Wireless Communications and Mobile Computing

Recent Advances in Wireless Communication Protocols for Internet of Things

Lead Guest Editor: Jiangchuan Liu
Guest Editors: Feng Wang, Xiaoqiang Ma, and Zhe Yang

Recent Advances in Wireless Communication Protocols for Internet of Things

Wireless Communications and Mobile Computing

Recent Advances in Wireless Communication Protocols for Internet of Things

Lead Guest Editor: Jiangchuan Liu
Guest Editors: Feng Wang, Xiaoqiang Ma, and Zhe Yang

Editorial Board

Contents

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 8791485, 2 pages
https://doi.org/10.1155/2017/8791485

Editorial

Recent Advances in Wireless Communication Protocols for Internet of Things

Jiangchuan Liu,[1] Feng Wang,[2] Xiaoqiang Ma,[3] and Zhe Yang[4]

[1]*College of Natural Resources and Environment, South China Agricultural University, Guangzhou, Guangdong 510640, China*
[2]*Department of Computer and Information Science, University of Mississippi, University, MS 38677-1848, USA*
[3]*School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China*
[4]*Northwestern Polytechnical University, Xi'an, Shaanxi 710065, China*

Correspondence should be addressed to Jiangchuan Liu; jcliu@cs.sfu.ca

Received 18 October 2017; Accepted 18 October 2017; Published 16 November 2017

Internet of Things (IoT) is one of the hottest research fields nowadays and has attracted huge interests and research efforts from both academia and industry. IoT can connect a large number of sensors, actuators, devices, vehicles, buildings, and/or other objects to form a network where data can be collected from the physical world, exchanged and processed in the cyber world, and then fed back into the physical world through actuations. This makes IoT one of the key foundations towards the vision of smart cities, with many promising applications such as environmental monitoring, infrastructure management, manufacturing, energy management, medical and healthcare, building and home automation, and transportation.

Recently, the advances in various wireless communication protocols in technologies such as 5G, RFID, Wi-Fi-Direct, Li-Fi, LTE, and 6LoWPAN have greatly boosted the potential capabilities of IoT and made it become more prevalent than ever, which also accelerate the further integration of IoT with emerging technologies in other areas such as sensing, wireless recharging, data exchanging, and processing. Yet, how these technologies especially the corresponding wireless communication protocols can be well aligned with IoT to maximize their benefits on such performance as scalability, service quality, energy efficiency, and cost effectiveness is still open to investigation and thus calls for novel solutions. And the involved privacy and security issues also need to be carefully examined and addressed.

This special issue aims to summarize the latest development in wireless communication protocols for Internet of Things and how such development can enable other emerging technologies to be further integrated with IoT and boost its capabilities. We classify the accepted papers into four major focuses: (1) energy efficiency; (2) QoS awareness; (3) localization; and (4) security and privacy.

The energy efficiency focus consists of 2 papers.

The paper "Energy Harvesting for Internet of Things with Heterogeneous Users" studies the energy harvesting problem in the Internet of Things with heterogeneous users under multiuser MISO broadcast channels.

The paper "Profiling Energy Efficiency and Data Communications for Mobile Internet of Things" aims at monitoring the power consumption behaviors of the smartphones, profiling both individual applications and the entire system, to make better decisions in power management. The authors design a cloud orchestration architecture as an epic predictor of behaviors of smart devices by extracting their application characteristics and resource utilization.

The QoS awareness focus consists of 4 papers.

The paper "QoE-Driven D2D Media Services Distribution Scheme in Cellular Networks" proposes a novel media service scheme based on different QoE models that jointly solve the massive media content dissemination issue for cellular networks, where the authors investigate the so-called Media Service Adaptive Update Scheme (MSAUS) framework to maximize users' QoE satisfaction and derive the popularity and priority function of different media service QoE expression. A Media Service Resource Allocation (MSRA) algorithm is also designed to schedule limited

cellular networks resource, which is based on the popularity function to optimize the total users' QoE satisfaction and avoid D2D interference.

The paper "Det-WiFi: A Multihop TDMA MAC Implementation for Industrial Deterministic Applications Based on Commodity 802.11 Hardware" proposes Det-WiFi, a real-time TDMA MAC implementation for high-speed multihop industrial application, which can support high-speed applications and provide deterministic network performance via combining the advantages of high-speed IEEE802.11 physical layer and a software Time Division Multiple Access (TDMA) based MAC layer.

As network coding is emerging as a promising technique that can provide significant improvements in the throughput of Internet of Things, the paper "Efficient Network Coding with Interference-Awareness and Neighbor States Updating in Wireless Networks" proposes a novel network coding scheme to achieve a higher throughput improvement with lower computational complexity and buffer occupancy than traditional greedy-based schemes, where topology knowledge is utilized to minimize interference and obtain more throughput, and a more reliable broadcast protocol is exploited to alleviate the decoding failure caused by the inherent error ratio in ETX.

The paper "Exploiting Delay Aware Load Balance for Scalable 802.11 PSM in Crowd Event Environments" presents ScaPSM (Scalable Power-Saving Mode scheduler), a design that enables scalable competing background traffic scheduling in crowd event 802.11 deployments with Power-Saving Mode (PSM) radio operation. The key novelty behind ScaPSM is that it exploits delay aware load balance to control judiciously the qualification and the number of competing PSM clients before every beacon frame's transmission, which helps to mitigate congestion at the peak period with increasing the number of PSM clients.

The localization focus consists of 3 papers.

As the Receiving Signal Strength (RSS) is the key foundation for IoT communication resource allocation, localization, interference management, sensing, and so forth, the paper "Compressed RSS Measurement for Communication and Sensing in the Internet of Things" proposes a compressive sensing-based RSS measurement solution, which takes advantage of compressive sensing theory to enable simultaneous measurement in the same channel, so as to achieve conflict-tolerant, time-efficient, and accuracy-guaranteed without any model-calibrate operation.

The paper "Dealing with Insufficient Location Fingerprints in Wi-Fi Based Indoor Location Fingerprinting" targets one but common situation when the collected measurements on received signal strength information are insufficient and shows limitations of existing location fingerprinting methods in dealing with inadequate location fingerprints. The authors then introduce a novel method to reduce noise in measuring the received signal strength based on the maximum likelihood estimation and compute locations from inadequate location fingerprints by using the stochastic gradient descent algorithm.

The paper "SmartFix: Indoor Locating Optimization Algorithm for Energy-Constrained Wearable Devices" presents SmartFix, an optimization algorithm for indoor locating based on Wi-Fi RSS, which utilizes user motion features and extracts characteristic value from history trajectory and corrects deviation caused by unstable Wi-Fi signals.

The security and privacy focus consists of 2 papers.

Given that covert communication is applied in many IoT scenarios for information transmission security, the paper "Collaborative Covert Communication Design Based on Lattice Reduction Aided Multiple User Detection Method" applies the lattice reduction theory to multiple access interference cancellation of spread spectrum communication and proposes a novel lattice reduction aided multiple user detection method to better design a covert communication system.

The paper "Privacy-Preserving Meter Report Protocol of Isolated Smart Grid Devices" proposes an efficient privacy-preserving meter report protocol for the isolated smart grid devices, which consists of an encryption scheme with additively homomorphic property and a linearly homomorphic signature scheme suitable for privacy-preserving data aggregation.

Acknowledgments

We would like to thank all the reviewers who have participated in reviewing the articles submitted to this special issue.

