency. These extensions, including jitter tolerance, bandwidth assignment delay tolerance and protection switching delay tolerance, are being added to the existing PON standard^[5].

As for the latency due to quiet window opening, one proposal is to use a dedicated wavelength for ONU activation and registration. This dedicated activation wavelength (DAW) may be a newly defined wavelength, a separate PON system operating on a different wavelength on the same ODN, or a subset of wavelength channels in a TWDM PON. Once an ONU is activated in the DAW channel, it is handed over to the low latency operating wavelength channel to begin data transmission.

Another proposal is to use WDM PON, which does not require DBA nor ONU ranging. The latencies depend purely on the processing delays in the end nodes of the PON system and typically range below 10 μ s. As such, the ITU-T Q2/SG15 group began a project in February 2020 to standardize WDM PON. The initial target requirements are 20-pair of C-band wavelength channels each at 25 Gbit/s for up to 20 km distance.

5 Quiet Window Elimination Using Dedicated Activation Wavelength

In this section, we describe in more detail the process of using DAW to eliminate the quiet window for TDM PON activation, which is being studied in the ITU-T G.hsp.ComTC project^[13]. Experiment results are also shown as a proof of concept.

A DAW could be a newly defined wavelength not used in current PON systems or a legacy PON wavelength. In the latter case, the activation process needs to coordinate the quiet windows and exchange ranging information between the new PON and the legacy PON. Note that the DAW is only used for the activation purpose, not as a service wavelength as in TWDM or WDM PON system. The management of this wavelength channel is much less stringent and different from a service channel.

Here we give the example of the activation process for the scenario using a newly defined wavelength as DAW for a 50G PON, as shown in Fig. 4. In this scenario, three wavelengths are being used:

- λ_{50Gd} : Downstream (DS) wavelength of 50G PON;

- λ_{50Gu} : Upstream (US) wavelength of 50G PON;
- λ_{DA} : DAW in the US.

The activation process for this scenario is as follows.

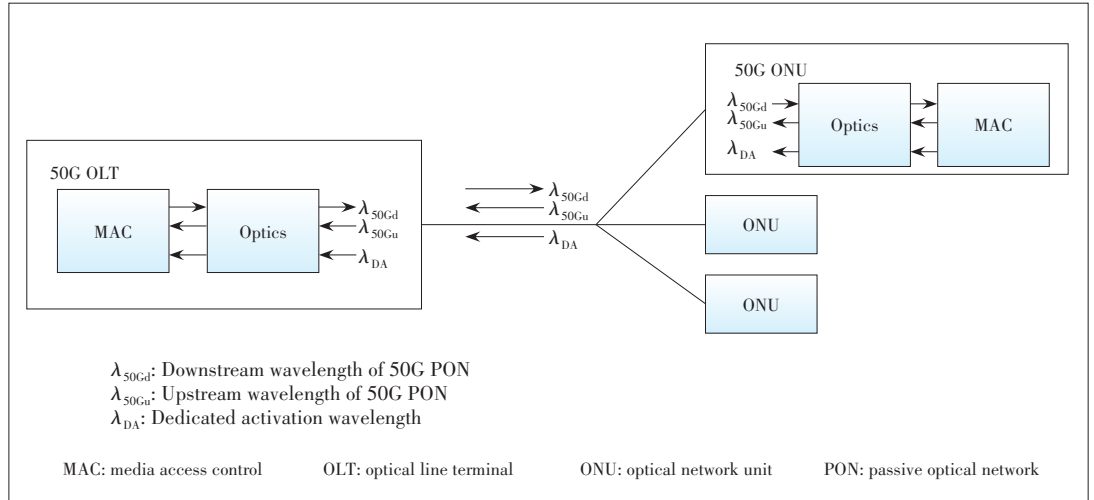
- 1) Upon power on, the ONU works at λ_{50Gd} and λ_{DA} , and listens to the serial number (SN) request at λ_{50Gd} .
- 2) The OLT opens a quiet window at λ_{DA} and broadcasts the SN request at λ_{50Gd} .
- 3) The ONU responds with its SN at λ_{DA} .
- 4) Once the OLT receives the SN response, it opens the quiet window at λ_{DA} and sends a ranging request at λ_{50Gd} directly to the ONU.
- 5) The ONU responds with the ranging response at λ_{DA} .
- 6) After receiving the ranging response, the OLT calculates the ranging results at $\lambda_{50Gd}/\lambda_{DA}$ based on the timing difference between the request and the response. The OLT further calculates the ranging results at $\lambda_{50Gd}/\lambda_{50Gu}$ using the ranging results at $\lambda_{50Gd}/\lambda_{DA}$ based on the dispersion difference between λ_{50Gu} and λ_{DA} . The OLT then sends the ranging results at $\lambda_{50Gd}/\lambda_{50Gu}$ to the ONU.
- 7) The ONU applies the ranging result at $\lambda_{50Gd}/\lambda_{50Gu}$ and starts working on λ_{50Gu} . The ONU tunes from λ_{DA} to λ_{50Gu} in case of tunable ONUs, or switches from λ_{DA} to λ_{50Gu} in case of dual-wavelengths ONUs.
- 8) The OLT assigns US bandwidth with burst profile of long preamble to the newly activated ONU at λ_{50Gd} .
- 9) The ONU sends Acknowledge PLOAMu message to the OLT.
- 10) The OLT assigns a directed US bandwidth with burst profile of short preamble to the ONU at λ_{50Gd} .
- 11) The ONU enters the operational state.

Another parameter to consider is the number of burst allocations per ONU within a PHY frame (125 μ s) in a BWmap. The current ITU-T standard specifies a maximum 16 burst allocations per ONU per 125 μ s,

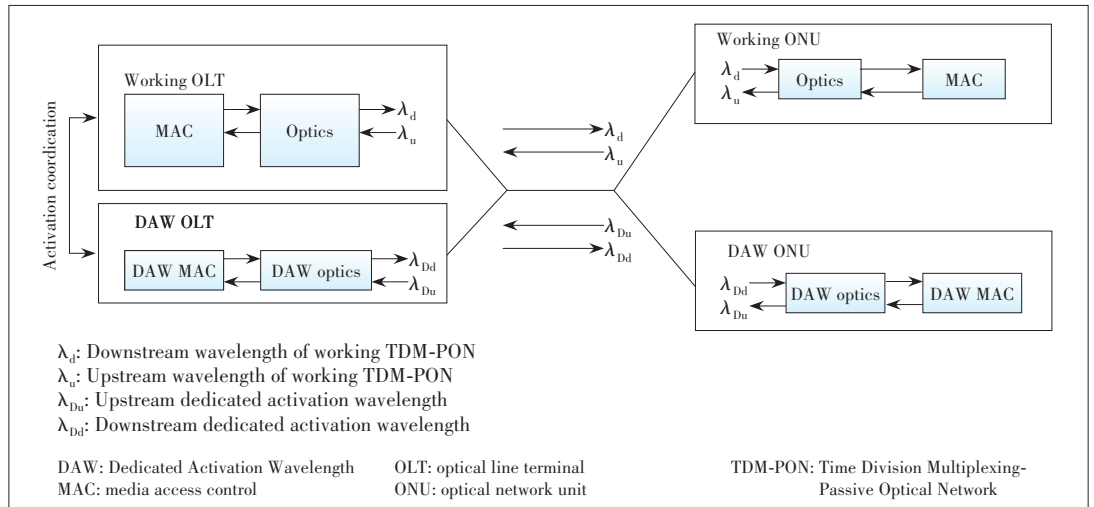
which corresponds to an ONU buffering delay of 7.8125 μ s. When more burst allocations in 125 μ s are allowed, lower buffering delay can be achieved, e.g., 4 μ s if 31 bursts are allocated^[14]. This latency reduction comes at the cost of bandwidth efficiency due to guard time and preamble per burst.

An experiment was set up to test low latency in a TDM-PON using DAW. The experimental configuration is shown in Fig. 5. Two pairs of wavelength channels are used in this system. One pair is working wavelengths and the other is DAW. The latency of quiet window is eliminated from the working wavelengths.

Furthermore, in this experiment, the fixed bandwidth for the ONU is split into multiple (N) mini-slots. The gap between mini-slots as well as the latency due to DBA is reduced when N increases. In this experiment, the traffic flow per direction is 950 Mbit/s, the total US bandwidth is 1 000 Mbit/s, and the fiber distance between the ONU and OLT is less than 100 m. Two cases of packet sizes are measured: a fixed length of 128 bytes and random length between 64 bytes and 1 518 bytes.



▲ Figure 4. Dedicated activation wavelength for 50G PON



▲ Figure 5. Experimental set-up

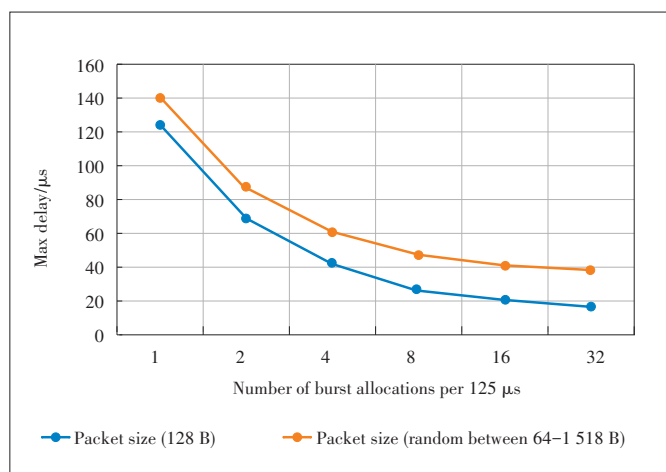
The results of maximum delay versus the number of burst allocations are shown in Fig. 6. The results show that the maximum latency reduces when the number of burst allocations per 125 μs increases in both cases of packet sizes. The reduction of maximum latency becomes more gradual as the number of burst allocations increase. Note that the maximum delay in current TDM-PON system is up to system configurations, e.g., the DBA duration is $M \times 125 \mu\text{s}$ ($M \geq 1$, typically 4) and the quiet window size is up to the differential distance (about 250 μs when the differential distance is 20 km). The typical maximum delay in TDM-PON system is higher than 750 μs . Obviously, it is much higher than that in this experiment and not shown in Fig. 6.

6 Conclusions

In summary, we provided an overview of low latency services and their corresponding requirements. As a potential solution to these requirements, PON technologies to support 5G xhaul transport are presented. Recent progress in PON standardization projects by the ITU-T Q2/SG15 is discussed. A proof-concept experiment and its results are described.

Acknowledgment

We would like to acknowledge Dr. Jun Shan WEY for the discussion and extensive review of the manuscript, and colleagues in the fixed media research and product teams of ZTE Corporation who contribute to the low latency PON project.



▲ Figure 6. Maximum latency under different numbers of burst allocations per 125 μs

References

- [1] ITU-R. IMT vision – framework and overall objectives of the future development of IMT for 2020 and beyond: ITU-R M.2083-0 [S]. 2015
- [2] WEY J S. The outlook for PON standardization: a tutorial [J]. Journal of light-wave technology, 2020, 38 (1): 31 – 42, 2020
- [3] The Full Service Access Networks Group. About FSAN [EB/OL]. [2021-03-01]. <http://www.fsan.org>
- [4] GlobalData and ZTE Corporation. White paper: precision 5G transport—the foundation of future mobile network [EB/OL]. (2021-02-26) [2021-03-01]. https://www.zte.com.cn/global/solutions/201905201708/201905201738/GlobalData_Precision_5G_Transport
- [5] ITU-T. 40-gigabit-capable passive optical network (NG PON2): transmission convergence layer specification: ITU-T G.989.3 recommendation amendment 3 [S]. 2020
- [6] ITU-T. 40-gigabit-capable passive optical network (NG PON2): G.989.x series of recommendations [R]. 2015
- [7] ITU-T. 5G wireless fronthaul requirements in a PON context, edition 3.0: ITU-T G.Sup66 supplement [S]. 2020
- [8] ITU-T. Higher speed bidirectional single-fiber point to point optical access systems: ITU-T G.9806 recommendation amendment 1 [S]. 2020
- [9] IEEE. IEEE P802.3cp bidirectional 10 Gbit/s, 25 Gbit/s, and 50 Gbit/s optical access phys task force [EB/OL]. [2021-03-01]. <http://www.ieee802.org/3/cp>
- [10] WEY J S, LUO Y, PFEIFFER T. 5G wireless transport in a PON context: an overview [J]. IEEE communications standards magazine, 2020, 4(1): 50 – 56. DOI: 10.1109/MCOMSTD.001.1900043
- [11] eCPRI Transport Network Group. Common public radio interface: requirements for the eCPRI transport network, v1.2 [EB/OL]. (2018-06-25) [2021-03-01]. http://www.cpri.info/downloads/Requirements_for_the_eCPRI_Transport_Network_V1_2_2018_06_25.pdf
- [12] ITU-T SG 15 (G.Sup.CODBA). ITU-T rec. series G supplement CO DBA: OLT capabilities for supporting CO DBA [Z]. 2021
- [13] ITU-T SG 15 (G.hsp.ComTC). Higher speed passive optical networks: common transmission convergence layer specification living list [Z]. 2021
- [14] OU H, TAKAHASHI K, NAKAMURA H. A proposal to change a maximum number of burst allocation series for improving upstream latency in G.989.3, contribution D3 [C]//ITU-T e-meeting. Geneva, Switzerland: ITU, 2016

Biographies

ZHANG Weiliang (zhang.weiliang@zte.com.cn) received his Ph.D. degree in communication and information system from Tsinghua University, China in 2001. He is currently a senior expert of fixed networks at ZTE Corporation and engaged in the research, standardization and product planning of fiber access and home networking. He has led or participated in a number of national “863” projects, as well as provincial and ministerial key projects. He has published more than ten papers and held more than 100 authorized patents.

YUAN Liquan received his master’s degree in information engineering from Harbin Institute of Technology, China in 1999. As a project leader at ZTE Corporation, he is responsible for pushing forward the research and standardization in fiber access and home networking of the company, cooperating with the standardization bodies including ITU-T SG15/IEEE802.11/IEEE802.3/Broadband Forum/CCSA/ETSI. As an editor, he has participated in drafting more than 10 optical access standards and held more than 40 authorized patents.



Differentially Authorized Deduplication System Based on Blockchain

Abstract: In architecture of cloud storage, the deduplication technology encrypted with the convergent key is one of the important data compression technologies, which effectively improves the utilization of space and bandwidth. To further refine the usage scenarios for various user permissions and enhance user's data security, we propose a blockchain-based differential authorized deduplication system. The proposed system optimizes the traditional Proof of Vote (PoV) consensus algorithm and simplifies the existing differential authorization process to realize credible management and dynamic update of authority. Based on the decentralized property of blockchain, we overcome the centralized single point fault problem of traditional differentially authorized deduplication system. Besides, the operations of legitimate users are recorded in blocks to ensure the traceability of behaviors.

Keywords: convergent key; deduplication; blockchain; differential authorization

ZHAO Tian¹, LI Hui¹, YANG Xin¹, WANG Han¹, ZENG Ming², GUO Haisheng², WANG Dezheng²

(1. Shenzhen Graduate School, Peking University, Shenzhen 518055, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202102009

<http://kns.cnki.net/kcms/detail/34.1294.TN.20210519.1626.002.html>, published online May 20, 2021

Manuscript received: 2021-01-13

Citation (IEEE Format): T. zhao, H. Li, C. Yang, et al., "Differentially authorized deduplication system based on blockchain," *ZTE Communications*, vol. 19, no. 2, pp. 67 – 76, Jun. 2021. doi: 10.12142/ZTECOM.202102009.

1 Introduction

In recent years, with the development of cloud storage technology, user data are uploaded to the cloud server and many copies of the data are repeatedly stored by different users, resulting in the waste of storage space. This makes the deduplication an urgent problem to be solved^[1].

However, if the file data is directly stored in the cloud storage server, it faces a series of risks such as data theft. Therefore, we consider storing the ciphertext of the data in the cloud storage server.

In traditional encryption and decryption algorithms, the keys are generated independently by users leading to various

ciphertexts of the same data, which makes the deletion of duplicate data difficult. If the cloud storage server generates the key and encrypts the data uniformly, the security of user data cannot be guaranteed once the cloud storage server is maliciously attacked and becomes untrustworthy.

In order to achieve deduplication under the premise of data security, a deduplication system based on convergent keys has been proposed in Ref. [2]. The encryption key $H(F)$ is obtained by hashing the data, and used to encrypt the user data. The convergent encryption makes the consistent ciphertext of the same file or data block, and the cloud storage server or external attackers cannot see the original data. It not only guarantees the confidentiality of the data, but also facilitates the cloud storage server to perform data deduplication by using the original data to generate a convergence key.

Because the encryption method of the convergent key is vulnerable to offline brute force cracking, semantic security cannot be guaranteed^[3, 4]. In recent years, researchers in deduplication for convergent encryption have proposed a series of improvements. BELLARE et al.^[5] proposed an information lock encryption scheme, which optimized key calculations and en-

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. 2019ZTE03-01, National Keystone R&D Program of China under Grant No. 2017YFB0803204, National Natural Science Foundation of China (NSFC) under Grant No. 61671001, Guangdong Provincial R&D Key Program under Grant No. 2019B010137001, Shenzhen Research Programs under Grant Nos. JCYJ20190808155607340, JSGG20170406144032901, JSGG20170824095858416 and JCYJ20170306092030521, and PCL Future Regional Network Facilities for Large-scale Experiments and Applications under Grant No. PCL2018KP001.

encryption methods. PUZIO et al.^[6] designed the first repetition based on double-layer encryption in the deduplication scheme. The inner layer uses the convergent encryption schemes mentioned above, and the outer layer is outsourced to a trusted third party. In addition, BELLARE et al.^[7] also described the DupLESS scheme, which adds an additional key to the convergent key generation process to invalidate the dictionary attack. LI et al.^[8] proposed to use a deterministic secret sharing scheme instead of convergent encryption.