Jiangchuan Liu
Feng Wang
Xiaoqiang Ma
Zhe Yang

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 6562915, 15 pages
https://doi.org/10.1155/2017/6562915

Research Article

Profiling Energy Efficiency and Data Communications for Mobile Internet of Things

Peramanathan Sathyamoorthy,[1] Edith C.-H. Ngai,[1] Xiping Hu,[2,3] and Victor C. M. Leung[4]

[1] *Department of Information Technology, Uppsala University, Uppsala, Sweden*
[2] *Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China*
[3] *The Chinese University of Hong Kong, Shatin, Hong Kong*
[4] *Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada*

Correspondence should be addressed to Xiping Hu; xp.hu@siat.ac.cn

Received 15 April 2017; Accepted 2 October 2017; Published 7 November 2017

Academic Editor: Xiaoqiang Ma

This paper proposes a novel power management solution for resource-constrained devices in the context of Internet of Things (IoT). We focus on smartphones in the IoT, as they are getting increasingly popular and equipped with strong sensing capabilities. Smartphones have complex and asynchronous power consumption incurred by heterogeneous components including their on-board sensors. Their interaction with the cloud allows them to offload computation tasks and access remote data storage. In this work, we aim at monitoring the power consumption behaviours of the smartphones, profiling both individual applications and the system as a whole, to make better decisions in power management. We design a cloud orchestration architecture as an epic predictor of behaviours of smart devices by extracting their application characteristics and resource utilization. We design and implement this architecture to perform energy profiling and data analysis on massive data logs. This cloud orchestration architecture coordinates a number of cloud-based services and supports dynamic workflows between service components, which can reduce energy consumption in the energy profiling process itself. Experimental results showed that small portion of applications dominate the energy consumption of smartphones. Heuristic profiling can effectively reduce energy consumption in data logging and communications without scarifying the accuracy of power monitoring.

1. Introduction

IoT is a convergence of number of technologies such as sensors, IPv6, wireless communication, and the Internet. Any real-world objects become smart just by satisfying a few conditions not limited to (1) being uniquely identifiable, (2) being able to sense or actuate, and (3) being able to communicate [1]. The growth of smart objects is posing challenges to the research community in energy management, security [2], and data analytics. Among these challenges, security and privacy issues affect not only the technical system design level, but also the ethical, behavioural, and policy levels. In terms of data analytics, we have powerful analytical tools available with advanced data analysis algorithms [3]. On the other hand, energy management is more complex and chaotic considering different applications and usage patterns, which is our focus in this paper.

Berkeley National Laboratory defined energy efficiency as using less energy to provide the same service [4]. The need for energy efficiency is highly inevitable in almost every type of organisations and industries in different sectors including Information and Communications Technology (ICT). Energy management in Internet of Things (IoT) aims at reducing the electricity, which is beneficial for many industries to reduce their electricity bills. As the smart objects become smaller in size, their small sized batteries provide limited power for operations. Even though the smart appliances are idle, they could indirectly waste huge amount of energy in the long term and eventually increase the electricity bills too. Although ICT can enable energy efficiency across

all sectors, at present there is little market incentive to ensure that network-enabled devices themselves are energy-efficient. In fact, up to 80% of their electricity consumption is used just to maintain a network connection. Even though the amount of electricity used by each device is small, the anticipated massive deployment and widespread uses make the cumulative consumption considerable, as reported by International Energy Agency in [5].

Hereafter we narrow our focus on smartphones, which are smart devices that increase in exponential order over the last ten years. Modern smartphones provide heterogeneous functionalities including a number of sensors. They are one of the most representative and popular smart objects in the IoT. Nevertheless, smartphones are resource constraint with respect to battery, memory, and computation. It is common for them to offload computation and access remote data storage on the cloud servers via network. Cloud computing in the IoT leads to thousands of cloud supported applications and is growing steeply. As a consequence, smartphones are consuming a lot of energy for communication with the cloud. Due to the size limitation, effort of making powerful batteries is not able to satisfy the increasing energy demand in the smartphones. It is important to reduce energy consumption when developing new kind of applications.

In the last decade, the rising trend in the popularity of smartphones motivated software developers to increase application functionality. Based on previous study [6], most of the power of the smartphones is consumed by wireless communications and display (e.g., backlight for screen). In addition, increasing application functionality demands extra power budget that as a result decreases smartphone battery lifetime [7]. Smartphones are usually running multiple applications with different operations at a time. It is very difficult to understand and identify the cause of high energy consumption in this asynchronous power consuming environment. It is necessary to provide profiling of power consumption from different levels, including system level as a whole, individual applications, and system calls in operation level.

In this paper, we propose the first iterative and novel solution using *cloud orchestration* for power management on smartphones. Cloud orchestration aggregates power profiling data from the smartphones and coordinates data storage, data analysis, learning, and decision-making. By profiling data, the orchestrator learns the power consumption behaviours and the usage pattern of the participating smartphones. It can answer questions like the following:

(i) Which applications are most energy consuming on the smartphones?

(ii) What are the characteristics in applications that consume most of the energy?

These findings can be used to further optimize the energy monitoring framework. For example, the energy profiler can predict and collect only the most important power consumption data logs on the smartphones. Our cloud orchestrator framework supports dynamic workflow of processes and adaptive services fitting the needs of different users. It aims for providing overall system power management rather than making part of the system efficient. The cloud orchestration services can be selected and configured dynamically depending on the application characteristics, usage pattern, time, and location context. Both offline and real-time services can be supported to provide long-term and large data analytic, or to give real-time alert on unusual events. Profiling energy consumption creates an opportunity to perform better energy management and increase battery lifetime. It helps to understand energy consumption behaviours of a wide range of applications for optimal battery resource use.

2. Related Works

A lot of efforts have been made to enable energy efficiency in smartphones and IoT in general. There are a range of solutions tried out in the *hardware architecture* level [8, 9], *data communication* level [10, 11], *network infrastructure* level [12], and *protocols* optimization [13]. Different tools have been developed to measure energy consumption on smart devices and smartphones. For example, power monitor meter has been used to provide the current with constant voltage 3.7 V to the smartphone instead of using the battery [14]. This hardware setup can provide accurate power consumption measurements, but it is a bulky solution not suitable for ordinary users in their daily usage.

As Intel summed up in [15], *Software Energy Efficiency* has the significance towards achieving *computational efficiency*, *data efficiency*, *context awareness*, and *idle efficiency* in broader sense. Nevertheless, current solutions which try to characterize power consumption on the smartphones usually focus on specific operations, such as communications [10] or interactions with certain hardware components, such as LCD or GPS. There are several common problems in most of the existing solutions, including the following: (1) system as a whole was not considered; (2) trade-off between components was not properly considered; (3) interdependence of the components was not properly studied; (4) the existing solutions are suboptimal. In order to address the above problems, we need a comprehensive approach to understand the energy consumption of individual applications as well as their interdependency and significance in the whole system. A comprehensive analysis can help the users to identify the most power consuming applications or operations on their smartphones. This can also make the power monitoring process more adaptive to the user behaviour and more energy-efficient in a long run.

Measuring power consumption of resources such as CPU, memory, display, and communication is extremely useful in finding the energy-hungry applications, background services, and processes. In [16], the energy costs for task offloading over IEEE 802.11 WiFi and 3G have been modelled mathematically. For WiFi network, it uses protocol parameters such as data rate, base rate, and contention window size to derive the formula. For 3G/4G, Radio Resource Controller (RRC) states are considered and the total energy consumption is calculated by three parts, including promotion signalling, data transfer, and tail energy.

Many existing solutions for power monitoring are running on the smartphones nowadays [17, 18]. Software energy profilers are common tools to measure the energy consumption of mobile devices, applications running on those devices, and various hardware components. They adopt different modelling and measurement techniques [19]. These energy profilers can monitor the percentage of battery consumed by different applications. The advantage of these solutions is simple to use, but the limited memory and computation capability of smart devices make it hard to support more advanced data analysis. It is then difficult to support large-scale and long-term analysis of energy consumption data for both personalized or crowd-based monitoring. Regarding measuring energy consumption, solid background has been provided in [20]. Internet-of-Things Architecture is a consortium rigorously developing architectural reference models. These models serve as initial guidance potentially towards concrete architecture for the problem of interest and eventually towards the actual system architecture [21]. In [22], devices orchestration is explained from the business process point of view.

With respect to mobile-cloud paradigm, AppATP [23] leverages cloud computing to manage data transmissions for mobile apps, transferring data to and from mobile devices in an energy-efficient manner. Carat [24] presents a crowdsourcing approach for collection energy consumption data on smartphones and diagnosing energy anomalies from a community of clients. We share a similar concept of running data analysis in the cloud and further explore the opportunity of cloud orchestration services for smartphones. Instead of taking a black-box process-based approach, we propose a cloud orchestration approach for energy efficiency of smart devices. Cloud orchestration has the capabilities of coordinating different cloud services, such as data storage, analysis, and processing, in a comprehensive framework. Appropriate services can be selected according to the need of individual users. Heuristic profiling can be implemented to reduce the amount of log data and communication overheads. This approach is useful in reducing the energy consumption in the energy profiling process itself. Our framework can be extended easily to include new data mining techniques and new services contributed by other users. It supports both individual profiling for personalized services and community analysis using crowdsourced data.

3. Cloud Orchestration for Energy Efficiency

IoT initially has two visions: one is the *Things oriented* vision and the other is the *Internet oriented* vision. The Things vision emphasizes the sensing and communication capability of different types of smart devices, which can be standalone or embedded into different real-world objects. The Internet vision focuses on the connectivity of the smart devices and their interaction with the Internet. Connecting smart devices to the Internet enables large data storage and analysis that are not feasible on the resource-limited devices. The Internet vision has driven cloud computing for the IoT to provide advanced data processing and data management capabilities.

Later, when new challenges were introduced such as unique addressing and storing information, *semantic oriented* vision had arisen [25]. According to this new vision, the participating devices are categorized and the orchestration is configured to support scalable and controlled integrated solutions. In this work, we develop the idea of cloud orchestration to provide energy efficiency services for the IoT devices. A cloud orchestrator is a software system that manages the interconnections and interactions of different cloud-based services and processes. It supports dynamic workflows to connect various automated processes and associated resources according to the needs of users and the context environment [26].

3.1. Data Communications. Data communication in mobile IoT often occurs between smartphones (clients or peers) and the cloud (server) via interconnected telecommunication medium such as the Internet. The important components for data communication are data, client, server, and the infrastructure of network and protocols.

Our focus on energy profiling is more on the client side taking into account its applications and communication interfaces. Profiling energy consumption in data communications is complex. We should consider the protocols at various levels in the OSI (Open System Interconnectivity), especially the Internet suite of the TCP/IP protocols. The energy consumption of communication interfaces, such as WiFi, 3G, 4G LTE, Bluetooth, Near Field Communication (NFC), and GPS, could be measured via the APIs given by the smartphone operating system, such as Android. It is very important to capture the events related to communication interfaces for energy profiling.

3.2. Cloud Orchestration Design Goals. The main goal of our system design is to provide energy-efficient decision(s) back to the service enabled smartphones which are participating in the orchestration. *Orchestration*, the concept existing in the music world, was adopted in process automation of business world by automating, coordinating, and managing complex systems, middleware, and services. Energy profiling may impose vulnerability in energy efficiency. Let us give an illustrative example. Even a single and careless piece of code `(while(battery.percentage) println(battery .percentage))` may cause the system to run into an infinite loop and drain all the battery. This small mistake can make any effort of power saving become in vain. Hence, there is a need for intelligent system, which is capable of coordinating different components and finding and categorizing the energy errors. The system should have access to powerful *dynamic control system engine* for fixing such errors. To assist bug fixing, we may need insights from the big data of crowdsourced logs/operations over long period of time. Orchestration has the capabilities of integrating different types of clouds, processes, and services for power management, which is an ideal solution.

3.3. EEaaS Orchestration Architecture. We propose a cloud orchestration architecture, called EEaaS, for power management of IoT devices in Figure 1. The design is open and

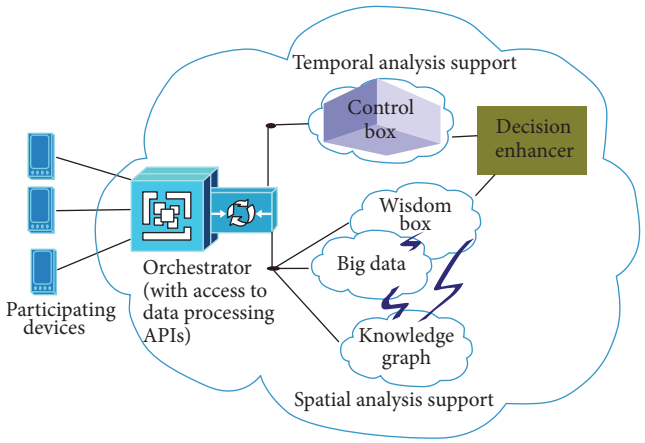


Figure 1: EEaaS cloud orchestration architecture for power management.

flexible, which makes it easy to add, remove, and merge with new models and services at any granular level in the orchestrator. The orchestrator coordinates the following components, including *participating devices*, *data processor*, *big data storage*, *knowledge graph*, *wisdom box*, *control box*, and *decision enhancer*.

3.3.1. Participating Devices and Smart Profiler. These are the smart devices in the IoT that are of interest in minimizing their energy consumption. Upon registration with the orchestrator, a fully customizable and lower energy consuming background *service application* is enabled in the smart devices. This service application sends low-level system-call logs periodically and reports abnormal system behaviours spontaneously. These abnormal events may include accidental system crash or unusual battery drain by specific application. To avoid security and privacy issues, logs are collected anonymously with unique device profile. Not only is this application a log collector, but also it acts as a local *self-controller* attempting to catch energy errors in time and optimize energy profiling in the long run. Its functionalities are regularly updated by the orchestrator. The participating devices report unusual events to the orchestrator in real time. Since the unusual events contain only small amount of data, the communication overhead is not so much. On the other hand, the large volume of logged data on resource utilization is reported only when the smartphones are connected to the computer or WiFi network. This is to avoid the continuous data communication through mobile cellular networks. Data filtering and heuristic profiling can be performed to reduce the amount of data samples in order to save energy.

3.3.2. Data Processor. Data processor is a collection of APIs for various data processing methods accessible to the orchestrator. According to the context and the need of user, appropriate data processing methods will be chosen by the orchestrator. The data processor supports both big data analysis and temporal data analysis. Advanced data mining techniques can be implemented in the data processor to perform data filtering and data aggregation. For example, it can characterize the energy consumption behaviour of different applications on the smartphones and identify the most power-hungry applications.

3.3.3. Big Data Storage and Modern Tools. The data produced by the smartphones would be in massive scale over time. In order to handle these data-intensive operations, we need big data storage and modern cloud programming paradigms such as *Hadoop* and *Apache flink*. For deep analysis of sample data, powerful computing languages such as *Python* and *R* are required.

3.3.4. Knowledge Graph. When interpreting large volumes of data logs, dynamic knowledge graph is built and keeps on updating. Knowledge graph is a knowledge base originally used by Google to enhance its search engine's search results with semantic-search information gathered from a wide variety of sources. Nodes are qualified classes and subclasses with attribute-value pairs, so that they provide a clear and structured view of data. Using the knowledge graph, it is then easy to get specific resource utilization and energy consumption data for analysis with respect to location, device model, Internet service provider, and various specifications.

3.3.5. Wisdom Box. Wisdom box contains a set of learning algorithms with primary focus on building location specific context information (in the spatial domain). It can capture the location of the device and correlate the location with various usage and communication patterns. The wisdom box acts as a predictor of trends in data, usage patterns, and system behaviour anomalies. It uses combination of statistical algorithms and machine learning algorithms to make energy-efficient decisions. The decisions that are independent of device, platform, and applications are stored in the *decision enhancer* in the orchestration. Device, platform, and application specific decisions are fused with the knowledge graph and the reference graph in the orchestration.

3.3.6. Control Box. Control box is a builder of real-time and dynamic self-controller for the participating devices. It can also raise alerts and give time-sensitive feedback (temporal domain) to the users. The self-controller is implemented as a service. To make the self-controller even more intelligent, context related decisions in the spatial domain are used. Feedback is received from the participating devices to evaluate the performance of data log collection.