The above schemes are designed to alleviate the data security problem, but they do not fully consider how to build a credible authority when there are authority differences between users. Since the blockchain has the advantages of convenient generation and non-tampering, we consider introducing the blockchain technology to solve this problem.

NAKAMOTO Satoshi published the first paper on blockchain in 2008^[9]. Although its main introduction focuses on Bitcoin, a digital currency payment system, blockchain has also caused extensive research in academia as its carrier. It allows any party that has reached an agreement to directly generate transactions without the participation of third-party intermediaries^[10]. The blockchain encapsulates the history of consensus transactions in the block, as well as the identities of participants and timestamps. Each block uses the Hash algorithm to generate an important identification header for sequential connection, forming a chained data structure, which can be used as a distributed transaction and log record of the entire system^[11].

Blockchain has the characteristics of decentralization, non-tampering, and traceability. Decentralization avoids the damage to the system caused by the evil master node in the traditional storage model; non-tampering ensures that if the attacker wants to tamper with a certain data in a block, he needs to recalculate the block and all the subsequent blocks; traceability guarantees that each user's operation can be located and tracked, which virtually increases its destruction cost.

The structure of the blockchain can be roughly divided into 6 levels, namely the data layer, network layer, consensus layer, incentive layer, contract layer and application layer. The data layer is at the bottom, which is mainly used to implement functions of data storage and transaction recording. The network layer is used to realize the functions of data transmission and verification. The consensus layer is the core part of the blockchain. It encapsulates various consensus algorithms. It is mainly used to achieve the consistency of block generation and transaction data. The incentive layer is mainly responsible for introducing incentive mechanisms, such as token distribution to miners who generate new blocks to encourage mining. The contract layer mainly includes various scripts written to enable the blockchain to obtain programmable application attributes. The application layer is the display of the blockchain in specific application scenarios, with many different manifestations. In this paper, we mainly use the first three lay-

ers of the blockchain. More specifically, the components of the consensus layer are utilized and improved.

According to different application scenarios, blockchains can be divided into the public blockchain, private blockchain and consortium blockchain. The public blockchain is completely open. Any user in the entire network is allowed to freely join or exit the blockchain system and everyone has equal rights. In the private blockchain, an organization has complete ownership of data on distributed nodes. The consortium blockchain is somewhere between the public and private blockchains, where several groups participate in the management of each node to grant different identities to jointly maintain the blockchain system.

We propose a differential authorization deduplication system based on blockchain, which alleviates the problems of single point of failure and inflexible permission changes. The user's public key and the permissions signed by the private key are written into the blockchain. It ensures the security of user permissions through the immutable modification of the blockchain and maintain each user's permission table. When the permission is changed, the blockchain directly generates a new block to cover the original permission, which is convenient for the dynamic modification of the permission. On the other hand, the blockchain can record each user's operation on the permission to ensure its traceability.

The rest of this paper is organized as follows. In Section 2, we describe the traditional differential authorization deduplication system and its problems. We also introduce the messaging process of the PoV algorithm. Then our differential authorization deduplication system based on blockchain is proposed in Section 3, followed by the performance analysis in Section 4 and experimental simulations in Section 5. Finally, conclusion and future works are drawn in Section 6.

2 Preliminaries

2.1 Traditional Structure

In this section, we introduce the traditional differential authorization deduplication system, including the main process of file upload and download, as well as its problems.

Let us consider a practical application scenario. In a company, subordinate relationships exist in different users leading to various permissions. For this reason, we have higher requirements for deduplication. Differential authorization deduplication should be implemented. Users with higher authority can upload and download data, while users with lower authority cannot download and access the data of high-level users.

The traditional differential authorization system^[12] mainly provides different permission sets for different users. It introduces a private cloud server to maintain the permission table, and adopts the hybrid cloud architecture to realize the deduplication of differential authorization.

The system is mainly composed of three parts: the public storage cloud server provider (S-CSP) responsible for storing encrypted user data, the private cloud server responsible for maintaining the user permission table, and the user who uploads and downloads files.

The specific workflow is as follows:

1) System Initialization Stage

Define the tag of file F as $\mathcal{O}_F = \text{TagGen}(F)$, and a label corresponds to a unique file data. Each permission p of the system has a corresponding permission key k_p . Define the token of file F as $\mathcal{O}'_F = \text{TagGen}(F, k_p)$, that is, only users with permission p can access file F .

Assuming that the permission set owned by user U is P_U , its corresponding permission key $\{K_p\}, P_i \in P_U$ will be sent to the private cloud server acting as a permission check server. The private cloud server maintains a table to store each user's public key pk_U and its corresponding permissions.

2) File Upload

The file upload process is shown in Fig. 1.

Suppose that the file owner U wants to upload a file F for access by users with permission $\{P_j\}, P_j \in P_F$. First, the user needs to use his private key sk_U to verify his identity with the private cloud server. If the verification is passed, the user needs to send the tag $\mathcal{O}_F = \text{TagGen}(F)$ of the file F to it. The private cloud server will return all initial file tokens $\{\mathcal{O}'_{F,p} = \text{TagGen}(F, k_p)\}, p \in P_U$ that match the user's permissions, then the user will send these tokens to the S-CSP.

If duplicate files are found in S-CSP during upload, S-CSP first runs the Proof of Ownership (PoW) algorithm^[8] to verify the user's ownership of the file. If it passes, it will return a file pointer to the user. A signature δ and a time stamp are appended to the token $\{\mathcal{O}'_{F,p}\}$ and returned to the user. The user

sends the token and the permission set $\{P_j\}$ of the file F to the private cloud server for verification. After the verification is passed, the private cloud server will calculate all the file tokens $\{\mathcal{O}'_{F,p} = \text{TagGen}(F, k_p)\}, p \in P_F - P_U$ and return to S-CSP. The permissions of file F at this time are the union of the permissions of P_F and other owners of the file.

If S-CSP does not find duplicate files when uploading, a signature δ and a time stamp are appended to the token $\{\mathcal{O}'_{F,p}\}$ and returned to the user. The user sends the token to the private cloud server for verification. After passing the identity verification, the private cloud server will calculate all the file tags $\{\mathcal{O}'_{F,p}\}$ within the P_F authority and return to the S-CSP, then the user can upload the data encrypted by the convergent key to the S-CSP.

3) File Download

The user sends a file download request to the S-CSP, and the S-CSP will verify the user's permissions. If it cannot download, S-CSP will return the download failure. If it can download, S-CSP will return the encrypted data. The user uses the locally saved convergent key to decrypt the file and get the original data.

However, the data deduplication solution has some problems:

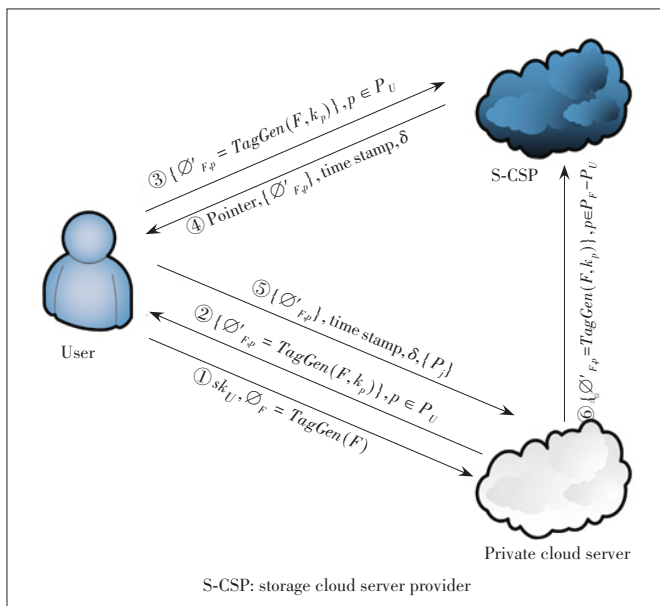
The first is the security assurance issue of the private cloud server. If the private cloud server is attacked and the user and corresponding permissions are tampered with, the system will not operate normally.

The second is the dynamic change of permissions. Once a file is uploaded, its permission is difficult to modify flexibly. When the user and file permissions change, for example, a user no longer has file permissions, the system cannot modify permissions in time.

2.2 Blockchain Consensus Algorithm

The design and implementation of blockchain involves many algorithms, the core of which is its consensus algorithm. For example, the consensus algorithm used by Bitcoin is the Proof of Work (PoW)^[13]. The main idea of the PoW algorithm is that each independent node in the network conducts competitive mining to solve mathematical calculation problems, thereby obtaining the following accounting rights and generating new blocks. However, with the continuous mining of bitcoins, the mathematical puzzles that need to be solved to generate new bitcoins have become more and more complicated, which has caused a huge waste of computing resources and lower efficiency^[14]. In addition, the important blockchain consensus algorithms include the Proof of Stake (PoS)^[15, 16], Delegated Proof of Stake (DPoS)^[17, 18], etc.

At present, there are many studies on the consensus algorithm of the blockchain. In this system, we use the blockchain based on the Proof of Vote (PoV) consensus^[19] to construct the blockchain in the system. The PoV consensus can well ease double-spending attacks, selfish mining, witch attacks and



▲ Figure 1. File upload

other attack methods. It also can well guarantee the security of user permissions. There are four types of nodes in this system, including commissioners responsible for voting, butlers responsible for accounting and production blocks, butler candidates, and ordinary user nodes that can apply to become butler candidates. This system allows concurrent roles to a certain extent, as shown in Fig. 2.

The consensus process of block generation is carried out jointly by a butler and all committee members. The butler is called duty butler, and the duty butler is determined by the butler number selected by the committee members. Assuming there are n commissioners, namely C_1, C_2, \dots, C_n ; and m butlers, namely B_1, B_2, \dots, B_m , the consensus process of the block is shown in Fig. 3.

PoV divides a round of consensus into four phases: Prepare, Ready, Commit, Confirm. Among them, the Confirm stage is for the block placing, with no need to send messages. Each block contains a block header and an indefinite number of transactions. In the Prepare phase, the butler on duty takes a certain number of transactions from the transaction pool, packs them into pre-blocks and sends the pre-blocks to all committee nodes. The difference between the pre-block and the official block is that the pre-block does not have a timestamp, committee signature, and the number of the next butler on duty. The committee node needs to verify the block header

of the received pre-block and the information contained in the transaction. If the verification passes, it will sign the pre-block header and send the signature to the duty butler.

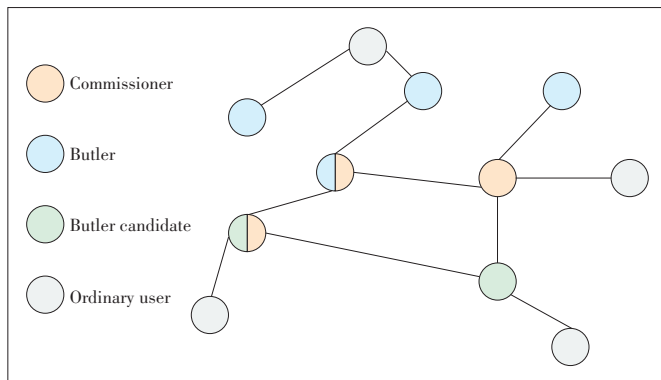
The duty butler can complete the pre-block and release the official block after collecting signatures of more than $n/2$ committee members. The node that receives the newly released block stores the block in the local blockchain and updates the relevant variables including the number of the duty butler, thereby replacing the duty butler who is responsible for the next round of consensus. The information supplement of the pre-block header depends on the signature of the committee member. The PoV stipulates that the latest member signature time of the signature is issued as the generation time of the block, and the next housekeeper number is generated by hashing the signature. These regulations make use of the randomness and unforgeability of signatures. Since the signatures generated by each committee node are random, the next butler number calculated is also random.

3 Differentially Authorized Deduplication System Based on Blockchain

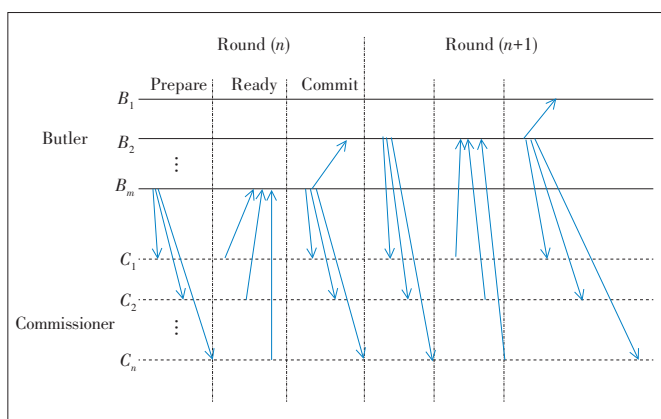
In order to solve the above problems, this paper designs and implements a differentially authorized deduplication system based on blockchain. In more complex specific application scenarios, such as several companies working together to develop a project, they need to implement data deduplication on the same cloud storage server, which requires the system to credibly record the behavior of users to facilitate accountability. On the other hand, it is necessary to implement differential authorization of files according to different user identities and also to be able to make changes in time when users or file permissions change. Using the immutability and traceability of the blockchain to ensure the security of user permissions and accountability of behavior, these requirements can be met well. At the same time, when the user's file management authority changes, we can write the authority change into a new block. Because the blockchain is based on the record in the newly generated block, we can achieve dynamic changes in permissions.

The system is divided into three parts: the public cloud server S-CSP that stores encrypted data, the users who upload and download files, and the blockchain that saves permissions and upload and download records.

Blockchain is a distributed ledger and used in the Bitcoin currency transaction system. The blockchain network system maintains an orderly data block that keeps growing without a center. Each data block has a timestamp and a pointer to the previous block. Once the data is on the chain, it cannot be changed. Blockchain can be analogous to a distributed database technology. By maintaining a chain structure of data blocks, it can maintain a continuously growing, non-tamperable data record. The blockchain of our system is constructed



▲ Figure 2. PoV network model



▲ Figure 3. Message transmission process of Proof of Vote (PoV) consensus

using the PoV algorithm described in the previous section, and its immutability is mainly guaranteed by the consensus mechanism.

As mentioned in Section 1, data deduplication achieved by convergent encryption can be performed in file-level data and block-level data respectively. In order to further save storage space and efficiently use bandwidth, we can encode the file into n data blocks $\{Bi\}$. When the files are not the same but the content is not much different, the data block deduplication check is used to complete the deduplication.

We will separately discuss the upload and download of file-level data and block-level data.

3.1 File-Level Data Upload and Download

1) System Initialization

Define the file F label as $\varnothing_F = \text{TagGen}(F)$, and a label corresponds to a unique file data. The user's permission set is $\partial = \{\partial_1 \dots \partial_n\}$, where we define its number from small to large as the permission from high to low. Those with high-level permissions can access files uploaded by people with low-level permissions and modify file permissions. In the initial state of the system, the blockchain will have an authority table signed with private key S_{hp} of the negotiated highest authority owner P to declare the authority level of each user. Any legal user in the system can use its public key p_k to check the authority. At the same time, the blockchain will also store the label \varnothing_F of the file F and the encrypted file permission ∂_F , which is convenient for S-CSP query. Suppose user U wants to upload a file with permission ∂_F , the user's private key is S_{ku} and S-CSP is initially blank.

2) File-Level Data Upload

The file-level data upload process is shown in Fig. 4.

The user sends the tag $\varnothing_F = \text{TagGen}(F)$ of the file to upload encryption by the private key S_{ku} , the user name U and its own authority ∂_U , and the S-CSP will query whether there is the tag $\varnothing_F = \text{TagGen}(F)$ of the file and the permission ∂_F of the file on the blockchain. If the file exists and the permission is lower than or equal to user permission ∂_U , the user needs to verify that he owns the file with S-CSP using the POW algorithm. At this time, the server will return a pointer to the user indicating that the server already has the file and the user has no need to upload repeatedly.

If the file does not exist or has no right to access the file, the user needs to write the uploaded relevant information to the blockchain, including the file tag \varnothing_F , the user name U , the user name U signed by the private key S_{ku} and the file at the access level $\{\varnothing_F, S_{ku}(U, \partial_F)\}$. After verifying the identity of the user and S-CSP, the blockchain writes the record into a new block, and then the user can send the data of the file encrypted by the convergent key to the S-CSP.

3) File-Level Data Download

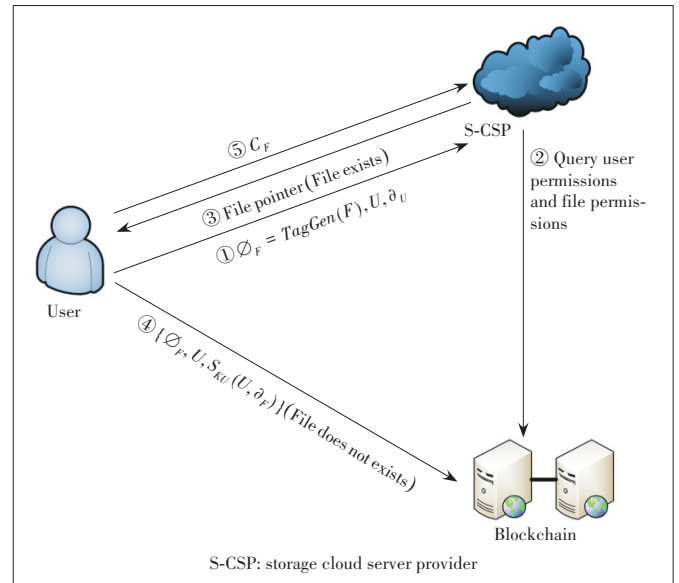
The file-level data download process is shown in Fig. 5.