3.4. Energy-Efficient Data Profiling and Communication. Given that data communication is one of the greatest challenges for energy efficiency, cloud-assisted data profiling would seem counterintuitive. Our system design takes smart logging with minimal data communications to reduce the overhead. It provides an alternative solution to fine-grained and real-time profiling, which is both data- and computation-intensive. We study existing profiling techniques and identify the most important energy features to design a heuristic profiler. This heuristic profiler collects data intelligently and maintains minimal communications with the cloud. Our

Table 1: Indicators of energy consumption in EEaaS.

Indicator types	Definitions	Example resources
Key energy indicators (KEIs)	Primary and key resources in system	CPU and memory
Secondary KEI (SKEI)	Not frequently used resources but important	GPS and Bluetooth
Relative KEI (RKEI)	Active only when certain applications are running	Camera and activity sensors
Potential cause indicator (PCI)	Learned by the orchestrator over time	

approach can effectively reduce the communication overhead and energy consumption in the energy profiling process itself.

The orchestrator is designed to able to learn the trends and patterns of the sample data. After learning the characteristic of different applications, the sampling rates in the profiler can be adjusted adaptively to reduce the data samples of less active and less energy consuming applications. Heuristic profiling and data filtering allow more focused sampling on key applications that consume most of the energy, while reducing the total number of data samples for less energy consuming applications.

3.4.1. Key Energy Indicators. Managing and analysing big data is not the major cause of energy consumption. The real bottleneck is collecting data through frequent and intensive communications from the participating devices. Instead of sending all the data logs to the orchestration for analysis, we classify certain system calls and resource utilization as primary key factors, called *key energy indicators* (KEIs), such as CPU and memory usage. We further categorize the KEI into *secondary key energy indicators* (SKEI) and relative key energy indicators (RKEI). SKEI are not so frequently used but are important when they are active, such as GPS and Bluetooth. RKEI are active only when specific applications are running or occur in certain context. Examples of RKEI include camera and activity sensors. Table 1 summarizes the definition of the indicators.

3.4.2. Potential Cause Indicators. KEIs capture the energy consumption behaviours and identify the causes of energy consumption. Then comes the *potential cause indicators* (PCI), which can be learned by the orchestration over time. When the system becomes matured enough, the participatory devices collect and report less amount of data. Crowdsourcing further makes it easy to distribute the workload of data collection. In the long run, less logs and less communications from participatory devices are required. More sophisticated context-aware models in the orchestration are created as a result. The workflows in the orchestration then become more energy-efficient, as it can adapt to the usage pattern and context dynamically (see Figure 2).

In Figure 2, the cloud-based smart profiler (top-left) sends the logged data or control messages from the participating devices to the cloud. The logged data are sent together with the *context header*, which is used by the orchestration to indicate the message type. For example, if the context header is *control*, then the orchestrator knows that there are no data logs in the message but events (issues) being reported from the participating devices. In this case, the events are forwarded to control box in order to address the issues. Otherwise, the other messages (data logs) are forwarded to the big data module supported by the spatial data processing unit for classification, prediction, and context generation. The data processing unit helps the orchestration to learn what the KEIs are, so that less significant data can be identified and removed in the future. When the system becomes mature enough, the participatory devices will collect mainly the KEI, so that they can report less amount of data logs to save energy. Through iterative learning, the orchestration can build more sophisticated context-aware energy models for making better prediction in energy profiling.

3.5. Energy-Efficient Techniques. In addition to energy-efficient profiling, we present additional techniques that can help to reduce energy consumption on the mobile phones. Based on our observations, a lot of mobile applications are running as background applications, such as Facebook, Google Maps, WeChat, and WhatsApp. It is recommended to close the applications when the user is not using them. Running background applications consumes additional energy and many of these applications are unnecessary. In modern Android OS, there is optimization function that could be used to reduce battery consumption. For example, a user may choose to close background applications when the screen is locked to help saving energy. If a user may not want to miss any emails or messages, he or she may choose to close down background applications selectively to fit his or her needs.

Moreover, it is beneficial to turn off mobile data and GPS when they are not needed. This is because data communication is one of the major sources of battery consumption. In particular, GPS and mobile data consume much more energy than WiFi and Bluetooth. Note that Android keeps location based applications, such as Google Maps, running in the background, which constantly drains the battery. To reduce battery consumption, it is best if the user can turn off unnecessary hardware radios, such as LTE, NFC, GPS, WiFi, and Bluetooth.

Another effective way to reduce energy consumption is to check the battery consumption of the mobile using Android or other monitoring mobile applications (e.g., Carat, PowerTutor, and Trepn). These applications help to identify power-hungry applications and detect bugs that cause unusual high energy consumption on the mobile device. Figure 3(a) shows the power-intensive applications identified by Carat [24], a mobile application to generate personal recommendations for improving battery life. This screenshot shows four applications that are correlated with higher energy

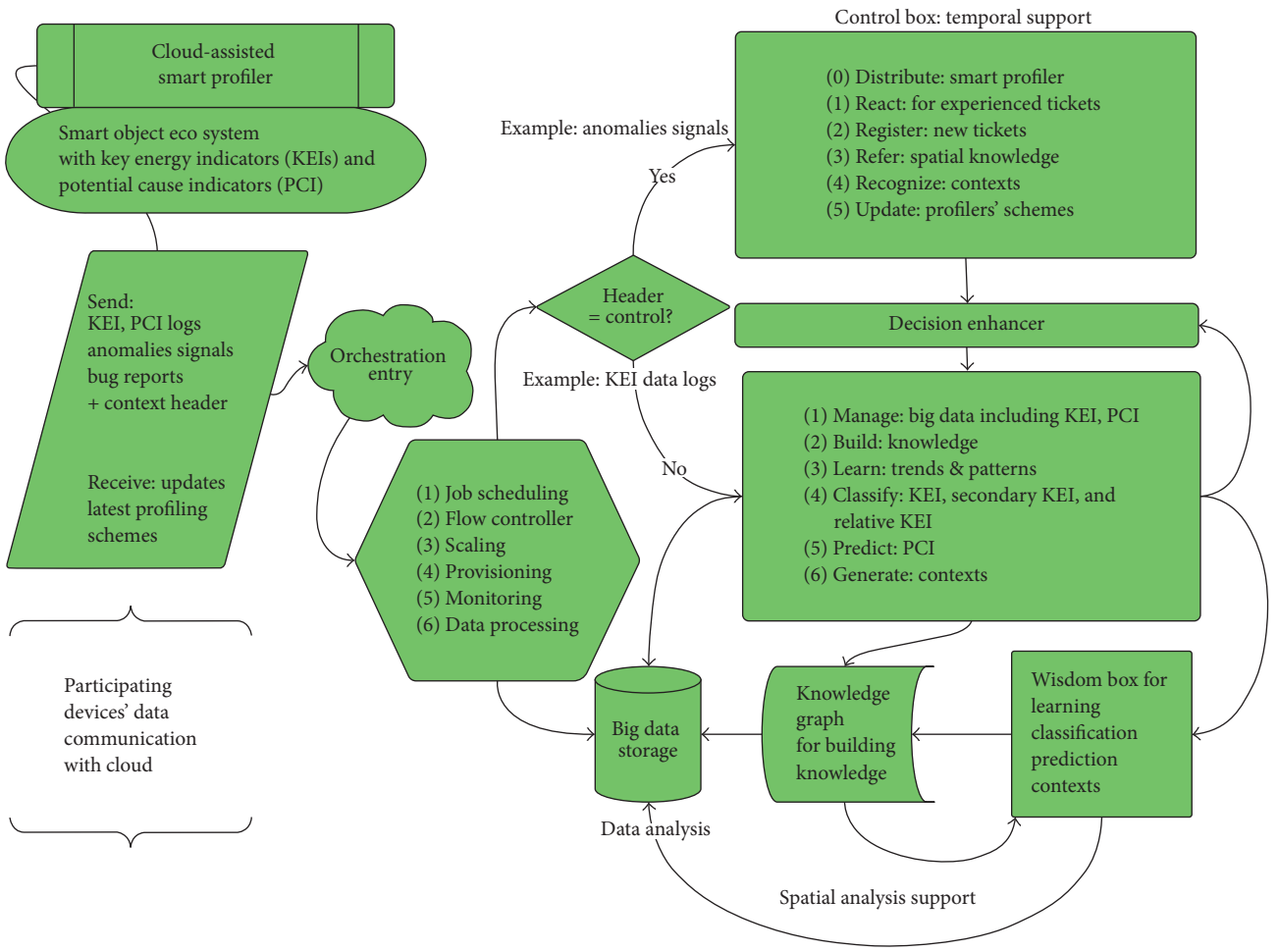


Figure 2: EEaaS orchestration inside the box.

user across many devices. Carat recommends that closing these power-intensive applications may improve battery life.

Similarly, Figure 3(b) shows a list of background applications that are suggested to be closed during lock screen by Android. These power-intensive applications include Geo Tracker, Google Maps, WeChat, and WhatsApp. Users may consider uninstalling power-hungry applications or better control of their operations. For example, it is possible to reduce polling from emails, Facebook, or Twitter to save energy by reducing refresh frequency or enabling manual polling instead of using automatic and constant polling.

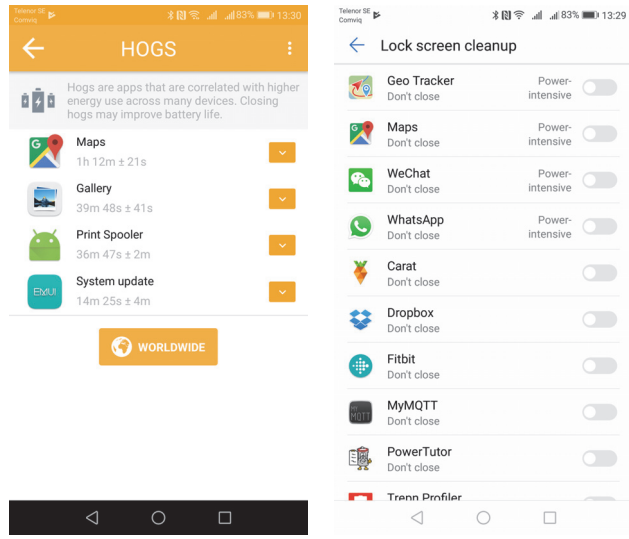
4. Implementation

In this section, we describe the prototype development, the tools used, and the implementation details.

4.1. Data Inspection. We first explored the data logs of two smartphones, Samsung SIII running Android version 4.4 and Nexus 5 running Android version 5.0. They are installed with Android platform tools, such as logcat [27], for logging system debug information, and DDMS (Dalvik Debug Monitor Server) [28] for debugging applications. These tools help us to capture the behaviours of the system and the applications running on the platform.

We use Qualcomm's Trepn Profiler [29] for collecting data in suitable formats, namely, SQLite and CSV. Unfortunately, the eclipse plug-in provided by Trepn suffers from poor visualization. The data in CSV format requires a lot of cleaning to use. It also has great limitations on the number of parameters it can profile. Nevertheless, the data brings a lot of insights. We also develop a web application, called EnergyApp, that helps to study the data in SQLite format. Figure 4 is a screenshot of a list of mobile applications and their statistics in the EnergyApp.

4.2. Key Energy Indicators. We soon realize that crowdsourcing methodology results in a lot of data, so that smart big data management and analysis are needed to reduce communication and computation overheads. One method is to select key profiling features and ignore the less influencing features. We derive a methodology to extract the KEIs to reduce the overheads. It relies on two data mining techniques, decision trees and random forest, to figure out the variables of importance by Gini index. Gini index measures the inequality among values of a frequency distribution. A Gini index of



(a) Power-intensive apps shown by Carat

(b) Power-intensive apps shown by Android

Figure 3: Power-intensive applications identified.

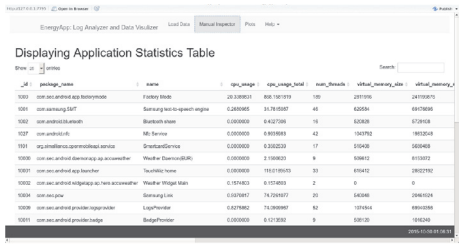


Figure 4: A list showing the app statistics in the EnergyApp.

Figure 5: Screenshot of the energy data administration tool developed in the cloud.

zero expresses perfect equality, while that of one expresses maximal inequality among values. Gini index is a common metric used for determining the splits of a decision tree. This service is running as software as a service application in the cloud. A screenshot of the energy data administration tool is shown in Figure 5. It shows a decision tree on the left, which predicts the energy consumption using a set of resource utilization parameters. The right side of the figure shows the results of random forests on different resources in the system.

4.3. Energy-Efficient Cloud Profiler. We further develop a prototype for cloud profiler with the following key characteristics:

(i) Real-time database with autosynchronization support, such as Firebase, which works reliably with low latency and energy consumption

(ii) Distributed data protocol, which supports publish-subscribe services for mobile and web applications

(iii) Client-centric data response, such as GraphQL, which can reduce the amount of data.

We explain in detail client-centric response because of importance compared to the others. Mobile and web applications are heavily using HTTP RESTful services and ad hoc endpoints to obtain data nowadays. The core problem with this approach is that the response data is entirely decided by the server. The server-side application might be responding to the client application with more data than required. For example, a simple client application showing current weather is getting more information than what the application displays on the screen. Facebook is one of the most energy consuming applications due to user-engaged usage, according to our measurements. GraphQL [30] is designed as remedy to the above-mentioned problem. It is a query data language for the API and a server-side runtime for executing queries. It provides a complete and understandable description of the data in the API, giving clients the power to ask for exactly what they need and nothing more.

5. Experimental Results

Smartphone is a system-on-chip architecture with three key components, including *application processor* to handle user

Table 2: ACPI/OSPM defined power states.

Global system states	Device power states	Processor power states
G0 working	D0 fully on	C0
G1 sleeping	D1	C1
G2/S5 soft off	D2	C2
G3 mechanical off	D3	C3
	D3 off	

applications, *modem processor* to handle transmission and reception, and *peripheral devices (I/O)* to interact with the users. In smartphones, the power consumption of any I/O component is often higher than the power consumption of the CPU or at least comparable. In [31], the drawbacks of power models derived from external power meters and software modelling are well explained. Software power modelling does not address the tail power states, which occur when the components remain powered on and consume energy even though the CPU is idle. This problem can be addressed by system-call tracing to check each component's power state, though it may consume more energy. Nevertheless, energy profiling is a very important first step to characterize the power consumption on smartphones.

The Advanced Configuration and Power Interface (ACPI) specification [32] has been evolving as a common hardware interface in Operating System directed configuration and Power Management (OSPM) for both the end devices and the entire systems. When profiling an individual application or entire platform, it is useful to fetch information about the states of system, device, and processors, so that better decision can be made to achieve energy efficiency. Table 2 lists the key global system states, device power states, and processor power states. For instance, the process power state C0 indicates that the CPU executes instructions, while C1–C3 are processor sleeping states where the CPU consumes less energy than C0.

5.1. Energy Consumption in Hardware and Software. We conduct an experiment to study energy consumption of the hardware and software on a newly installed Android mobile phone (Huawei Honor 8). The main purpose of this experiment is to evaluate the energy consumption of basic operating system and hardware on a mobile phone. Table 3 shows the energy consumption on the mobile during the day. We consider a mobile phone with basic functionality with minimum usage of power-hungry applications (e.g., Facebook, YouTube, or games). We can see that the hardware on the mobile consumes around 34% of the battery and the software consumes the remaining 66% on the mobile phone. The screen takes up most energy consumption in the hardware during daytime. The remaining energy in the hardware is taken by the 3G/4G communication and other short-range communications (e.g., WiFi and Bluetooth). We also observe that the voice call takes very small amount of power consumption compared with other hardware components. On the software side, we see that this mobile user is a frequent user of Google Maps due to his need of driving. Other than that, the Android Operating System and the Google Account Manager consume most of the energy on the software side.