The user sends the tag $\varnothing_F = \text{TagGen}(F)$ of the file that will

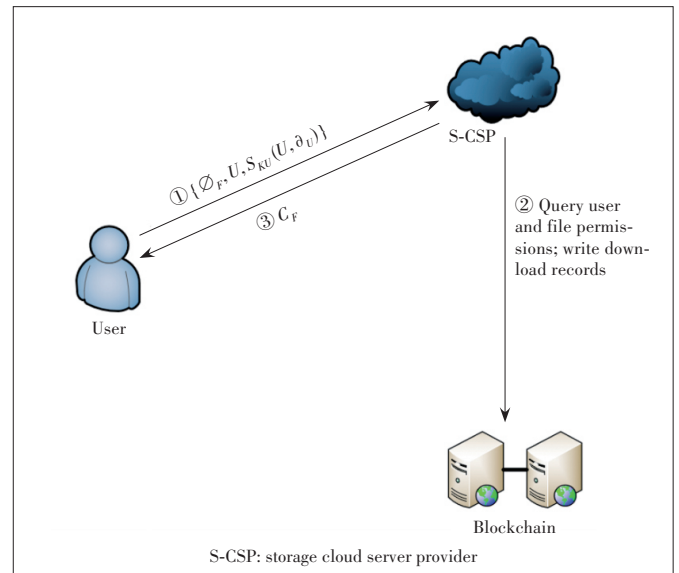
be downloaded to the S-CSP. The S-CSP will query whether the file exists. If the file does not exist, the S-CSP will return a prompt message to the user.

If the file exists, the user U sends the file tag \varnothing_F , the user name U , and the user name and authority information encrypted by its own private key $S_{ku} : \{\varnothing_F, U, S_{ku}(U, \partial_U)\}$ to S-CSP; S-CSP uses the user's public key P_{ku} to decrypt for obtaining authority, and then uses the public key of the highest authority to query the authority table to confirm whether the authority matches.

After passing, the S-CSP will send a confirmation message to the blockchain. After the blockchain verifies the identity of the user and S-CSP, the download record is saved in the block, and the S-CSP returns the file ciphertext C_F encrypted



▲ Figure 4. File-level data upload



▲ Figure 5. File-level data download

by the convergent key to the user. The user uses the locally stored convergent key to decrypt the file. LI et al.^[20] have also done related research on the storage of convergent keys, but this is not the focus of this article. For simplicity, this article uses the traditional method of saving locally.

3.2 Block-Level Data Upload and Download

The block-level data upload and download process is similar to the file-level data upload and download process, as follows:

1) System Initialization

It is roughly the same as the system initialization requirements in Section 3.1, except that the file is divided into $\{Bi\}$ data blocks for upload and download.

2) Block-Level Data Upload

The user first sends the file F and all the tags \varnothing_F and $\{\varnothing_{Bi}\}$ of the data block $\{Bi\}$, the user name U , and the own authority ∂_U to the S-CSP.

S-CSP will query whether there is a label for the file. If label \varnothing_F of the file exists, it will turn to the processing flow of the file label in the previous section. The user can prove that he owns the file through PoW, and then S-CSP returns the corresponding pointer to inform the user that the file already exists. If the data block exists, the user needs to use the PoW algorithm to verify to the S-CSP that he owns the data block. At this time, the server will return a pointer to the user indicating that the data block already exists in the server, and there is no need to upload it repeatedly. S-CSP will add a new record to the blockchain, that is, the label of the newly added file corresponding to the data block, which is convenient for the repeatability check of the next file and data block.

If the data block does not exist, the user needs to write the uploaded relevant information to the blockchain, namely the file tag \varnothing_F , the data block tag \varnothing_{Bi} , the user name U , the user name signed by its own private key and the file can be accessed permission level: $\{\varnothing_F, \varnothing_{Bi}, U, S_{KU}(U, \partial_F)\}$. After verifying the identity of the user and the S-CSP, the blockchain writes the information into the block, and then the user can send the data of the data block encrypted by the convergent key to the S-CSP.

3) Block-Level Data Download

The user sends the file label that will be downloaded to S-CSP. S-CSP will query whether there is a label for the file on the blockchain. If the file does not exist, S-CSP will return a prompt message to the user. If the file exists, the user sends the file label, data block label, user name and authority information encrypted by own private key: $\{\varnothing_F, \varnothing_{Bi}, S_{KU}(U, \partial_U)\}$ to S-CSP. S-CSP uses the user's private key to decrypt for obtaining the authority, and then uses the public key of the highest authority to query the authority table to confirm whether the authority matches. S-CSP will send a confirmation message to the blockchain; after passing the identity verification, the download record is saved in the block. The S-CSP returns

the data block ciphertext encrypted by the convergent key to the user, and the user uses the locally stored convergent key to decrypt the data block. Then the user can restore the original data file $F=\{Bi\}$.

The encryption method of the convergent key has its inherent flaw, that is, it cannot resist offline dictionary attacks. Specifically, if external attackers know the ciphertext C_F and can infer the file set $\{F\}$, they can directly generate the convergent key and encrypt the corresponding files for comparison. If the ciphertext is the same, attackers may obtain the original data file.

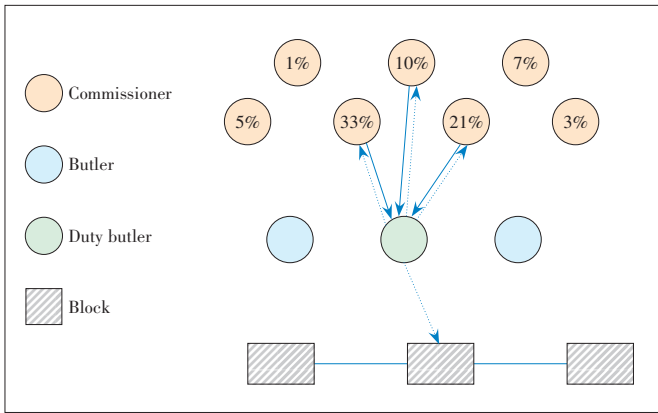
In response to this defect, this article proposes to use block-level data upload when storing high-privilege files, and use double-layer encryption for more important data blocks (in the application scenario of deduplication, the number of important data blocks in the file is less), that is, use another *Hash* function to generate the convergent key in the outer layer of the encrypted data block to encrypt again. After S-CSP receives the data, it sends a message to all users higher than the authority, calculates the convergent key of the *Hash* function and saves it. When the user needs to download a file, the convergent key is used to decrypt the outer layer, and then the original convergent key is used to decrypt the inner layer. This mode can be turned on or off according to the requirements of the system application scenario.

3.3 Specific Application of PoV Algorithm in The System

This section mainly introduces the specific consensus and generation process of the blockchain used to store user authority information. In this system, we introduce the PoV consensus algorithm in Section 2.2 to generate our blockchain. The consensus process is roughly the same as the foregoing. Specifically, we can implement the PoV algorithm in the initial state with the company's leadership as the committee node and ordinary employees as the ordinary nodes. The ordinary employees can submit a campaign request, and the members of the leadership will vote to select trusted butler nodes and the current duty butler nodes. The initial block maintains the permissions of all initial users. After the initial block is confirmed by all members, the selected butler node on duty is responsible for generating a new block to record the permissions of all users and the upload and download records of files. The specific block generation process is shown in Fig. 6. The percentage in the figure indicates the weight of commissioners.

Different from the traditional PoV scheme, we no longer require the replacement of the duty butler after a specific block is generated. Instead, the committee members vote to elect a new butler on duty. This is because in actual application scenarios, a housekeeper on duty is probably not online. At this time, if you need to upload or download files or change permissions, committee members need to select a new butler on duty to handle the request.

In addition, we have also changed the voting weight of the



▲ Figure 6. Block generation process

committee members, that is, the voting weight of each committee member is no longer equal, but the voting weight of each committee member is different according to their different rights. This improvement has brought the following benefits: First, it is in line with the logic of inter-company affairs and people with higher status can have more voice to determine file access and permission changes; second, when there is an emergency, it only needs to notify a few committee nodes to make the voting weight more than half to complete the fast processing of the transaction; finally, when the user scale is expanded to a large scale, the voting scheme can speed up the consensus completion time and improve the transaction processing effectiveness.

4 Theoretical Analysis

In this section, we analyze the performance of the system, including functional analysis and safety analysis.

4.1 Functional Analysis

4.1.1 Differential Access Control

The main process of uploading and downloading data of the system has been introduced. The specific differential authorization is embodied in the user uploading a file, which can be easily written into the file's permissions, including reducing or increasing file permissions. For example, a fourth-level permission can upload a fifth-level permission file for low-privilege access; it can also upload a high-level file such as the second-level authority. Users with a level greater than or equal to the second-level authority can access the file, and at the same time, the third-level user has no right to access the file, which realizes the system's differential authority access control.

4.1.2 Dynamic Changes in Permissions

The system can also solve the problem of dynamic changes in permissions.

When the user authority changes, we can update the record in the block, generating a new block record. According to the

traceability of the blockchain, when all nodes confirm the block, the user will have the new authority. At this time, if the authority level is increased, the user can access and download the high-level authority file. If the authority is reduced, the user cannot access the original authority file.

When the file authority changes, the high-level authority or the file uploader can send information to the blockchain again, rewrite the file authority level, and realize the dynamic management of file authority.

4.1.3 Single-Enterprise Environment

Although the system mainly considers the differential authorization deduplication between multiple institutions, it is also applicable to a single-enterprise environment. Using the typical structure of blockchain, the generation of new blocks is relatively simple. At the same time, the PoV consensus mechanism has strong consistency, which means that when the user or file permissions change, the change can be quickly confirmed, thus ensuring the efficiency of system operation.

As mentioned earlier, in the context of multi-institution, using this scheme can ensure the security of authority management. If you use this solution internally, you can deploy the blockchain on the cloud. Nodes located in different data centers act as butler nodes, so that when a butler node fails, the entire blockchain can still operate stably, which greatly improves the availability of the system.

4.2 Performance Analysis

Compared with the traditional scheme, the additional cost of this system mainly includes the following aspects, namely, the time for generating a new block by consensus, the time for permission query, and the time required for outer encryption.

For simplicity, we use some symbols to represent these variables, as shown in Table 1.

It can be seen that in our solution, the time required to complete the upload and download of the entire file is:

$$T_A = T_{CB} + T_{PQ} + T_{FL} + T_{CE} + T_{FT}$$

In the traditional solution, because the permission table is maintained on the private cloud server, the permission query speed is very low and can be ignored. At this time, the total time required by the traditional solution is:

$$T = T_{FT} + T_{FL} + T_{CE}$$

▼ Table 1. Parameters and their description

Parameter	Description
T_{FL}	Time required for file label generation
T_{CE}	Time it takes for the file or data block to convergent encryption
T_{FT}	File transfer time
T_{PQ}	Time required for blockchain permission query
T_{OE}	Time required for outer encryption
T_{CB}	Time required to generate a block
T_A	Total time
T_{AE}	Enhancement plan total time

We set $\omega_1 = (T_{GB} + T_{PQ})/T$. At this point, if ω_1 is small enough, it proves that our solution achieves dynamic management of permissions without introducing obvious additional overhead.

In the enhancement scheme, the time we need to complete the entire process is $T_{TE} = T_A + T_{OE}$.

We set $\omega_2 = T_{OE}/T_A$. It can be seen that if ω_2 is small enough, that is the number of blocks that need to be double-layered encryption is relatively small, or the time used for outer encryption is relatively small, it proves that we can ensure the security of the system without significantly increasing the time overhead.

4.3 Security Analysis

The security of permissions is mainly guaranteed by the security of the blockchain, that is, the tolerance of attacking nodes.

Define the number of blockchain nodes as n and the number of attackers as f , when the weight of each node is the same, our tolerance for attacking blocks is:

$$f \leq \frac{n-1}{2}.$$

That is, it can tolerate no more than half of the nodes being attacked, which better guarantees that the permissions cannot be tampered with.

At the same time, the double-layer encryption scheme for confidential data blocks mentioned in Section 3 can also better prevent offline dictionary attacks. The encryption method we use for convergent encryption is Advanced Encryption Standard (AES) encryption, and the key length is 256 bit. AES encryption has good resistance to brute force cracking. If 10 000 collision attacks are executed every nanosecond, it will take 1.8×10^{56} years to crack^[21].

5 Experimental Simulation

We implemented the model of the system and ran it in our own experimental environment. The code was implemented in C++. Table 2 shows the hardware conditions for the implementation environment.

We tested the performance of this model. The main measurement indicators are the ratio of the time used for permission query, file transfer, file label generation, and convergent encryption when the upload and download file sizes are different.

We used AES encryption as the encryption algorithm for

convergent encryption. The key is a 256-bit hash value generated by the SHA-256 algorithm; the file label is generated by the SHA-1 algorithm. When the permission changes, we can manually set the new block generation time to 1 s, which is the time required for the permission change.

We tested the performance of the system when the file size was 100 MB, 200 MB, 300 MB, and 400 MB (Fig. 7).

Because the authorization query time is too short, the figure shows the total time used for 10 000 queries. It can be seen from the experimental results that compared to data transmission and convergent encryption, the time required for user permission query and record writing to the blockchain is shorter. Our system does not significantly increase system overhead while achieving differential authorization deduplication.

At the same time, we also tested the enhancement scheme proposed in Section 3. The main performance indicators are the ratios of the time used for double-layer encryption of important data blocks, file label generation, convergent encryption and file transfer.

In this enhanced scheme, the outer layer encryption uses the AES encryption algorithm, and the key is a 128-bit hash value generated by the MD5 hash algorithm. The inner layer encryption uses the AES encryption algorithm, and the key is a 256-bit hash value generated by the SHA-256 algorithm.

We tested the situation where the number of important data blocks (requiring double-layer encryption) accounted for different proportions of the total number of data blocks. For convenience, we set the number of important data blocks to 1, the size of 100 MB, and the size of ordinary data blocks to 100 MB. The experimental results are shown in Fig. 8.

It can be seen from the experimental results that when the important data is relatively few, our enhancement scheme will not significantly increase system overhead while improving data security.

In summary, from the experimental results, the maximum value of ω_1 is less than 1%, and the maximum value of ω_2 is less than 5%, indicating that our solution does not significantly increase the overhead.

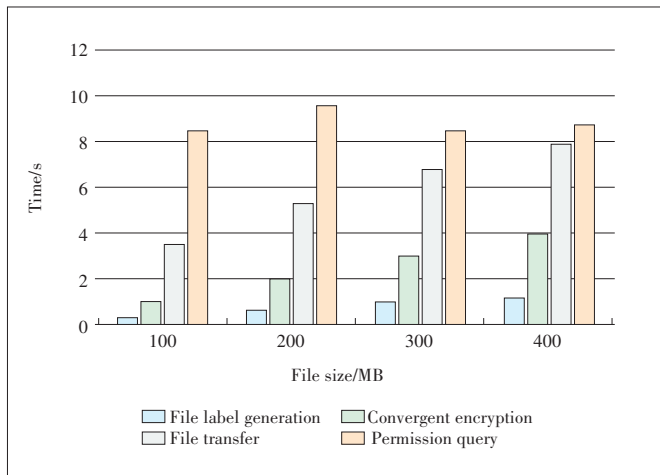
6 Conclusions and Future Work

The differentially authorized deduplication system based on blockchain system proposed in this paper writes the user permission table and file permissions into the blockchain, using the immutable modification of the blockchain, and it better solves the vulnerability of public cloud servers and private cloud servers. It can also overcome the single point of failure problem caused by the original private cloud server due to centralization.

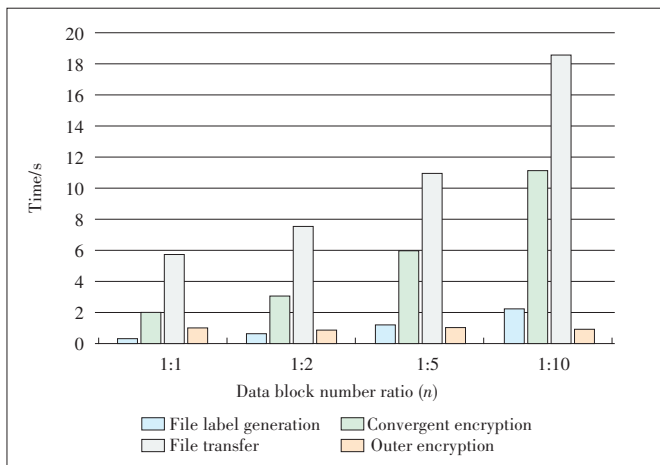
At the same time, the blockchain read and write efficiency is high, and the latest written information will prevail. Therefore, it can solve the problem that the permissions of users and

▼Table 2. Hardware conditions for the implementation environment

Hardware	Setting
Operating system	Ubuntu16.04
CPU brand	AMD
CPU frequency	1.7 GHz
Memory size	8 GB
Network bandwidth	1 000 Mbit/s



▲ Figure 7. Permission query time and main process time



▲ Figure 8. Time required for outer encryption and the main process time

files in the original system cannot be dynamically modified in time, and realize the management of dynamic changes in user and file permissions. At the same time, the experimental results show that the extra cost of our proposed scheme and enhancement scheme is less than 5%, indicating that the system overhead is not significantly increased.

In the future work, we will continue to improve the existing convergent key generation method or explore other encryption schemes to replace the existing convergent key encryption to overcome its vulnerability to offline dictionary attacks.