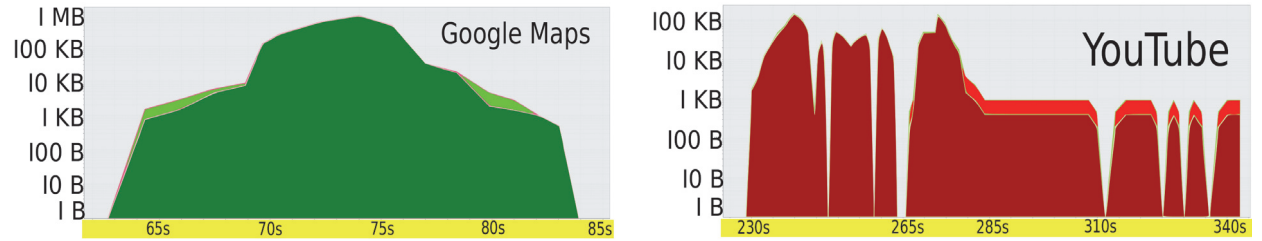
Table 3: Energy consumption in daytime.

Hardware (34%)	Screen (22.05%)
	Phone idle (7.66%)
	Mobile standby (3.12%)
	WiFi (1.14%)
	Bluetooth (0.05%)
	Voice call (<0.01%)
Software (66%)	Google Maps (19.01%)
	Android OS (18.63%)
	Android System (14.34%)
	Google Account Manager (10.16%)
	Clock (2.47%)
	WeChat (1.89%)
	Facebook (1.53%)
	Exchange Services (1.42%)
	WhatsApp (1.14%)
	Media server (1.09%)
	Chrome (1.06%)
	Phone (0.94%)
	System UI (0.73%)
	Health (0.56%)
	Email (0.55%)
	System UI (0.47%)
	Gmail (0.38%)
	Gallery (0.33%)

We repeat the same experiment on the phone during the night time. The purpose is to study the basic operation consumption on the mobile without using any mobile applications by the user. Table 4 shows that the hardware of the phone consumes around half of the power consumption. The software consumes the remaining half, mainly for running the Android System, Android OS, and Google Account Manager. While the user is sleeping, there is very little power consumption on mobile applications, such as WhatsApp and WeChat. This indicates that these applications run in the background mode and only check for new messages or updates from the application servers.

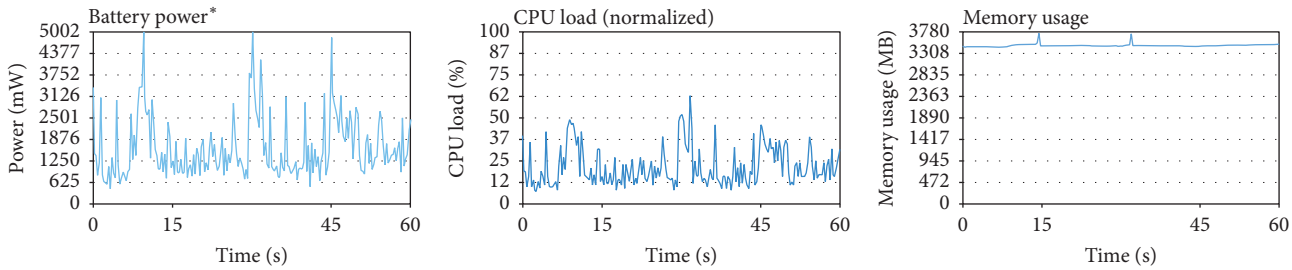
5.2. Profiling Data Communications of Mobile Applications. We use Qualcomm's Trepn Profiler [29] to study the data communications and resources that the applications consume on the mobile, including CPU usage, memory usage, and data usage.

As data communication is a major cause of fast energy drain in the smartphones, we show an interesting example of profiling data communication patterns of popular mobile applications. Figure 6 shows the data communication usage patterns of two popular applications, *Google Maps* and *YouTube*. From the data logs, we observe that both Google Maps and YouTube are running two threads in their applications. The data communications of the two threads in



(a) Google Maps (b) YouTube

Figure 6: Snapshots of profiled data communication patterns.



(a) Energy consumption of YouTube (b) CPU utilization of YouTube (c) Memory utilization of YouTube

Figure 7: Battery, CPU, and memory utilization of YouTube.

Table 4: Energy consumption in night time.

Hardware (53%)	Mobile standby (23.33%)
	Phone idle (17.99%)
	Screen (8.87%)
	WiFi (2.49%)
	Bluetooth (0.27%)
	Voice call (<0.01%)
Software (47%)	Android System (15.58%)
	Android OS (10.98%)
	Google Account Manager (8.81%)
	Clock (1.91%)
	WhatsApp (1.23%)
	WeChat (1.17%)
	Media server (1.15%)
	Chrome (0.89%)
	System UI (0.78%)
	Health (0.63%)
	Exchange Services (0.69%)
	Phone (0.47%)
	Email (0.39%)
	Google Maps (0.32%)
	Huawei Home (0.27%)
	Google (0.1%)
	Gmail (0.06%)
	Facebook (0.05%)

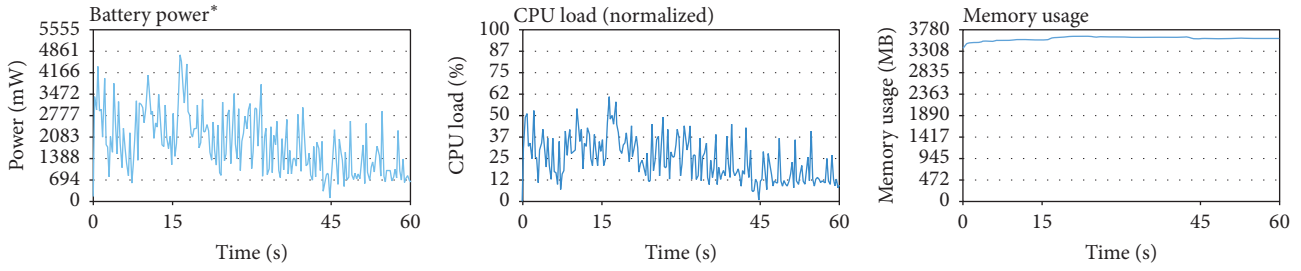
each application share similar patterns, which are indicated by dark and light colors in the figure.

Figure 6(a) shows a snapshot of data communication profiling of Google Maps. The y-axis is plotted in log scale, showing the size of data communication at different time. We observe that there is a sudden increase of data communication occurring at time interval 70 s–75 s. Through careful inspection, we find that it is due to the action of zooming in the map triggered by the user.

While profiling YouTube, a video is randomly picked for playing in the full screen mode. From Figure 6(b), we observe high initial data communications due to prefetching. Then, the video is played smoothly with constant data communication from 285 s. We believe that the intermittent communication pattern is due to the communication protocol and the reliability of the network.

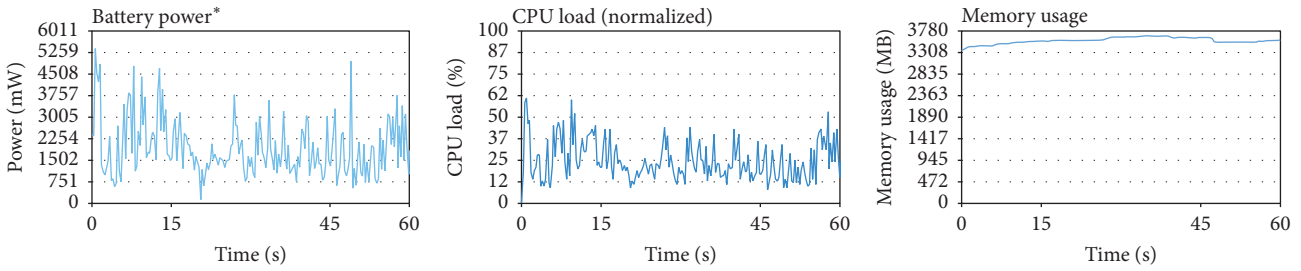
5.3. Profiling Battery, CPU, and Memory Utilization. We continue to profile the battery power consumption, CPU load, and memory utilization of three representative mobile applications, including YouTube, Google Maps, and Facebook. Figure 7(a) shows the battery power consumption of YouTube. At $t = 10$ s, a new video is played on YouTube, which triggers a peak on battery consumption. Similarly, at $t = 30$ s, the user selects another video, which triggers another high consumption of battery. At $t = 45$ s, the user rotates the screen to show the video in full screen, which leads to an increase of energy consumption. Figure 7(b) shows that the CPU load increases when new videos were played at $t = 10$ s and $t = 30$ s. We also observe an increase of memory usage with similar patterns in Figure 7(c).