References

[1] HARNIK D, PINKAS B, SHULMAN-PELEG A. Side channels in cloud services: Deduplication in cloud storage [J]. IEEE security & privacy, 2010, 8

(6): 40 – 47. DOI: 10.1109/MSP.2010.187

[2] DOUCEUR J R, ADYA A, BOLOSKEY W J, et al. Reclaiming space from duplicate files in a serverless distributed file system [C]//22nd International Conference on Distributed Computing Systems. Vienna, Austria: IEEE, 2002: 617 – 624. DOI: 10.1109/ICDCS.2002.1022312

[3] LIU J, ASOKAN N, PINKAS B. Secure deduplication of encrypted data without additional independent servers [C]//22nd ACM SIGSAC Conference on Computer and Communications Security. Denver, USA: ACM, 2015: 874 – 885. DOI: 10.1145/2810103.2813623

[4] LIU X F, SUN W H, LOU W J, et al. One-tag checker: message-locked integrity auditing on encrypted cloud deduplication storage [C]//IEEE Conference on Computer Communications (INFOCOM). Atlanta, USA: IEEE, 2017: 1 – 9. DOI: 10.1109/INFOCOM.2017.8056999

[5] BELLARE M, KEELVEEDHI S, RISTENPART T. Message-locked encryption and secure deduplication [C]//Annual International Conference on the Theory and Applications of Cryptographic Techniques: Berlin/Heidelberg, Germany: Springer, 2013: 296 – 312. DOI: 10.1007/978-3-642-38348-9_18

[6] PUZIO P, MOLVA R, ÖNEN M, et al. CloudDedup: secure deduplication with encrypted data for cloud storage [C]//5th International Conference on Cloud Computing Technology and Science. Bristol, UK: IEEE, 2013: 363 – 370. DOI: 10.1109/CloudCom.2013.54

[7] KEELVEEDHI S, BELLARE M, RISTENPART T. Dupless: server-aided encryption for deduplicated storage [C]//22nd USENIX Conference on Security. Washington D. C., USA: USENIX, 2013: 179 – 194

[8] LI J, CHEN X F, HUANG X Y, et al. Secure distributed deduplication systems with improved reliability [J]. IEEE transactions on computers, 2015, 64(12): 3569 – 3579. DOI: 10.1109/TC.2015.2401017

[9] NAKAMOTO S. Bitcoin: a peer-to-peer electronic cash system [EB/OL]. (2008-10-31)[2020-01-01]. <https://bitcoin.org/bitcoin.pdf>

[10] KHALIL R, GERVAIS A. Revive: rebalancing off-blockchain payment networks [C]//ACM SIGSAC Conference on Computer and Communications Security. Dallas, USA: ACM, 2017: 439 – 453. DOI: 10.1145/3133956.3134033

[11] BHATTACHARYA R, WHITE M, BELOFF N. A blockchain based peer-to-peer framework for exchanging leftover foreign currency [C]//Computing Conference. London, UK: IEEE, 2017: 1431 – 1435. DOI: 10.1109/SAI.2017.8252284

[12] HALEVI S, HARNIK D, PINKAS B, et al. Proofs of ownership in remote storage systems [C]//18th ACM Conference on Computer and Communications Security. Chicago, USA: ACM, 2011: 491 – 500. DOI: 10.1145/2046707.2046765

[13] GERVAIS A, KARAME G O, WÜST K, et al. On the security and performance of proof of work blockchains [C]//ACM SIGSAC Conference on Computer and Communications Security. Vienna, Austria: ACM, 2016: 3 – 16. DOI: 10.1145/2976749.2978341

[14] EYAL I, SIRER E G. Majority is not enough: Bitcoin mining is vulnerable [C]//International Conference on Financial Cryptography and Data Security. Berlin/Heidelberg, Germany: Springer, 2014: 436 – 454. DOI: 10.1007/978-3-662-45472-5_28

[15] KIAYIAS A, RUSSELL A, DAVID B, et al. Ouroboros: a provably secure proof-of-stake blockchain protocol [C]//Advances in Cryptology—CRYPTO 2017, Cham, Switzerland: Springer, 2017: 357 – 388. DOI: 10.1007/978-3-319-63688-7_12

[16] LI W T, ANDREINA S, BOHLI J M, et al. Securing proof-of-stake blockchain protocols [C]//Data privacy management, cryptocurrencies and blockchain technology. Cham, Switzerland: Springer, 2017: 297 – 315. DOI: 10.1007/978-3-319-67816-0_17

[17] ZHENG Z B, XIE S A, DAI H N, et al. An overview of blockchain technology: Architecture, consensus, and future trends [C]//International Congress on Big Data (BigData Congress). Honolulu, USA: IEEE, 2017: 557 – 564. DOI: 10.1109/BigDataCongress.2017.85

[18] SANKAR L S, SINDHU M, SETHUMADHAVAN M. Survey of consensus protocols on blockchain applications [C]//4th International Conference on Advanced Computing and Communication Systems (ICACCS). Coimbatore, India: IEEE, 2017: 1 – 5. DOI: 10.1109/ICACCS.2017.8014672

[19] LI K J, LI H, WANG H, et al. PoV: an efficient voting-based consensus algorithm for consortium blockchains [J]. Frontiers in blockchain, 2020, 3: 11 DOI: 10.3389/fbloc.2020.00011

- [20] LI J, CHEN X F, LI M Q, et al. Secure deduplication with efficient and reliable convergent key management [J]. IEEE transactions on parallel and distributed systems, 2014, 25(6): 1615 – 1625. DOI: 10.1109/TPDS.2013.284
- [21] KAHATE A. Cryptography and network security [M]. New Delhi, India: Tata McGraw-Hill Education, 2013

Biographies

ZHAO Tian is a postgraduate of the Shenzhen Graduate School, Peking University, China. His main research directions are big data applications, blockchain, and distributed storage.

LI Hui (huilihuge@163.com) is currently a professor and Ph.D. supervisor with the Shenzhen Graduate School, Peking University, China. His research fields include cyberspace security and block-chain technology, artificial intelligence and future network systems, distributed storage coding theory and systems, intelligent big data analysis, and data standards.

YANG Xin received the B.Eng. degree from the Department of Computer Science and Engineering, South China University of Technology, China in 2016. She is currently pursuing the Ph.D. degree with the School of Information Science, Peking University, China. She is also the student of the Peng Cheng Laboratory, China. Her research interests include cyber security, future network ar-

chitecture, and distributed storage systems.

WANG Han received the B.Eng. degree from the Department of Communication Engineering, Jilin University of Technology, China in 2017. She is currently pursuing the Ph.D. degree with the School of Information Science, Peking University, China. Her research interests include distributed systems and cyber security.

ZENG Ming received his bachelor's degree from University of Electronic Science and technology, China, majoring in computer science and technology. He is a system architect of ZTE Corporation and has been engaged in software development and architecture design for more than 16 years. His research interests include big data, blockchain and other related technologies fields.

GUO Haisheng received his master's degree from Nanjing University of Aeronautics and Astronautics, China. He is a project manager of ZTE Corporation and has been engaged in software development, architecture design, project management for nearly 20 years. His research interests include blockchain and other related technologies fields.

WANG Dezheng received his master's degree in computer science from Zhejiang University, China. He is a chief engineer of the Center Institute, ZTE Corporation and has been engaged in architecture design for more than 20 years. His research interests include big data and blockchain.



A Novel De-Embedding Technique of Packaged GaN Transistors

Abstract: This paper presents a novel de-embedding technique of packaged high-power transistors. With the proposed technique, the packaged model of the power amplifier (PA) tube can be divided into the frequency independent de-embedded intrinsic device (DID) and the frequency dependent internal parasitic network (IPN), which is of great help in reducing the design complexity of a broadband PA. Different from the conventional technique of parasitic extraction, the proposed technique only requires external measurements. The frequency independent characteristic of DID is verified and the IPN is modeled and calibrated for a 50 W gallium-nitride (GaN) transistor. At last, a broadband Doherty PA is fabricated with the de-embedding technique. According to the measured results, the PA exhibits satisfactory power and efficiency performance.

Keywords: de-embedding; power amplifier; intrinsic device; parasitic network

WEI Xinghui, CHEN Xiaofan,
CHEN Wenhua, ZHOU Junmin

(Tsinghua University, Beijing 100084, China)

DOI: 10.12142/ZTECOM.202102010

<http://kns.cnki.net/kcms/detail/34.1294>.

TN.20210408.1512.002.html, published online
April 09, 2021

Manuscript received: 2020-03-04

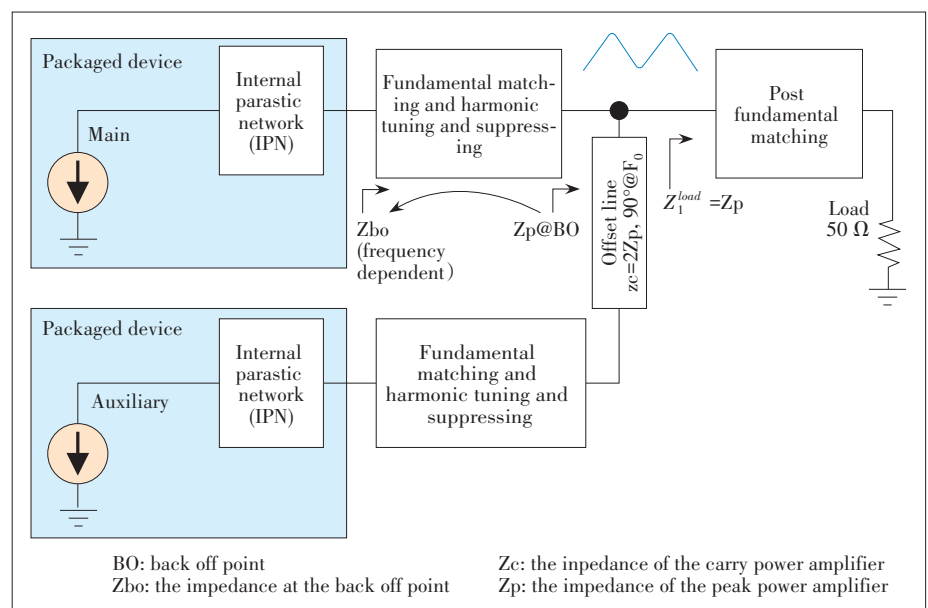
Citation (IEEE Format): X. H. Wei, X. F. Chen, W. H. Chen, et al. "A novel de-embedding technique of packaged GaN transistors," *ZTE Communications*, vol. 19, no. 2, pp. 77 – 81, Jun. 2021. doi: 10.12142/ZTECOM.202102010.

1 Introduction

With the rapid development of wireless communications, especially the emerging 5G, the power amplifier (PA) is expected to handle broad bandwidth, high power, and high efficiency simultaneously. With advantages like large power capacity, high efficiency, and high breakdown voltage, gallium-nitride (GaN) power devices make themselves ideal candidates for such applications^[1].

However, the traditional PA design is based on the packaged model of the power amplifier tube, as shown in Fig. 1. The output impedances of the packaged model vary significantly with frequency due to the parasitic characteristic of the packaged model, which complicates the matching network's design and limits the efficiency and bandwidth. While lots of the improved technologies were proposed to keep the parasitic low^[2-6], it cannot be eliminated, and many parasitic compensation methods have been introduced^[7-9].

For packaged power devices, the use of the parasitic de-embedding network to allow waveforms at the intrinsic plane is required. It is crucial when designing a complex power amplifier and investigating its performance. Several approaches were proposed to extract the extrinsic (package) parasitics, as



▲ Figure 1. Conventional Doherty power amplifier (DPA) design

well as the intrinsic parasitics^[10–14]. A de-embedding network of CGH40010F at 0.9 GHz was presented in Ref. [12]. It was derived through a combination of datasheet parameters and the manufacture's passive packaged models. It has been proved accurate enough at 2.1 GHz, but would result in an unacceptable error at a higher frequency. In Ref. [13], a de-embedding network of CGH40025F was proposed for applications up to 2 GHz. The previous work did not consider the effect of either the feedback capacitance or the inductance due to the gate and drain metallization. At higher frequencies, up to 3 GHz, these effects will impact the de-embedding accuracy. In Ref. [14], the parasitic model was only suitable for narrow bandwidth applications.

This work presents a novel de-embedding technique of packaged GaN transistors, which allows for accurate extraction of the device parasitic. The validation is carried out on a 50 W GaN transistor, which proves the effectiveness of the technique.

2 Proposed De-Embedding Technique

The packaged model of the packaged transistor can be divided into two parts through the de-embedding technique: the frequency independent nonlinear de-embedded intrinsic device (DID) and the frequency dependent linear internal parasitic network (IPN). The IPN can be absorbed into the matching circuit^[14–15], as shown in Fig. 2. In this way, the design of the PA can be simplified and more accurate. In order to validate the technique, a 50 W GaN transistor is used as a validation in this work.

2.1 Parasitic Network Modeling

In order to extract the parasitic parameters of a dual-port network, the conventional method assumes that the IPN is completely unknown, and thus it needs to be measured in two

cases, namely open and short circuits, as shown in Fig. 3a. Measurements should be taken inside the packaged model with special equipment, and C_{ds} , the capacitance between the drain and the source, cannot be calculated.

In this paper, we propose a novel single-port parasitic extraction method as shown in Fig. 3b. The parasitic network is modeled by passive components. The reflection coefficient of the external port Γ_{open} is measured with the internal port of IPN open, and the values of components in the IPN model are obtained when Γ_{mod} is equal to Γ_{open} . This method requires only external measurements to model a parasitic network without internal measurements. As shown in Fig. 4, the input and output parasitic networks of the 50 W GaN transistor are obtained with the proposed technique. The values of the components are listed in Table 1. Fig. 5 gives the small signal simulation results of the networks.

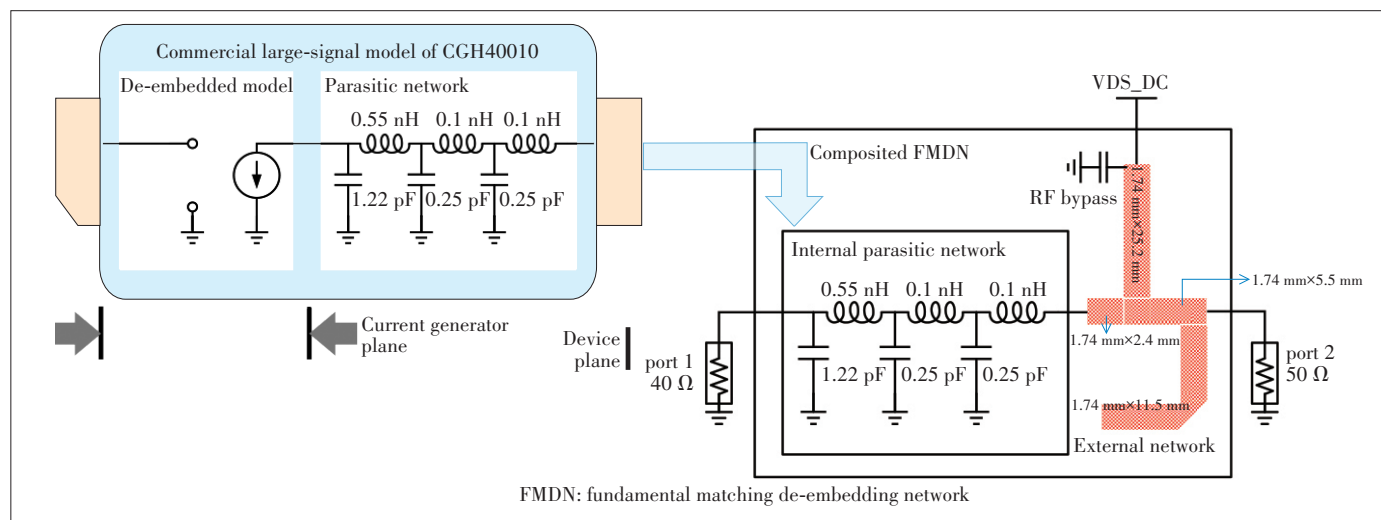
2.2 DID Model Calibration

In the proposed de-embedding technique, the parasitic parameter network is obtained through the large signal parametric model simulation. However, the accuracy of IPN is poorer compared with that of DID. In order to improve the accuracy, the model should be calibrated. We can measure the packaged transistor with a precise fixture, and thus an accurate parasitic parameter model is obtained by parasitic parameter modeling.

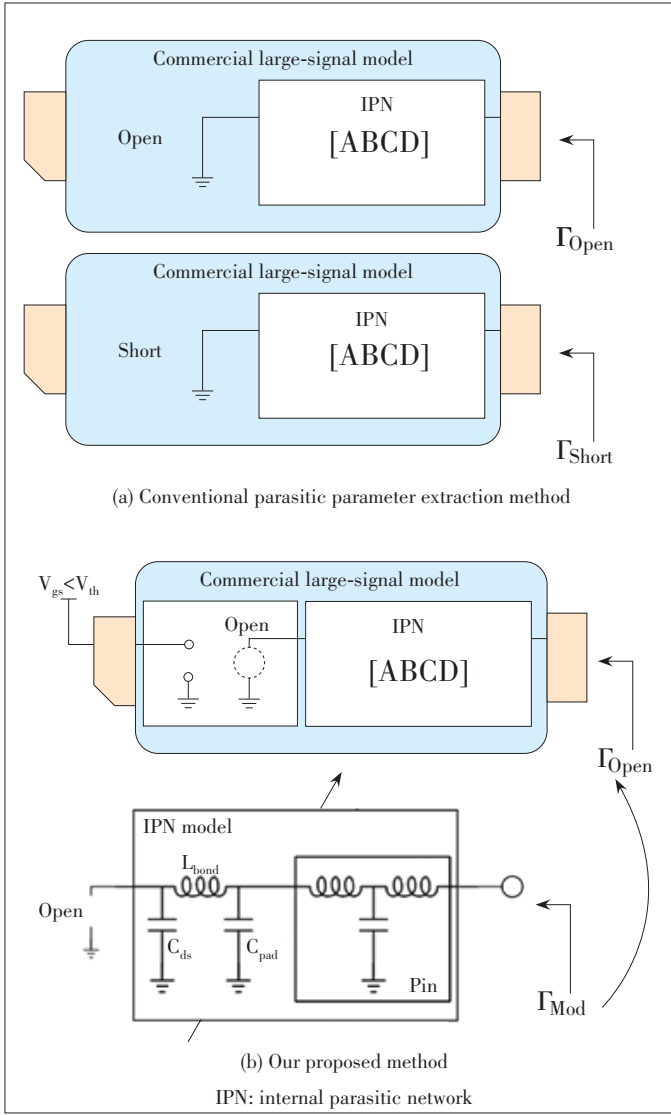
The 50 W GaN transistor is measured using a 6.9 Ω through reflection line (TRL) measurement fixture, as shown in Fig. 6. The calibrated accurate parasitic parameter model is shown in Fig. 7. The values of the components are listed in Table 2.