Figure 8(a) shows the battery power consumption of Google Maps. At $t = 0$ s, the user enters a new search on a destination. At $t = 15$ s, the user starts the journey to the destination. We observe that the battery consumption is higher when user starts going on a new route. Figures 8(b)



(a) Energy consumption of Google Maps (b) CPU utilization of Google Maps (c) Memory utilization of Google Maps

Figure 8: Battery, CPU, and memory utilization of Google Maps.



(a) Energy consumption of Facebook (b) CPU utilization of Facebook (c) Memory utilization of Facebook

Figure 9: Battery, CPU, and memory utilization of Facebook.

and 8(c) show the corresponding CPU load and memory usage. We see that the CPU load follows similar pattern to the battery power, while memory usage remains more or less the same.

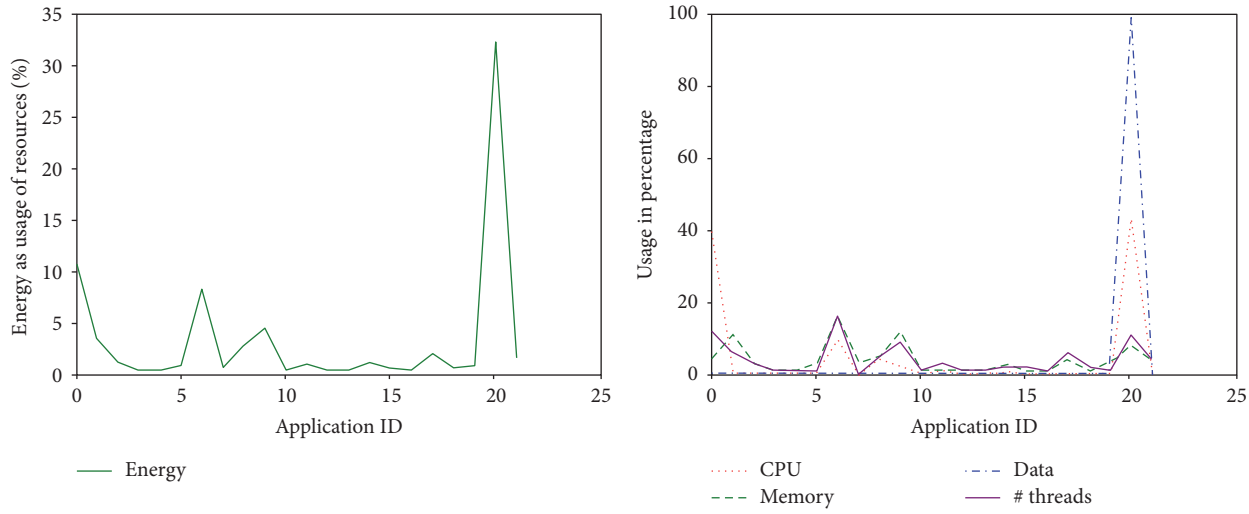
Figure 9(a) shows the battery power consumption of Facebook. At $t = 0$ s, the user logs into Facebook and then starts browsing at different posts of videos and images. As the user scrolls down the Facebook user interface, we observe different peaks in the battery power and the CPU load. These are believed to be the loading of new images and videos. Figures 9(b) and 9(c) show the corresponding CPU load and memory usage. We see that the CPU load follows similar pattern to the battery power, while memory usage remains roughly the same. The level of memory usage in Facebook is similar to those in YouTube and Google Maps.

5.4. Cloud-Based Data Profiling and Analysis

5.4.1. Energy Consumption. In order to understand and visualize the energy consumption pattern, the following experiment has been conducted. We collected the data from four users over a three-month period from 1 March 2015 to 31 May 2015. The users have been using Samsung S4 or NEXUS 5 smartphones. The data has recorded the energy consumption and resource usage of the smartphones that were idle or running actively in daily use. We have chosen twenty-two applications that were run by all the four users in this data analysis. These 22 applications range from social media applications, messaging applications, and navigation applications to personal management applications. The data has been cleaned up and processed, so that every resource utilization of each application has been summarized as a mean value in an hourly basis. Then, the summarized data are further aggregated to give an overview of energy consumption among all the applications.

We compare the total energy consumption and resource utilization distribution values among different applications in this experiment. This result helps us to identify the most power consuming applications and understand what resources have been utilized to make them so power hungry. Figure 10(a) compares the energy consumption values of the 22 applications. The y-axis shows the total energy consumption of each application in percentage (among all applications). The x-axis shows the application ID from 0 to 21. From the figure, we observe that there are four to five applications, which consume the most energy compared with the others. We rank the energy consumption percentage of these 22 applications in Table 5. It shows that the most power consuming applications are Facebook, followed by Android Operating System, Google Contacts Sync, and Google App.

We further investigate four key energy indicators (KEIs) in these applications, including CPU load, memory, data communication, and number of threads. Figure 10(b) shows the resource utilization of the applications in percentage. From the figure, we can see that the applications that consume most energy usually have high utilization in all four kinds of resources. Take Facebook as an example, it has the highest data communication and CPU usage among the 22 applications. It also has relatively high number of threads and memory usage compared with other applications. We observe similar resource utilization patterns for applications that have high power consumption. The shapes of the curves



(a) Energy consumption of applications

(b) Resource utilization of applications

Figure 10: Energy consumption and resource utilization.

Table 5: Ranked energy consumption of applications.

ID	Application name	Energy consumption (%)
20	Facebook	32.270
0	Android System	11.076
6	Google Contacts Sync	8.238
9	Google App	4.490
1	com.qualcomm.qcrilmsgtunnel	3.464
8	System UI	2.788
17	YouTube	2.006
21	Messenger	1.656
2	Nfc Service	1.204
14	Google Keyboard	1.146
11	CaptivePortalLogin	1.008
5	Media Storage	0.808
19	ES File Explorer	0.804
15	Maps	0.616
18	Google Connectivity Services	0.604
7	Google Dialer	0.602
13	Hangouts	0.410
16	Google+	0.406
10	Calendar	0.404
3	Calendar Storage	0.402
4	User Dictionary	0.400
12	Fit	0.400

in Figures 10(a) and 10(b) follow very similar patterns. It implies that these four selected resources are very important when profiling energy consumption for the smartphones.

5.4.2. CPU Load. We also observe different CPU load patterns in the applications. Figure 11 shows the CPU load of the Facebook App. We can see that the Facebook App has high CPU use when the app is started. After that, the CPU load is quite random depending on the operations triggered by the user, such as uploading photos or sending messages. Figure 12 shows the CPU load pattern of Google Maps, which is quite periodic due to the regular update of GPS locations.

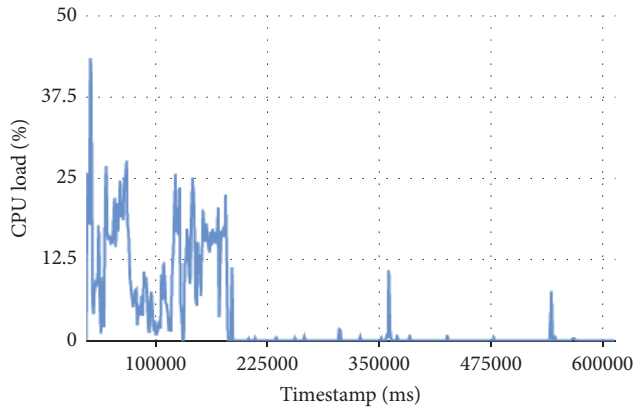


Figure 11: CPU load pattern of Facebook App.