2.3 DID Model Verification

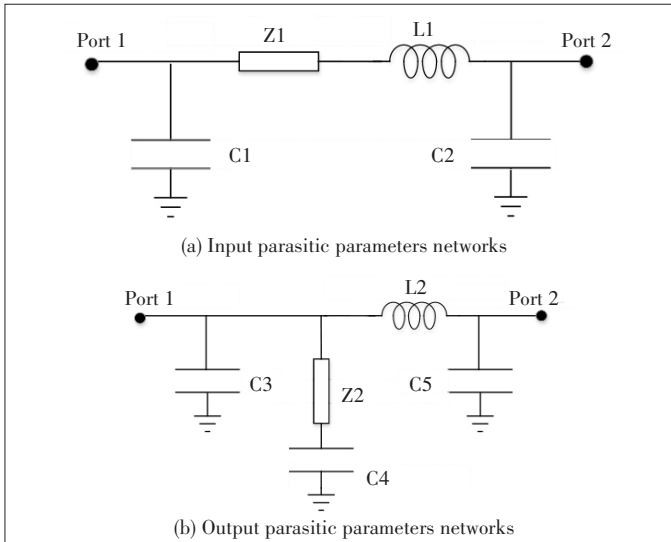
According to the parasitic network model, the DID model can be obtained. Considering the DID model is frequency in-



▲ Figure 2. De-embedding technique of packaged model



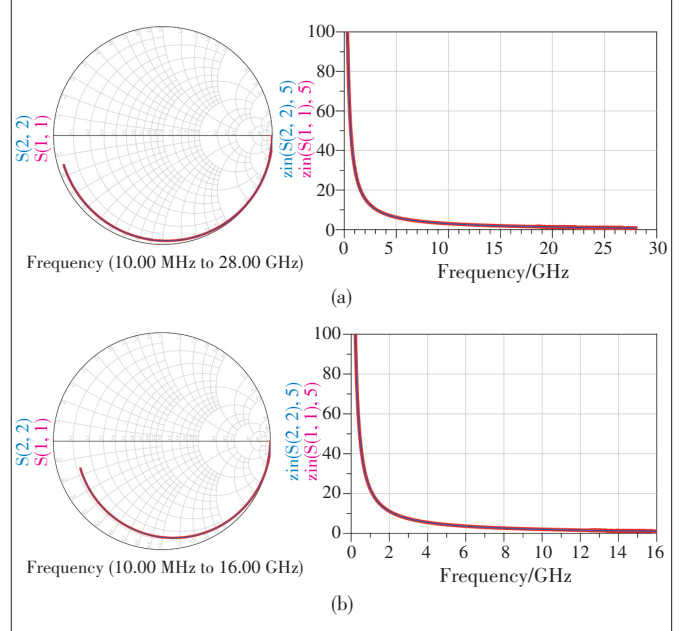
▲ Figure 3. Methods to extract parasitic parameters



▲ Figure 4. Input and output parasitic parameters networks

▼ Table 1. Values of the components in the parasitic model

Component	Value	Component	Value
Z1	0.36 Ω	C2	11.93 pF
Z2	6.05 Ω	C3	1.88 pF
L1	3.44 pH	C4	609.68 fF
L2	5.29 pH	C5	61.41 fF
C1	4.18 fF		



▲ Figure 5. Small signal simulation results of (a) the input network and (b) the output network



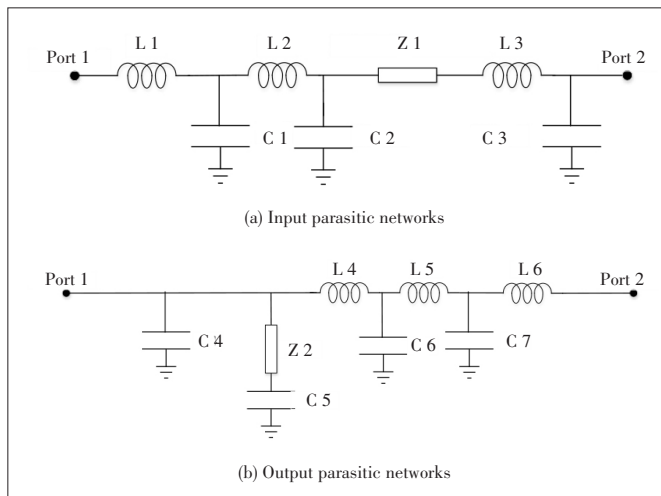
▲ Figure 6. The 6.9 Ω through reflection line (TRL) measurement fixture

dependent, the design of broadband DPA can be significantly simplified. In order to verify the frequency independent characteristics of DID, the DID model of the 50 W GaN power tubes is simulated over a wide frequency range of 500 – 5 000 MHz. The simulation results are shown in Fig. 8. It can be seen from Fig. 6 that the output power, efficiency, and gain performance are independent of frequency in the wide frequency range of 500 – 5 000 MHz. The DID model gives the ideal frequency independent characteristics and the proposed de-embedding

technique is confirmed.

3 DPA Design

In order to verify the effectiveness of the de-embedding technique, a broadband Doherty PA using the 50 W GaN transistor is designed and fabricated utilizing the de-embedded model. The demonstrating 200 W DPA is designed at the range of 1.7 – 2.7 GHz. The fabricated DPA is measured using continuous wave (CW) signals and the measured drain efficiency and power gain are shown in Fig. 9. The designed PA exhibits satisfactory power and efficiency performance, which



▲ Figure 7. Calibrated accurate parasitic parameter model

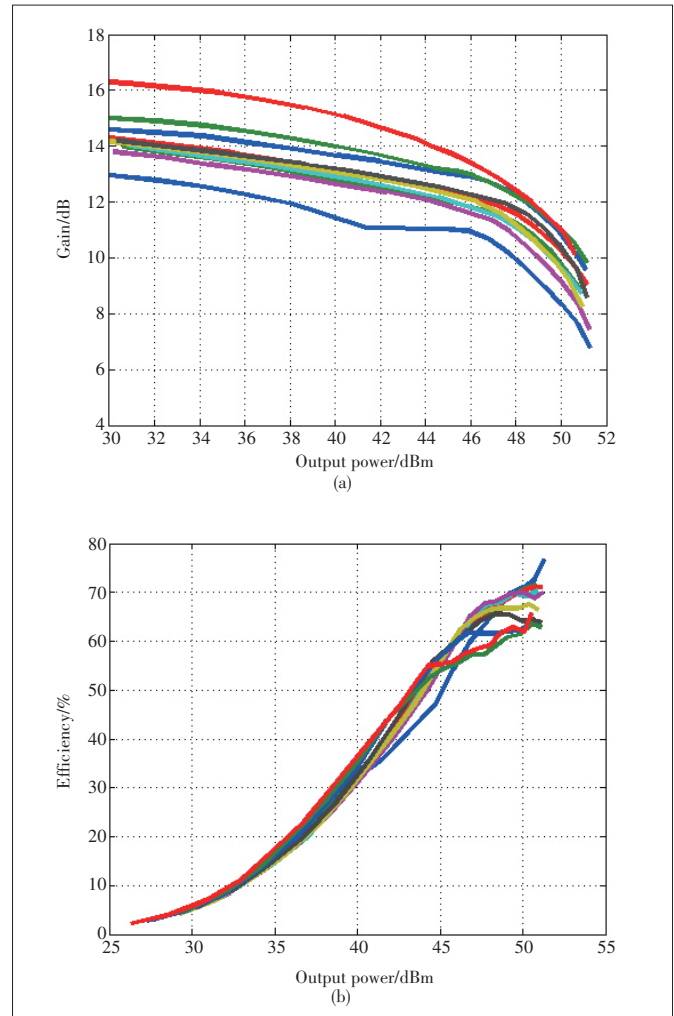
▼ Table 2. Values of the components in the calibrated model

Component	Value	Component	Value
Z1	0.56 Ω	C1	1.90 pF
Z2	26.39 Ω	C2	196.27 fF
L1	128.22 pH	C3	9.97 pF
L2	492.62 pH	C4	3.73 pF
L3	15.04 pH	C5	17.53 fF
L4	460.77 pH	C6	1.47 pF
L5	52.70 pH	C7	571.46 fF
L6	108.9 pH		

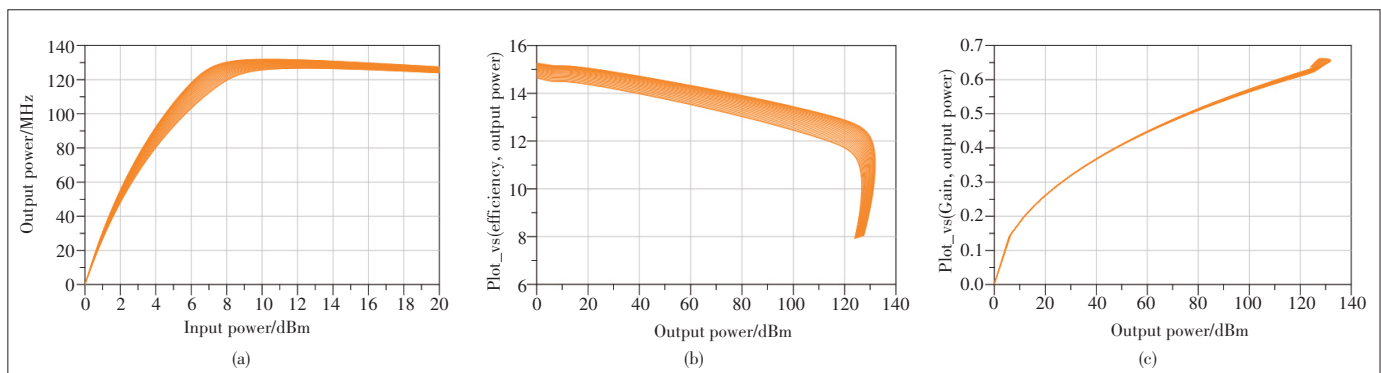
could be a great demonstration of the proposed technique.

4 Conclusions

In this paper, a novel de-embedding technique of packaged GaN transistors is proposed and verified. The technique di-



▲ Figure 9. Measured (a) drain efficiency and (b) gain of the fabricated DPA



▲ Figure 8. Simulation results: (a) output power; (b) efficiency; (c) gain

vides the packaged device into the frequency independent DID and frequency dependent IPN, which only requires external measurements. The IPN can be absorbed into the matching circuit, thus the PA design is only based on the DID. Considering the frequency independent characteristic of DID, the method greatly simplifies the design of efficient wideband PA.

Regarding the rationale, this method could be expanded to other packaged microwave power devices.

References

- [1] PENGELLY R S, WOOD S M, MILLIGAN J W, et al. A review of GaN on SiC high electron-mobility power transistors and MMICs [J]. IEEE transactions on microwave theory and techniques, 2012, 60(6): 1764 – 1783. DOI: 10.1109/MTT.2012.2187535
- [2] HO S W, YOON S W, ZHOU Q, et al. High RF performance TSV silicon carrier for high frequency application [C]//Electronic Components and Technology Conference. Lake Buena Vista, USA, 2008: 1946 – 1952. DOI: 10.1109/ECTC.2008.4550249
- [3] WU J H, ALAMO J ADEL. Fabrication and characterization of through-substrate interconnects [J]. IEEE transactions on electron devices, 2010, 57(6): 1261 – 1268. DOI: 10.1109/TED.2010.2045671
- [4] DONG S, TASKER P J, WANG G, et al. A practicable de-embedding network at higher frequency of a packaged GaN device [C]//IEEE MTT-S International Wireless Symposium (IWS). Shanghai, China: IEEE, 2016: 1 – 4. DOI: 10.1109/IWS.2016.75854735
- [5] KAKKAD P, AGRAWAL E, RAWAT K. De-embedded model based class-E power amplifier using waveform engineering [C]//International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, India, 2017: 1 – 4. DOI: 10.1109/ICCCNT.2017.8204012
- [6] DONG S W, LEES J, TASKER P J. A de-embedding network at higher frequency and its error analysis [C]//2016 IEEE International Conference on Electronic Information and Communication Technology (ICEICT). Harbin, China: IEEE, 2016: 512 – 514. DOI: 10.1109/ICEICT.2016.7879754
- [7] LEE Y S, LEE M W, JEONG Y H. High-efficiency class-F GaN HEMT amplifier with simple parasitic-compensation circuit [J]. IEEE microwave and wireless components letters, 2008, 18(1): 55 – 57. DOI: 10.1109/LMWC.2007.912023
- [8] LIU C, CHENG Q F. A novel compensation circuit of high-efficiency concurrent dual-band class-E power amplifiers [J]. IEEE Microwave and Wireless Components Letters, 2018, 28(8): 720 – 722. DOI: 10.1109/LMWC.2018.2842686
- [9] SAYED A S, AHMED H N. Wideband high efficiency power amplifier design using precise high frequency GaN-HEMT parasitics modeling/compensation[C]//2019 IEEE Topical Conference on RF/Microwave Power Amplifiers for Radio and Wireless Applications (PAWR). Orlando, FL, USA: IEEE, 2019: 1-4. DOI: 10.1109/PAWR.2019.8708722
- [10] DAMBRINE G, CAPPY A, HELIODORE F, et al. A new method for determining the FET small-signal equivalent circuit [J]. IEEE transactions on microwave theory and techniques, 1988, 36(7): 1151 – 1159. DOI: 10.1109/22.3650
- [11] YUAN Y, ZHONG Z, GUO Y X, et al. A novel hybrid parameter extraction method for GaAs/GaN HEMT modeling with electromagnetic analysis [C]//2015 Asia-Pacific Microwave Conference (APMC). Nanjing, China: IEEE, 2015: 1 – 3. DOI: 10.1109/APMC.2015.7413224
- [12] TASKER P J, BENEDIKT J. Waveform inspired models and the harmonic balance emulator [J]. IEEE microwave magazine, 2011, 12(2): 38 – 54. DOI: 10.1109/MMM.2010.940101
- [13] CHEN K L, PEROULIS D. Design of highly efficient broadband class-E power amplifier using synthesized low-pass matching networks [J]. IEEE transactions on microwave theory and techniques, 2011, 59(12): 3162 – 3173. DOI: 10.1109/TMTT.2011.2169080
- [14] KURODA K, ISHIKAWA R, HONJO K. Parasitic compensation design technique for a C-band GaN HEMT class-F amplifier [J]. IEEE transactions on microwave theory and techniques, 2010, 58(11): 2741 – 2750. DOI: 10.1109/TMTT.2010.2077951
- [15] WU D Y T, BOUMAIZA S. A modified Doherty configuration for broadband amplification using symmetrical devices [J]. IEEE transactions on microwave theory and techniques, 2012, 60(10): 3201 – 3213. DOI: 10.1109/TMTT.2012.2209446

Biographies

WEI Xinghui (784999315@qq.com) received the B.S. degree in electronic information engineering from Xidian University, China in 2017, and the M.Sc. Degree in electronics and communication engineering from Tsinghua University, China in 2020. Her research interests include the technologies of high efficiency and high linearity Doherty power amplifiers.

CHEN Xiaofan received the B.S., M.Sc., and Ph.D. degrees in electronic engineering from Tsinghua University, China in 2005, 2014, and 2017, respectively. From 2005 to 2011, he was a microwave engineer with Tongfang Co., Ltd., China, where he was also with the Broadcasting Transmitters Group and designed high-power amplifiers. He is currently a Post-Doctoral Fellow with the Department of Electronic Engineering, Tsinghua University. His current interests include broadband/dual-band RF power amplifier (PA) design, linearization of RF PAs, and multimode/multiband transmitting systems.

CHEN Wenhua received the B.S. degree in microwave engineering from the University of Electronic Science and Technology of China in 2001, and the Ph. D. degree in electronic engineering from Tsinghua University, China in 2006. From 2010 to 2011, he was a Post-Doctoral Fellow with the Intelligent RF Radio Laboratory, University of Calgary, Canada. He is currently an Associate Professor with the Department of Electronic Engineering, Tsinghua University. He has authored or co-authored over 150 journal and conference papers. His current research interests include power-efficiency enhancement for wireless transmitters, PA predistortion, and smart antennas.

ZHOU Junmin received the B.S. degree in College of Optoelectronic Engineering from Chongqing University, China. His current research interests include broadband high-efficiency RF power amplifier design, and research on integrated passive device.



Flexible Multiplexing Mechanism for Coexistence of URLLC and EMBB Services in 5G Networks

Abstract: 5G mobile networks are envisioned to support both evolved mobile broadband (eMBB) and ultra-reliable and low latency communications (URLLC), which may coexist and interfere with each other in the same service cell in many scenarios. In this paper, we propose a dynamic 2-dimension bitmap resource indication to cancel eMBB services with a finer uplink cancellation granularity and a lower probability of false cancellation. Meanwhile, a resource indication based power control method is introduced to dynamically indicate different power control parameters to the user equipment (UE) based on different time-frequency resource groups and the proportion of overlapping resources, by which the reliability of URLLC transmission is guaranteed while the impact on the performance of the eMBB service is minimized. Furthermore, a dynamic selection mechanism is proposed to accommodate the varying cases in different scenarios. Extensive system level simulations are conducted and the results show that about 10.54% more URLLC UE satisfy the requirements, and the perceived throughput of eMBB UE is increased by 23.26%.

Keywords: 5G; eMBB; power control; resource indication; uplink cancellation; URLLC

XIAO Kai^{1,2}, LIU Xing^{1,2}, HAN Xianghui¹,
HAO Peng¹, ZHANG Junfeng¹,
ZHOU Dong¹, WEI Xingguang¹

(1. ZTE Corporation, Shenzhen 518057, China;
2. State Key Laboratory of Mobile Network and
Mobile Multimedia, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202102011

<http://kns.cnki.net/kcms/detail/34.1294.TN.20210419.1612.002.html>, published online
April 20, 2021

Manuscript received: 2021-02-18

Citation (IEEE Format): K. Xiao, X. Liu, X. H. Han, et al. "Flexible multiplexing mechanism for coexistence of URLLC and eMBB services in 5G networks," *ZTE Communications*, vol. 19, no. 2, pp. 82 – 90, Jun. 2021. doi: 10.12142/ZTECOM. 202102011.

1 Introduction

Wireless communication services cover more and more application scenarios with the development of social digitization. Among them, enhanced mobile broadband (eMBB), ultra-reliable and low latency communication (URLLC) and massive machine type of communication (mMTC) have become three major scenarios supported by 5G systems^[1-2]. Large flow mobile broadband services such as ultra-high definition and 3D video are mainly included in eMBB, providing the ultimate communication experience for people. The unmanned driving, industrial automation, etc. are mainly covered by URLLC, requiring low-latency data and an extremely reliable connection, i. e., one-way latency up to 1 ms with 10^{-5} outage probability. The scene of large-scale machine communication in the Internet of Things is mainly supported by mMTC^[3-4]. From the current development trend, 5G mainly focuses on eMBB for basic daily needs and URLLC for emerging industries. Generally, the priority of a URLLC service is higher than that of an eMBB

service^[5-7]. In the practical application, eMBB services and URLLC services will appear in the same network, and the scheduling of a URLLC service by next-generation Node B (gNB) will inevitably conflict with the eMBB service as the triggering of the URLLC service is sporadic^[8]. Once the URLLC service arrives, the gNB shall allocate appropriate uplink (UL) resources to it as soon as possible to meet the stringent latency requirements. However, the resources may have been allocated to uplink data transmission for eMBB in advance. The latency and reliability requirement of URLLC transmission can hardly be guaranteed due to no resource being used within a certain time interval^[9-10]. Therefore, it is valuable to study how to multiplex the transmission resources between a URLLC service and an eMBB service.

Currently, power control and UL cancellation indication (CI) mechanisms are introduced into the Third Generation Partnership Project (3GPP) protocol as two independent and basic solutions to URLLC and eMBB service multiplexing^[11]. However, these baseline methods have some inherent disad-

advantages. More specifically, the baseline cancellation indication (BCI) method is based on the semi-static 2-dimension (2D) bitmap implementation. However, the URLLC service is dynamic and mutative, and the semi-static pattern indication is difficult to meet the changing service requirements^[12]. For the baseline power control (BPC) method, the power boosting is based on a relatively fixed value. For some cases where the proportion of overlapping resources over all scheduled resources is very small, the fixed setting of power boosting value will cause power waste and degrade the eMBB performance^[13]. On the other hand, the BPC cannot further boost power to protect URLLC transmission in the case of poor channel quality^[14]. This paper is mainly to solve these existing technical problems mentioned above.

In this paper, a dynamic pattern cancellation indication (DPCI) is proposed for making up the shortcomings of BCI. The proposed DPCI method enhances the current 2D bitmap pattern from semi-static to dynamic, so that the indication pattern can be adjusted flexibly according to the service arrival to obtain a more accurate indication. This can reduce the false indication and protect the eMBB service. Then, a resource occupancy based power control (ROPC) is proposed to enhance the current BPC method. Based on ROPC, it becomes possible for gNB to dynamically indicate different power control parameters to user equipment (UE) on different sets of time-frequency resources, which will further ensure URLLC transmission performance and protect the normal eMBB transmission. Furthermore, a dynamic selection of DPCI and ROPC is proposed. Because the scene is complex in the real deployment, each multiplexing mechanism has its advantages and disadvantages, and a combination of these mechanisms will get more robust and better performance.

The paper is organized as follows. Section 2 introduces the service multiplexing system model. In Section 3, the proposed design for DPCI, ROPC and the dynamic selection mechanism is described in detail. Extensive system level simulation results are introduced in Section 4, and the conclusions of this paper are given in Section 5.

2 Service Multiplexing System Model

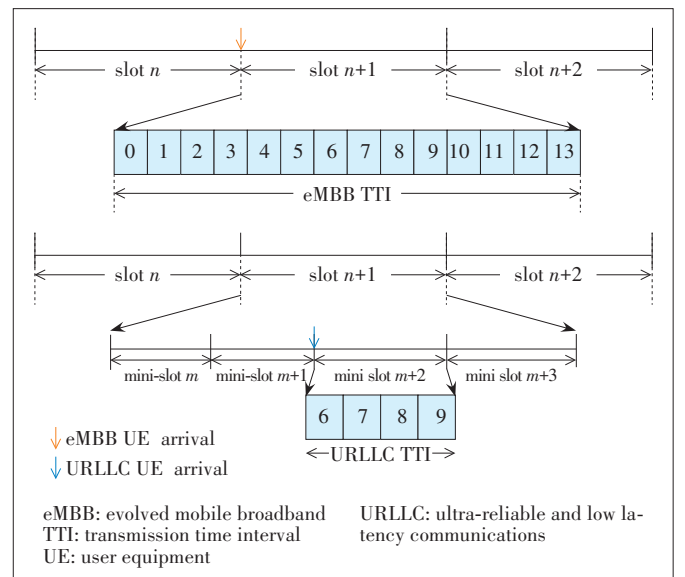
A 5G new radio (NR) uplink system is considered, where there are N cells, each equipped with K_r receiving antennas, and randomly distributed M user devices, each equipped with K_t transmitting antennas. Each cell includes two types of uplink transmission UE: URLLC UE and eMBB UE. The amount of URLLC UE and eMBB UE in each cell is M_{URLLC} and M_{eMBB} respectively, and $M_{\text{URLLC}} + M_{\text{eMBB}} = M$. The packets arrival for each eMBB user device is file transfer protocol (FTP) model 3 with Poisson arrival and the packet size is $B_{\min} - B_{\max}$ bytes with Pareto distribution^[17]. The packet arrival for each URLLC user device is sporadic with an average arrival rate of 1 packet per T ms and the packet size is B bytes. In the system model,

users are distributed indoors and outdoors in a random proportion, and $o\%$ of users are outdoors and $i\%$ of users are indoors, where $o + i = 100$.

A flexible frame structure is adopted for the service multiplexing system model, where URLLC and eMBB UE are scheduled with different transmission time intervals (TTIs). As an example in Fig. 1, the scheduling granularity is set to 14 orthogonal frequency division multiplex (OFDM) symbols for eMBB and 4 OFDM symbols for URLLC in order to achieve a latency reduction. The monitoring periodicity of UL cancellation signaling should be equal to the URLLC physical downlink control channel (PDCCH) monitoring interval, i.e. mini-slot level^[15]. In the frequency domain, the smallest scheduling unit is the resource block (RB) which is composed of 12 resource elements (RE).

For a UL CI based solution, the remaining part of the eMBB transmission is dropped by assuming the phase continuity of UL eMBB transmission cannot be guaranteed. For a UL power control based solution, P dB power boosting of URLLC transmission is assumed in case of overlapping with grant-based eMBB transmission respectively. In our simulation, each PDCCH monitoring occasion occupies one symbol with 32 CCEs. When considering reserving some candidates for eMBB scheduling, the PDCCH search space set configuration for UL cancellation signaling is assumed as the aggregation level (AL) = {1, 2, 4, 8, 16} with corresponding candidate numbers {4, 4, 2, 1, 1} respectively. The AL of the UL cancellation signaling is selected according to a PDCCH channel condition with a target block error rate (BLER) requirement. For a UL cancellation based solution, a group common PDCCH is adopted. In addition, the additional signaling caused by method improvement is carried by the PDCCH.

In this system model, the algorithm of the gNB receiver is



▲ Figure 1. TTI for scheduling URLLC and eMBB UEs

minimum mean square error-interference rejection combining (MMSE-IRC), which adopts MMSE criterion^[16]. The objective function is to minimize the mean square error between the transmitted signal vector \mathbf{s}_1 and the received signal vector linear combination \mathbf{W}_y^H , shown as :

$$\min_{\mathbf{W}} E \left[\left(\mathbf{W}_y^H - \mathbf{s}_1 \right)^H \left(\mathbf{W}_y^H - \mathbf{s}_1 \right) \right], \quad (1)$$

where \mathbf{s}_1 is the signal source symbol of a service cell, y is the signal received by the receiver, and \mathbf{W} is the $Kt \times Kr$ weighted matrix of dimension. When the gradient is used to find the optimal solution, the information of the known interference channel matrix is fully used, and the MMSE-IRC weighting matrix can be obtained as:

$$\mathbf{W}^H = \mathbf{H}_1^H \left(\mathbf{H}_1 \mathbf{H}_1^H + \frac{I_{oc}}{E_s} \mathbf{H}_2 \mathbf{H}_2^H + \frac{N_0}{E_s} \mathbf{I}_{K_i} \right)^{-1}, \quad (2)$$

where \mathbf{H}_1 represents the channel matrix from the service cell to the receiver, \mathbf{H}_2 represents the channel matrix from the interference cell to the receiver, E_s is the average power of the transmitting source symbol, and the noise power and interference power are I_{oc} and N_0 respectively. When there are multiple interference cells, the MMSE-IRC weighting matrix formula can be extended as:

$$\mathbf{W}^H = \mathbf{H}_1^H \left(\mathbf{H}_1 \mathbf{H}_1^H + \sum_n \frac{I_{oc}}{E_s} \mathbf{H}_n \mathbf{H}_n^H + \frac{N_0}{E_s} \mathbf{I}_{K_i} \right)^{-1}. \quad (3)$$

3 Multiplexing Methods

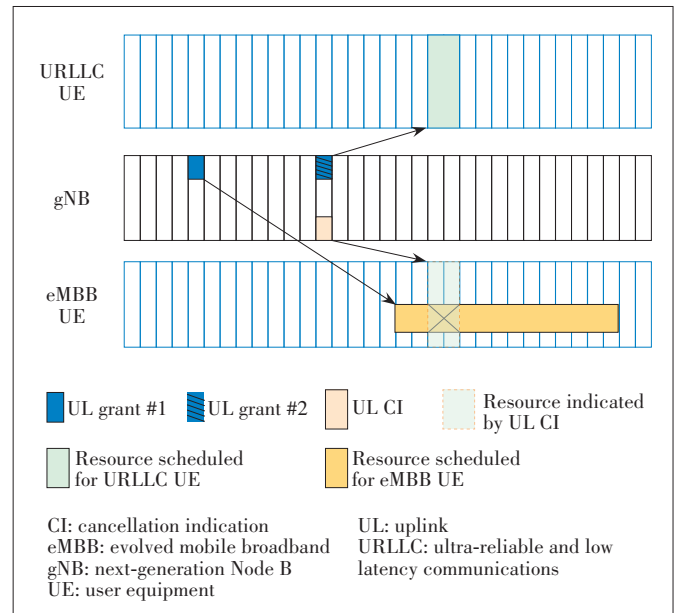
In Fig. 2, an example of BCI for UL multiplexing transmission is shown. The resource for grant-based eMBB and URLLC is scheduled by UL grant #1 and UL grant #2 respectively. Meanwhile, a URLLC resource indication can be transmitted to eMBB UE by the UL CI. The eMBB UE should cancel its uplink transmission when the UL CI is detected. In this case, the resource region in which the URLLC cancellation resource indicated by the UL resource indication is named as Reference Uplink Resource (RUR).

A BPC for UL multiplexing transmission is another alternative, i.e. boosting the URLLC transmission power on the colliding resources. When one user device is transmitting uplink data via an eMBB physical uplink shared channel (PUSCH) and another user device has urgent URLLC data to be sent on the same resource, relatively higher power can be applied than for the case without overlapping eMBB transmission. For the power control scheme, the gNB can still receive the eMBB transmissions. The URLLC transmission may have interference on the eMBB transmission, but it can still be possible for the gNB to decode the eMBB transmission block correctly without retransmission.

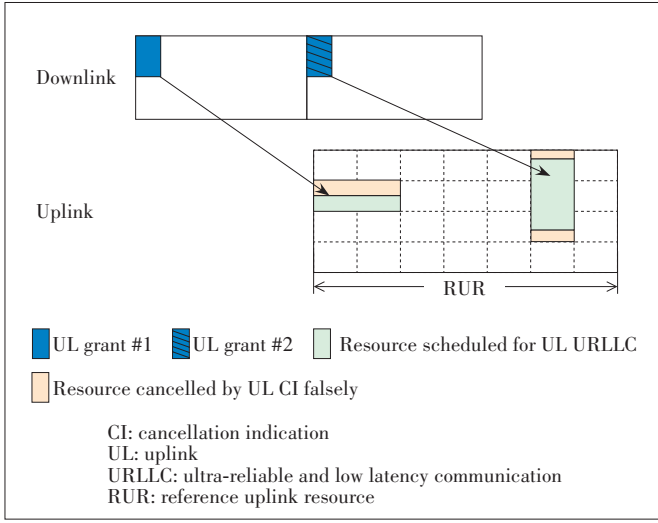
3.1 Design for DPCI

URLLC UE is randomly distributed at both the center and edge in a cell. For cell center UE, it is reasonable to allocate a “thin-tall” type of time-frequency resource to latency reduction. While for cell edge UE which is subject to greater inter-cell interference and larger path loss, higher power is expected to meet the reliability requirement. It tends to allocate less frequency resources to such UE due to a power limitation issue, and instead more symbols have to be scheduled for them. Then, a “fat-short” type of time-frequency resource is more suitable for cell edge UE. As a result, dynamic scheduled resources are different between cell center UE and cell edge UE. Fig. 3 shows an example of the gNB allocating resources to cell center UE and cell edge UE for the BCI method. In Fig. 3, the RUR is divided into 7×4 resource sub-blocks by a 2D bitmap with a size of 28 bits. The first green resource in RUR is allocated to a cell edge UE by gNB, which occupies 2 resource sub-blocks in the time domain and only 1 in the frequency domain. The second green resource block in the RUR is allocated to cell center UE by gNB, which occupies only 1 resource sub-block in the time domain and 3 resource sub-blocks in the frequency domain. Furthermore, different use cases are identified for URLLC services, such as power distribution, factory automation, transport industry, etc. In accordance with different traffic characteristics, different resource allocations are required for different use cases. For example, a service with a larger packet size and higher reliability requirement expects more resource allocation compared with a smaller packet size or lower reliability requirement. In the actual network deployment, various URLLC services could coexist in one cell.

The resource indication pattern under BCI is configured



▲ Figure 2. An example of baseline cancellation indication (BCI) multiplexing method



▲ Figure 3. Allocated resources to cell center UE and cell edge UE for baseline cancellation indication (BCI) method

semi-statically, e.g., 4×7 resource sub-blocks as shown in Fig. 3. A semi-static 2D bitmap pattern cannot provide a flexible frequency domain granularity indication, which causes a large number of eMBB transmissions to be cancelled falsely. From the overall performance, the loss outweighs the gain. In order to make better use of spectrum resources in different scenarios for URLLC and reduce the probability of eMBB being canceled by error, a dynamic configuration of the resource indication pattern should be supported. Instead of using the indication bits to indicate the frequency resource occupation uniformly for all time domain occasions, only the time domain occasions occupied by URLLC PUSCH are valid for further frequency indication in DPCI. Thus, the occupied time domain occasions are indicated firstly, and the time-frequency resource corresponding to the occupied time domain occasions is indicated by a dynamic 2D bitmap in DPCI. More specifically, the bit construction of DPCI is illustrated as follows.

- Q bits are used for indicating which time domain occasion is occupied, where “ Q ” equals the number of occasions in the time domain per RUR.

- $C_{m \times n}$ is a 2D bitmap for frequency domain indication, i.e., the occupied time domain occasions are divided into “ $a \times b$ ” portions, and each portion is indicated by a bit in the 2D bitmap, wherein a represents the number of occupied time domain occasions and b represents the frequency domain granularity. Both of them are determined according to the indication of Q bits dynamically.

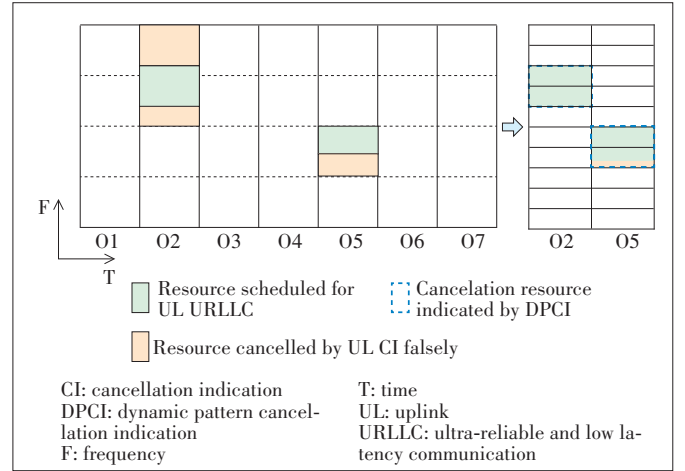
As shown in Fig. 4, the total number of occasion is 7, corresponding to $Q=7$; The number of occasion with scheduled resource for URLLC is 2, corresponding to $a=2$; The O2 and O5 occupied by URLLC service are divided into 10 parts in the frequency domain, corresponding to $b=10$. The total bit-number is 27. Compared with BCI, which requires 28 bits, this method can make the frequency domain indication granularity

(FDIG) finer with the same number of bits. This could reduce the false cancellation probability for better protection of eMBB PUSCH. As shown in Table 1, as long as the number of occupied time domain occasions (OTDOs) is less than 5, the minimum frequency domain indication granularity for each time domain occasion of DPCI is finer than that of BCI.

3.2 Design for ROPC

In order to ensure the flexibility of URLLC transmission, gNB should schedule the most appropriate time-frequency resources for URLLC UE without caring whether it overlaps with eMBB transmission. For BPC, once the resource of URLLC overlaps with that of eMBB, URLLC UE will perform P dB power boosting, i.e., 6 dB. However, the eMBB transmission will cause quite different interferences on URLLC time-frequency resources under some situations. For example, as shown in Fig. 5, the transmission power of different eMBB UE is different in the same RUR, and the proportion of overlapping resources to the total resources of URLLC services is different in different RURs. A fixed value of power boosting will not only lead to insufficient or serious power waste for URLLC transmission, but also affect the transmission performance of normal eMBB services.

Compared with fixed power adjustment according to resource multiplexing, the following two points are enhanced in ROPC: 1) defining different power control parameters for dif-



▲ Figure 4. Resource indicated by dynamic 2D bitmap

▼ Table 1. Minimum indication granularity with different numbers of the occupied time domain occasions

BCI	the number of OTDOs	1 - 7						
	FDIG	1/4						
DPCI	the number of OTDOs	1	2	3	4	5	6	7
	FDIG	1/21	≤1/10	1/7	≤1/5	≤1/4	≤1/3	1/3

BCI: baseline cancellation indication
DPCI: dynamic pattern cancellation indication
OTDO: occupied time domain occasion
FDIG: frequency domain indication granularity

ferent resource groups; 2) boosting power according to the proportion of overlapping resources to the total resources of URLLC services. More details are provided on the above two enhancements in the following subsections.

3.2.1 Different Power Control Parameters for Different Resource Groups

In an RUR, multiple groups of a time-frequency resource can be indicated by gNB to URLLC UE. Different groups of time-frequency resources correspond to different power control parameter sets. The URLLC transmission power is determined according to the power control parameter set corresponding to the group of time-frequency resource which overlaps with eMBB.

As shown in Fig. 6, the control information of ROPC includes at least one time-frequency resource indication field. Each time-frequency resource indication field can indicate a group of time-frequency resources. The power control parameters for each group of time-frequency resources will be configured via radio resource control signaling. If the resource scheduled for URLLC transmission overlaps with more than one group of time-frequency resources, transmission power will be calculated based on each power control parameter respectively, and a higher one or an average value will be selected.

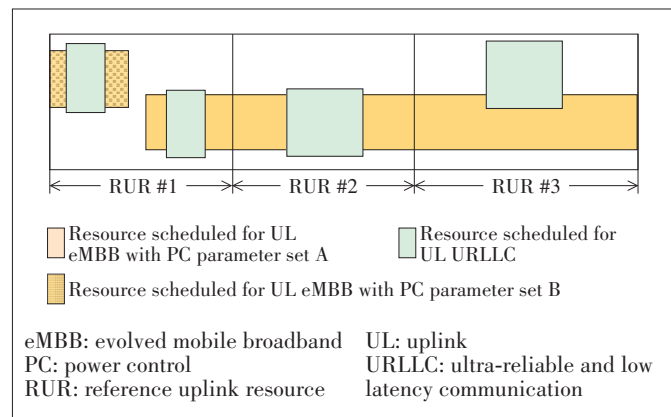
3.2.2 Boosting Power Based on Overlapping Resource Proportion

Multiple overlapping resource proportion thresholds are defined in advance, among which the overlapping resource proportion is defined as the proportion of overlapping resources to URLLC resources. The threshold includes 10%, 40% and 80%. Table 2 defines a mapping relationship between the actual overlapping resource proportion x and a power promotion value. For example, the overlapping resource proportion is 40%, if the URLLC resource is 10 RB and the overlapping resource is 4 RB. In such cases, the transmission power of URLLC will be increased by 3 dB.

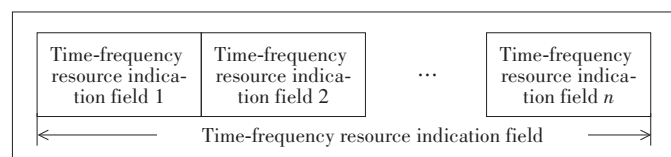
The execution procedure of ROPC is summarized as follows:

- 1) The gNB determines which group of URLLC time-frequency resources overlaps with the eMBB.
- 2) The gNB computes the overlap proportion between each group of URLLC time-frequency resources and eMBB time-frequency resources.
- 3) The gNB sends the control information carrying the index of power control parameters corresponding to each resource group according to the calculation result of the second step.
- 4) The URLLC UE receives and decodes the control information of ROPC.
- 5) The URLLC UE determines the power value to be enhanced for each group of time-frequency resources according to the index in the time-frequency resource indication field.

The introduction of overlapping resource proportion enables URLLC UE to adjust the transmission power to be optimal, while limiting the interference for the eMBB transmission.



▲ Figure 5. Various situations of overlapping between the URLLC physical uplink shared channel (PUSCH) and the eMBB PUSCH



▲ Figure 6. Time-frequency resource indication field

▼ Table 2. Power boosting value according to actual overlapping resource proportion

Index	Actual Overlapping Resource Proportion x	Power Boosting/dB
0	$x \leq 10\%$	0
1	$10\% < x \leq 40\%$	3
2	$40\% < x \leq 80\%$	6
3	$x > 80\%$	9

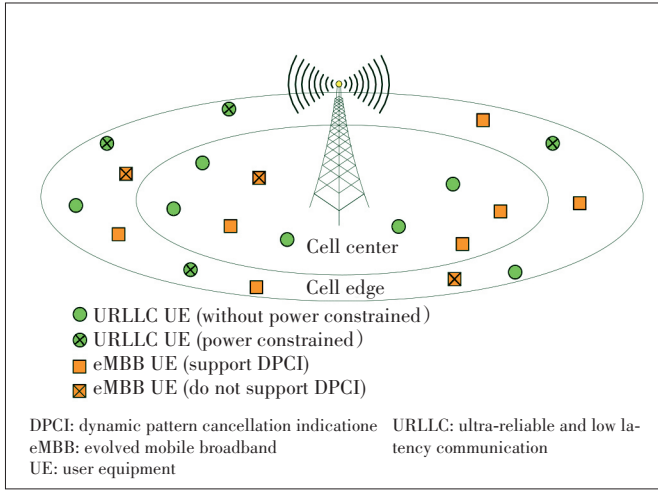
What's more, interpreting different power control parameters for different transmission time-frequency resources further improves the accuracy of power control.

3.3 Design for Dynamic Selection of DPCI and ROPC

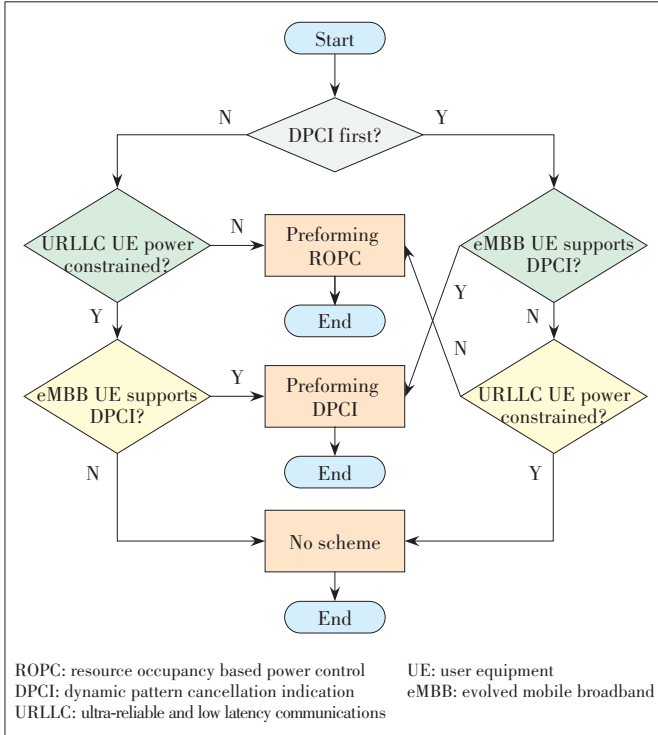
In this subsection, we combine DPCI with ROPC to obtain better performance. Two methods for the multiplexing application can be considered in the service multiplexing system model, and the most suitable method is selected for execution in one TTI. Fig. 7 shows an example for dynamic selection multiplexing methods based on the location and function configuration of the UE. There are both URLLC UE and eMBB UE in a cell. For URLLC UE, one kind of UE can perform ROPC without the power constrained, and the other kind cannot perform PC with the power constrained. For eMBB UE, one kind of UE supports DPCI and the other kind does not support DPCI. In practical application, there are three options for multiplexing scheduling: ROPC, DPCI and no scheme, and Fig. 8 shows the selection procedure of the multiplexing mechanism.

4 Simulation Results

In this section, system level simulation results based on dif-



▲ Figure 7. An example of dynamic selection



▲ Figure 8. The dynamic selection procedure

ferent multiplexing methods are provided. The simulation mainly includes the following four aspects:

- Percentage of time domain occasions occupied by URLLC per RUR;
- Minimum boosted power value for URLLC to meet the reliability requirement;
- System performance comparison for different multiplexing methods;
- System performance comparison for the dynamic selection mechanism and baseline methods.

The above four simulation aspects are based on a service multiplexing system model proposed in Section 2. The details

of simulation assumptions are listed in Table 3.

4.1 Percentage of Time Domain Occasions Occupied by URLLC per RUR

To compare the BCI and DPCI method, the percentage of the number of time domain occasions occupied by the URLLC per RUR is evaluated via system-level simulation. The duration of RUR is set as 1 slot. Within each RUR, there are 7 time domain occasions, and each of them has 2 OFDM symbols. As shown in Fig. 9, the number of occasions actually scheduled for URLLC transmission is relatively small. In each cell load setup scenario, the ratio of cases that the occupied RURs contain less than 3 time domain occasions occupied by URLLC is more than 94%. Together with the analysis in Table 1, we can infer that a finer frequency domain indication granularity can be expected by DPCI in most cases. In other words, DPCI has more accurate indication granularity.

4.2 Minimum Boosted Power Value for URLLC to Meet the Reliability Requirement

To prove that it is reasonable to grade multiple values for power boosting, we intercept 100 times of conflict between URLLC and eMBB in a system simulation. We repeated 10 times for each conflict with different power values, and se-

▼ Table 3. System-level simulation assumptions

Parameters	Value
Carrier frequency	4 GHz
Simulation bandwidth	40 MHz
SCS	30 kHz
Channel model	UMa in TR 38.901
Antenna configuration	4 receiving antenna ports 2 transmitting antenna ports
gNB receiver	MMSE-IRC
Cell load setup	$K_{\text{eMBB}}: 5, 10, 20, K_{\text{URLLC}}: 5, 10, 20$ $\Omega = (K_{\text{URLLC}}, K_{\text{eMBB}})$
TTI configuration	URLLC: 2, 3, 4 OFDM symbols eMBB: 14 OFDM symbols
HARQ	Max number of transmissions=4 with target BLER=0.01% (URLLC) or 10%(eMBB)
Traffic model	eMBB: • Packet arrival per UE: FTP Model 3 • Packet size: 50 – 600 bytes URLLC: • Packet arrival per UE: periodic with arrival rate of 1 packet per 2 ms • Packet size: 32 bytes
UE distribution	80% of users are outdoors 20% of users are indoors
eMBB UE function configuration	90% of users support DPCI 10% of users do not support DPCI

DPCI: dynamic pattern cancellation indication
eMBB: evolved mobile broadband
FTP: file transfer protocol
BLER: block error rate
BSgNB: next-generation Node B
HARQ: hybrid automatic repeat request
MMSE-IRC: minimum mean square error-in-
terference rejection combining
SCS: sub-carrier spacing
TTI: transmission time interval
UE: user equipment
uMA: uUrban Macro
URLLC: ultra-reliable and low latency communications

lected a minimum power value for URLLC UE to meet the reliability requirement. If URLLC UE cannot meet the reliability requirement, the maximum boosted power value will be selected. For the system simulation, the cell load setup is set $\Omega = (10, 10)$, and the minimum and maximum boosted power are 0 dB and 9 dB, respectively. The duration of RUR is assumed as 1 slot, which contains 4 time domain occasions. The simulation results are shown in Fig. 10.

As shown in Fig. 10, the most suitable boosted power value may not always be 6 dB, and we divide the power increase value into four levels with dotted lines equal to 0, 3, 6 and 9. In this experiment, 6 dB power boosting cannot meet the reliability requirement for URLLC in 17 out of 100 conflicts, and 6 dB power boosting becomes wasteful in 40 out of 100 conflicts. For ROPC, 9 dB can be boosted in scenarios where 6 dB cannot meet URLLC transmission requirements, while 3 dB and 0 dB can be boosted in scenarios of good channel condition quality to save power.

4.3 System Performance Comparison for Different Multiplexing Methods

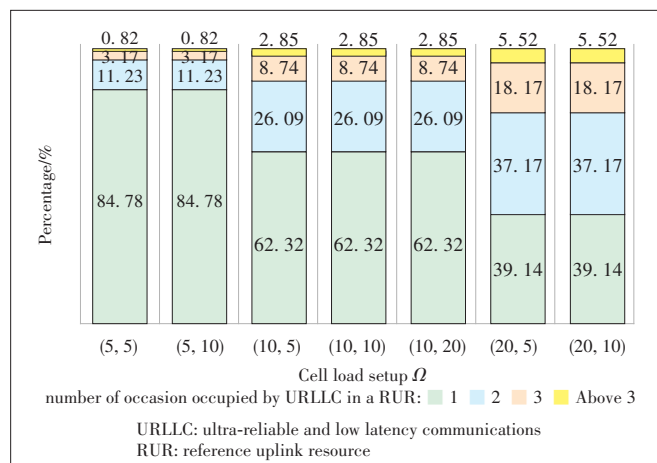
To compare the performance of different multiplexing methods as described above, the performance of the URLLC transmissions and eMBB UE perceived throughput (UPT) are evaluated. The corresponding simulation assumptions are shown in Table 3.

The scheduling granularity is set to 14 OFDM symbols for eMBB and 3 or 4 OFDM symbols for URLLC. For BCI, the 2D bitmap pattern is set as 4×7 , which means that the RUR is divided into 4 parts in the time domain and 7 parts in the frequency domain. For BPC, 6 dB power boosting of URLLC transmission is assumed in case of overlapping with eMBB transmission. For DPCI, 4 bits are used for indicating which time domain occasions is occupied, and 2D bitmap pattern is dynamically set based on the actual number of occasions occupied, such as 1×24 , 2×12 , 3×8 , and 4×6 . For ROPC, the power boosting value is set according to the actual overlapping resource proportion, and it is divided into 4 levels, such as 0 dB, 3 dB, 6 dB, and 9 dB. The system-level simulation results are shown in Fig. 11, Fig. 12 and Table 4. As a reference, URLLC performance of UL inter-UE multiplexing with no scheme is also listed.

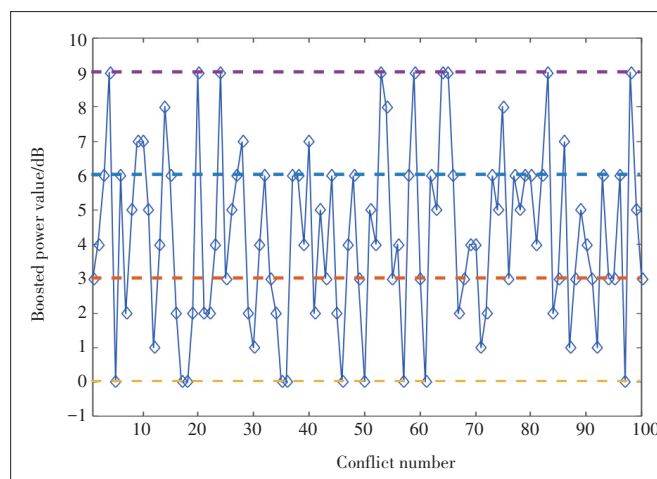
For the performance of eMBB transmission, we can see from Fig. 11 and Fig. 12 that with the increase of cell load, UPT of eMBB transmission shows a downward trend. In all cell load scenarios, ROPC has the largest UPT for eMBB transmission, which is mainly due to the dynamic selection of boosted power value. It can be observed that DPCI has a maximum gain of 13.78% compared with BCI, and ROPC has a maximum gain of 12.50% compared with BPC.

Although the cancellation method has a bigger impact on the eMBB transmission, it can effectively eliminate the interference of eMBB transmission on URLLC transmission. This

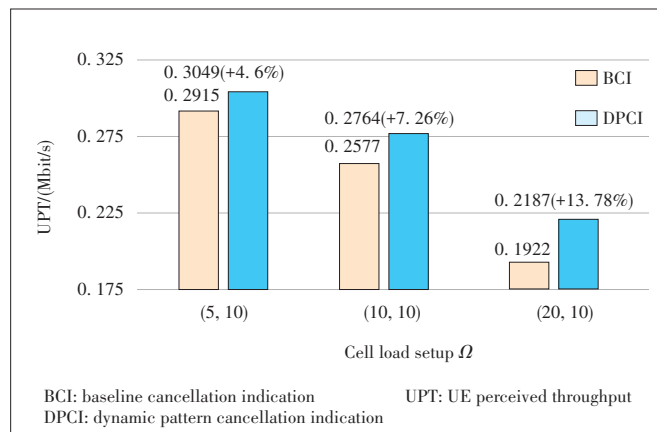
is proved by the simulation results in Table 4, where DPCI and BCI show better performance compared with power control-based methods for the performance of URLLC transmission. From Table 4, we can also see that the performance of the URLLC using DPCI is almost the same as that of BCI.



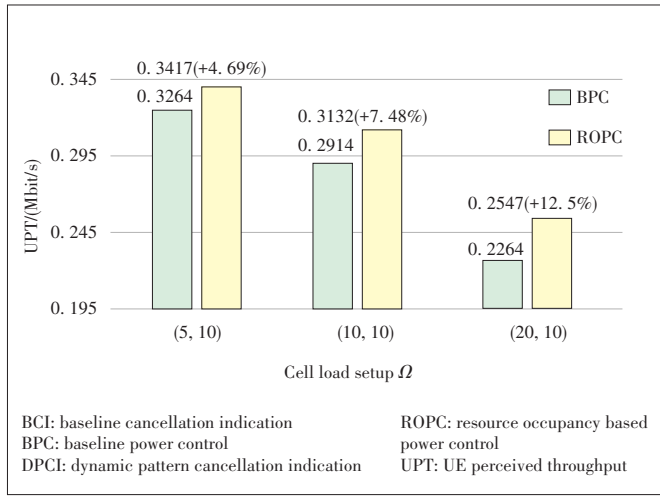
▲ Figure 9. Statistics of time domain occasions occupied by URLLC



▲ Figure 10. Actual boosted power values



▲ Figure 11. UPT of evolved mobile broadband (eMBB) transmission for BCI and DPCI



▲ Figure 12. UPT of evolved mobile broadband (eMBB) transmission for BPC and ROPC

▼ Table 4. Percentage of UE satisfying reliability and latency requirements for URLLC transmission in different multiplexing methods

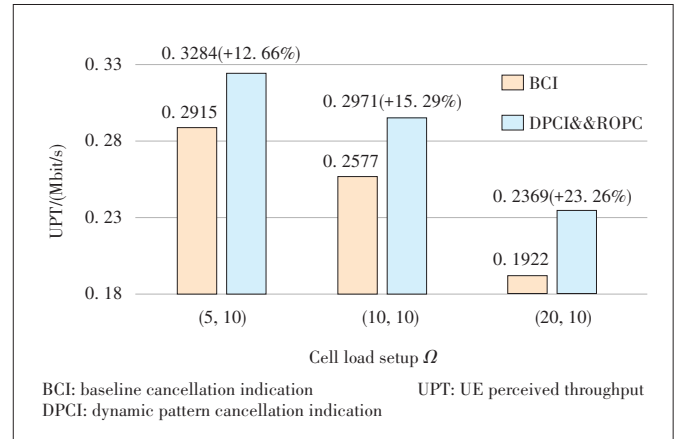
Multiplexing Method	$\Omega = (5,10)$	$\Omega = (10,10)$	$\Omega = (20,10)$
No scheme/%	84.37	78.64	66.71
BCI/%	93.33	89.87	80.64
DPCI/%	93.27	89.64	80.62
BPC/%	87.78	83.97	73.84
ROPC/%	88.34	86.47	76.77

This is because both DPCI and BCI can cancel the eMBB transmission, which means no interference on URLLC transmission. From Table 4, the URLLC UEs' satisfaction rate of ROPC is increased by 2.93% compared with BPC, which is mainly because ROPC can dynamically boost the power by 9 dB under the condition that 6 dB cannot ensure the normal URLLC transmission.

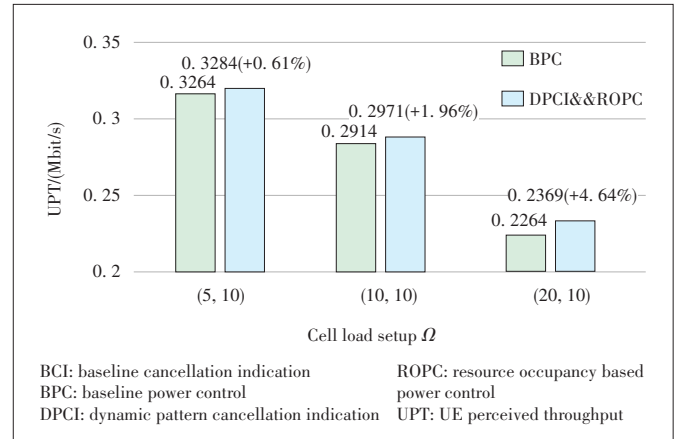
4.4 System Performance Comparison for the Dynamic Selection Mechanism and Baseline Methods

In this subsection, we provide the simulation results for the dynamic selection of DPCI and ROPC method. The simulation assumption is the same as that described in subsection 4.3. The system-level simulation results about UPT of eMBB transmission and the performance of the URLLC transmissions are shown in Fig. 13, Fig. 14 and Table 5.

As shown in Fig. 13 and Fig. 14, the dynamic selection method shows the best performance of eMBB transmission in all scenarios. For eMBB transmission performance, it can be observed that the dynamic selection mechanism has a maximum gain of 23.26% compared with BCI and a maximum gain of 4.64% compared with BPC. In Table 5, the percentage of URLLC UE satisfying the requirements of the dynamic selection mechanism is increased by 3.75% and 10.54% compared with BCI and BPC, respectively. It is mainly due to the two methods that can complement each other, which means another method can be used when one method is not supported.



▲ Figure 13. UPT of evolved mobile broadband (eMBB) transmission for BCI and the dynamic selection mechanism



▲ Figure 14. UPT of evolved mobile broadband (eMBB) transmission for BPC and the dynamic selection mechanism

▼ Table 5. Percentage of UE satisfying reliability and latency requirements for URLLC transmission in different baseline methods and the dynamic selection mechanism

Combination Case	$\Omega = (5,10)$	$\Omega = (10,10)$	$\Omega = (20,10)$
BCI/%	93.33	89.87	80.64
BPC/%	87.78	83.97	73.84
DPCI&ROPC/%	96.14	92.99	84.38

BCI: baseline cancellation indication
BPC: baseline power control
DPCI: dynamic pattern cancellation indication
ROPC: resource occupancy based power control
UE: user equipment
URLLC: ultra-reliable and low latency communications

5 Conclusions

To solve the coexistence problem of eMBB and URLLC UE in one service cell, the service multiplexing system model is provided. Based on the model, DPCI with a 2D bitmap resource indication and ROPC with dynamically indicating multiple levels of power control parameters are proposed for making up the shortcomings of the existing multiplexing methods. In addition, a dynamic selection mechanism based on DPCI and ROPC is proposed to accommodate the varying cases in different scenarios. Extensive system level simula-

tions and analyses are conducted, results of which show that about 10.54% more URLLC UE satisfies the requirements, and the user perceived throughput of eMBB UE is increased by 23.26%.

References

- [1] ITU-R. IMT Vision: Framework and overall objectives of the future development of IMT for 2020 and beyond: ITU-R M.2083-0 [R]. 2015.
- [2] QI R Z, CHI X F, ZHAO L L, et al. Martingales-based ALOHA-type grant-free access algorithms for multi-channel networks with mMTC/URLLC terminals Co-existence [J]. IEEE access, 2020, 8: 37608 - 37620. DOI: 10.1109/ACCESS.2020.2975545
- [3] ALSENWI M, TRAN N H, BENNIS M, et al. eMBB-URLLC resource slicing: A risk-sensitive approach [J]. IEEE communications letters, 2019, 23(4): 740 - 743. DOI: 10.1109/LCOMM.2019.2900044
- [4] GIDLUND M, LENNVALL T, ÅKERBERG J. Will 5G become yet another wireless technology for industrial automation? [C]//2017 IEEE International Conference on Industrial Technology (ICIT). Toronto, Canada: IEEE, 2017: 1319 - 1324. DOI:10.1109/ICIT.2017.7915554
- [5] ITU-R. Minimum requirements related to technical performance for IMT- 2020 radio interface(s): ITU-R M.2410-0 [R]. 2017
- [6] 3GPP. Study on new radio (NR) access technology physical layer aspects: TR 38.802 [S]. 2017
- [7] POPOVSKI P, NIELSEN J J, STEFANOVIC C, et al. Wireless access for ultra-reliable low-latency communication: principles and building blocks [J]. IEEE network, 2018, 32(2): 16 - 23. DOI: 10.1109/MNET.2018.1700258
- [8] NIKBAKHT H, WIGGER M, SHITZ S S. Mixed delay constraints in Wyner's soft-handoff network [C]//2018 IEEE International Symposium on Information Theory (ISIT). Vail, USA: IEEE, 2018: 1171 - 1175. DOI: 10.1109/ISIT.2018.8437572
- [9] ANAND A, DE VECIANA G, SHAKKOTTAI S. Joint scheduling of URLLC and eMBB traffic in 5G wireless networks [C]// IEEE Conference on Computer Communications. Honolulu, USA: IEEE, 2018: 1970 - 1978. DOI: 10.1109/INFOCOM.2018.8486430
- [10] KASSAB R, SIMEONE O, POPOVSKI P. Coexistence of URLLC and eMBB services in the C-RAN uplink: an information-theoretic study [C]//2018 IEEE global communications conference (GLOBECOM). Abu Dhabi, United Arab Emirates: IEEE, 2018: 1 - 6. DOI: 10.1109/GLOCOM.2018.8647460
- [11] KHAN H, BUTT M M, SAMARAKOON S, et al. Deep learning assisted CSI estimation for joint URLLC and eMBB resource allocation [C]//2020 IEEE international conference on communications workshops (ICC workshops). Dublin, Ireland: IEEE, 2020: 1 - 6. DOI:10.1109/ICCWorkshops49005.2020.9145297
- [12] YANG W, LI C P, FAKOORIAN A, et al. Dynamic URLLC and eMBB multiplexing design in 5G new radio[C]//2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC). Las Vegas, USA: IEEE, 2020: 1-5. DOI: 10.1109/CCNC46108.2020.9045687
- [13] MA T T, ZHANG Y, WANG F G, et al. Slicing resource allocation for eMBB and URLLC in 5G RAN [J]. Wireless communications and mobile computing, 2020, 2020: 1 - 11. DOI: 10.1155/2020/6290375
- [14] ESSWIE A A, PEDERSEN K I. Null space based preemptive scheduling for joint URLLC and eMBB traffic in 5G networks [C]//2018 IEEE Globecom Workshops. Abu Dhabi, United Arab Emirates: IEEE, 2018: 1 - 6. DOI: 10.1109/GLOCOMW.2018.8644351
- [15] PEDERSEN K I, BERARDINELLI G, FREDRIKSEN F, et al. A flexible 5G frame structure design for frequency-division duplex cases [J]. IEEE communications magazine, 2016, 54(3): 53 - 59. DOI: 10.1109/MCOM.2016.7432148
- [16] TAVARES F M L, BERARDINELLI G, MAHMOOD N H, et al. On the impact of receiver imperfections on the MMSE-IRC receiver performance in 5G networks [C]//2014 IEEE 79th Vehicular Technology Conference (VTC Spring). Seoul, Korea (South): IEEE, 2014: 1 - 6. DOI: 10.1109/VTCSpring.2014.7023014
- [17] ESSWIE A A, PEDERSEN K I. Capacity optimization of spatial preemptive scheduling for joint URLLC-eMBB traffic in 5G new radio [C]//2018 IEEE Globecom Workshops. Abu Dhabi, United Arab Emirates. IEEE, 2018: 1 - 6. DOI: 10.1109/GLOCOMW.2018.8644070

Biographies

XIAO Kai (xiao.kai@zte.com.cn) received the master degree from Xidian University, China in 2015 before joining ZTE Corporation. He is now responsible for research and standardization of latest wireless technologies as a standard pre-research engineer at ZTE Corporation. His research interests include initial access of wireless channels, multi-service resource multiplexing, dynamic spectrum sharing, and high-frequency wireless communication.

LIU Xing received the B.S. and M.S. degrees from Harbin Engineering University (HEU), China in 2007 and 2010, respectively. He has been working in ZTE Corporation as a pre-search engineer since graduation. His research interests include URLLC, multicast and broadcast services and cognitive radio.

HAN Xianghui received his M.S. degree in communication and information system in Beijing University of Posts and Telecommunications, China in 2015. Currently he is a senior researcher in ZTE Corporation. His research interests include interference coexistence, shortened TTI and ultra-reliable and low latency communications.

HAO Peng received his M.S. degree in communication engineering from Beijing University of Posts and Telecommunications, China. He joined ZTE Corporation in 2006 and worked on system and link level simulation of 4G system. He has also been involved in the research of key technologies of physical layer of 4G LTE and 5G NR system.

ZHANG Junfeng received the M.S. degree in communication and information technology from Tianjin University, China in 1999. After graduation, he joined ZTE Corporation as a senior engineer of the standardization department. His main research interests are initial access, reference signal, frequency hopping, Coordinated Multi-Point, URLLC, etc. He got two Golden Awards on the outstanding inventions from WIPO-SIPO.

ZHOU Dong received the M.S. degree in computer software and theory from Xi'an Jiaotong University, China in 2009 before joining ZTE Corporation. He is now the director of spectrum policy and regulatory affairs at ZTE Corporation. His research interests include radio regulations, radio communication technical policy, sharing and compatibility studies, cognitive radio and reconfigurable radio systems, high altitude platform systems, and satellite communication.

WEI Xingguang received the B.S. and M.S. degrees from Beijing University of Posts and Telecommunications, China in 2015 and 2018, respectively. He has been working in ZTE Corporation as a pre-search engineer since graduation. His research interests include URLLC, non-orthogonal multiple access and software defined radio.

ZTE Communications Guidelines for Authors

Remit of Journal

ZTE Communications publishes original theoretical papers, research findings, and surveys on a broad range of communications topics, including communications and information system design, optical fiber and electro-optical engineering, microwave technology, radio wave propagation, antenna engineering, electromagnetics, signal and image processing, and power engineering. The journal is designed to be an integrated forum for university academics and industry researchers from around the world.

Manuscript Preparation

Manuscripts must be typed in English and submitted electronically in MS Word (or compatible) format. The word length is approximately 3 000 to 8 000, and no more than 8 figures or tables should be included. Authors are requested to submit mathematical material and graphics in an editable format.

Abstract and Keywords

Each manuscript must include an abstract of approximately 150 words written as a single paragraph. The abstract should not include mathematics or references and should not be repeated verbatim in the introduction. The abstract should be a self-contained overview of the aims, methods, experimental results, and significance of research outlined in the paper. Five carefully chosen keywords must be provided with the abstract.

References

Manuscripts must be referenced at a level that conforms to international academic standards. All references must be numbered sequentially in-text and listed in corresponding order at the end of the paper. References that are not cited in-text should not be included in the reference list. References must be complete and formatted according to *ZTE Communications Editorial Style*. A minimum of 10 references should be provided. Footnotes should be avoided or kept to a minimum.

Copyright and Declaration

Authors are responsible for obtaining permission to reproduce any material for which they do not hold copyright. Permission to reproduce any part of this publication for commercial use must be obtained in advance from the editorial office of *ZTE Communications*. Authors agree that a) the manuscript is a product of research conducted by themselves and the stated co-authors; b) the manuscript has not been published elsewhere in its submitted form; c) the manuscript is not currently being considered for publication elsewhere. If the paper is an adaptation of a speech or presentation, acknowledgement of this is required within the paper. The number of co-authors should not exceed five.

Content and Structure

ZTE Communications seeks to publish original content that may build on existing literature in any field of communications. Authors should not dedicate a disproportionate amount of a paper to fundamental background, historical overviews, or chronologies that may be sufficiently dealt with by references. Authors are also requested to avoid the overuse of bullet points when structuring papers. The conclusion should include a commentary on the significance/future implications of the research as well as an overview of the material presented.

Peer Review and Editing

All manuscripts will be subject to a two-stage anonymous peer review as well as copyediting, and formatting. Authors may be asked to revise parts of a manuscript prior to publication.

Biographical Information

All authors are requested to provide a brief biography (approx. 100 words) that includes email address, educational background, career experience, research interests, awards, and publications.

Acknowledgements and Funding

A manuscript based on funded research must clearly state the program name, funding body, and grant number. Individuals who contributed to the manuscript should be acknowledged in a brief statement.

Address for Submission

<http://mc03.manuscriptcentral.com/ztecom>

ZTE COMMUNICATIONS

中兴通讯技术(英文版)

ZTE Communications has been indexed in the following databases:

- Abstract Journal
- Cambridge Scientific Abstracts (CSA)
- China Science and Technology Journal Database
- Chinese Journal Fulltext Databases
- Index of Copernicus
- Inspec
- Ulrich's Periodicals Directory
- Wanfang Data

ZTE COMMUNICATIONS

Vol. 19 No. 2 (Issue 74)

Quarterly

First English Issue Published in 2003

Supervised by:

Anhui Publishing Group

Sponsored by:

Time Publishing and Media Co., Ltd.

Shenzhen Guangyu Aerospace Industry Co., Ltd.

Published by:

Anhui Science & Technology Publishing House

Edited and Circulated (Home and Abroad) by:

Magazine House of ZTE Communications

Staff Members:

General Editor: WANG Xiyu

Editor-in-Chief: JIANG Xianjun

Executive Editor-in-Chief: HUANG Xinming

Editor-in-Charge: ZHU Li

Editors: REN Xixi, LU Dan, XU Ye, YANG Guangxi

Producer: XU Ying

Circulation Executive: WANG Pingping

Liaison Executive: LU Dan

Assistant: WANG Kun

Editorial Correspondence:

Add: 12F Kaixuan Building, 329 Jinzhai Road,
Hefei 230061, P. R. China

Tel: +86-551-65533356

Email: magazine@zte.com.cn

Website: http://journal13.magtechjournal.com/Jwk3_zte

Annual Subscription: RMB 80

Printed by:

Hefei Tiancai Color Printing Company

Publication Date: June 25, 2021

China Standard Serial Number: ISSN 1673-5188
CN 34-1294/TN