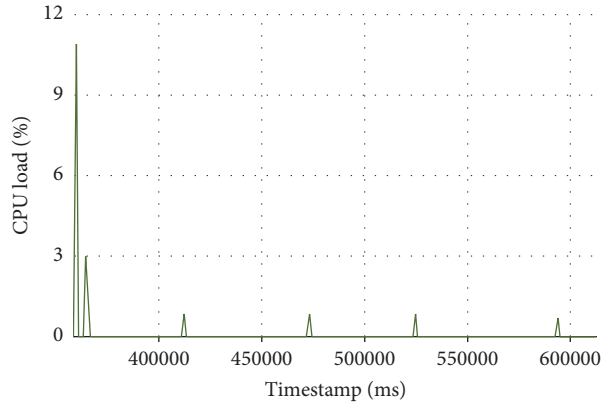


Figure 12: CPU load pattern of Google Maps.

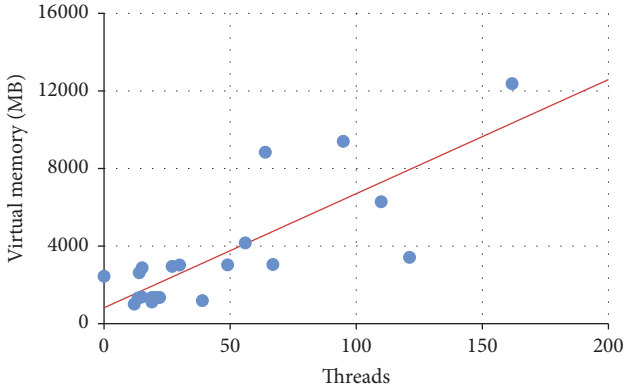


Figure 13: Threads and virtual memory use of applications.

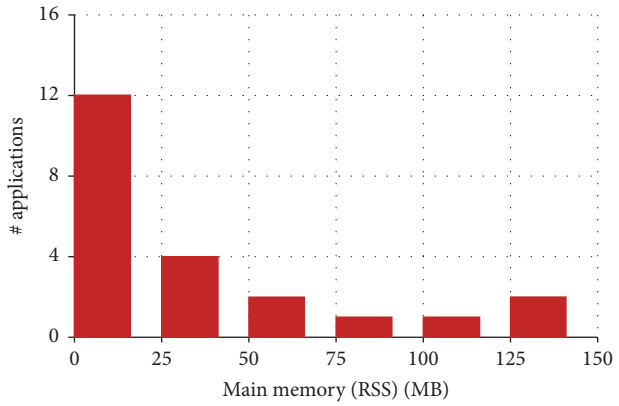


Figure 14: Main memory use of applications.

By observing the CPU patterns, it helps us to understand the operation characteristics and energy consumption of different applications.

5.4.3. Threads and Memory Use. Next, we analyse the correlation of threads and memory use in the applications. Figure 13 shows the average number of threads and the average virtual memory use of the 22 applications. As seen from the figure, there is a positive correlation between the number of threads and virtual memory use. Most of the applications use less than 50 threads. However, there are several applications using much more threads than the others. The top application, Google Contacts Sync, uses 162 threads and more than 12000 MB virtual memory. The applications with high number of threads are Android System and Facebook, which use more than 100 threads. Google App uses almost 100 threads and a lot of virtual memory as well.

Figure 14 shows the number of applications consuming different amount of main memory. We divide the memory use into different ranges and count the number of applications in each range. The figure shows that most of the applications consume less than 25 MB main memory. However, two applications, Facebook and Google Contacts Sync, consume almost 150 MB main memory. Google App also has high main memory use of 100 MB.

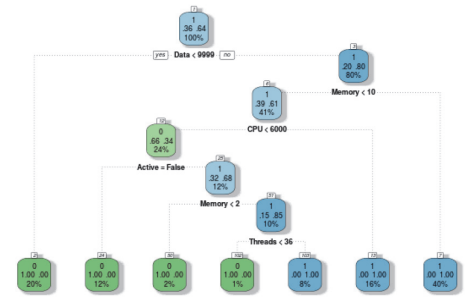


Figure 15: Screenshot of the energy data administration tool developed in the cloud.

5.4.4. Analysis for Energy-Efficient Profiling. Understanding the characteristics of applications and energy consumption patterns on the smartphones is very useful for reducing the energy consumption in the profiling process itself. Through simple analysis, we can make initial observations on what applications consume most energy and what applications consume insignificant amount. Our profiler can use this information to reduce the amount of data being collected on resource utilization. If we filter out the data from applications that consume less than 1% of the total energy in the system, we can greatly reduce the number of applications that require intense monitoring. Take the 22 applications in Table 2 as an example, we can reduce the number of data samples by 50%, while still keeping accurate data logs from applications that consume more than 94% of the total energy in the smartphones. In other words, it can save half of the energy in profiling without losing much accuracy in power monitoring. Even if the amount of data from smartphones to cloud is reduced to only 20%, the orchestrator can still capture almost 80% of the energy consumption from major applications through profiling. As the orchestrator learns about the characteristics of applications, it can configure the system dynamically to reduce the data samples of applications that are less frequently used or consume little energy. Thus, the amount of data that needs to be transferred from smartphones to the cloud will be significantly reduced.

6. Simulations

We have simulated smartphone data of 100 users with 20 applications. We prepared training data and test data samples containing 40,000 records and 20,000 records, respectively, for an hour of usage.

We first applied simple decision tree regression classifier, rpart (Recursive Partitioning and Regression Trees), available in R language. The decision tree classifier is tested against the test data set using 10-fold cross-validation, which results in 92% accuracy. To conduct the experiment using real data, the data must be formatted and normalized, and, most importantly, it requires sufficient amount of data. The experiment involves careful preprocessing of data to make sure of the validity of data. The data is affected by noise and various interruptions in real time.

Figure 15 shows an example decision tree, which can predict whether the power consumption is low or high given

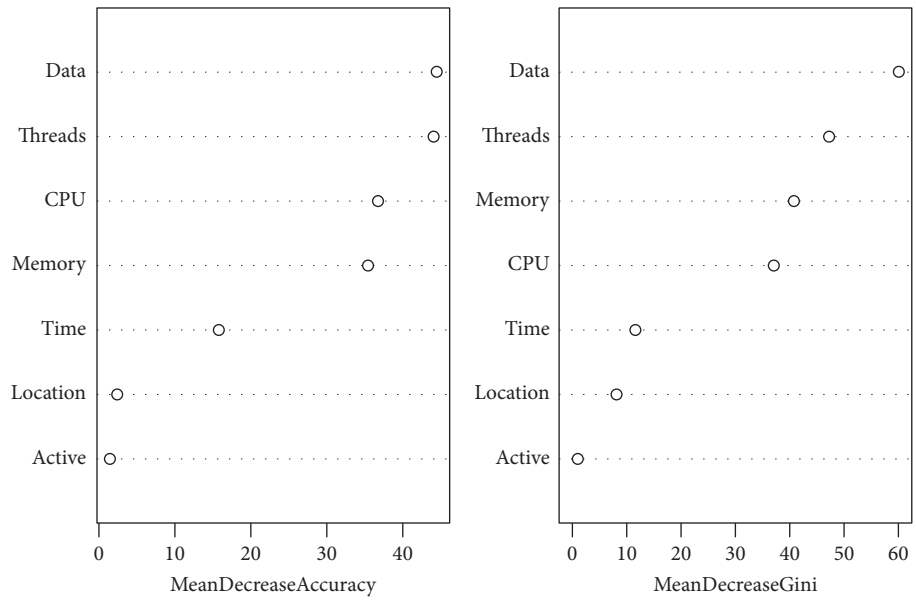


Figure 16: Results of random forest.

a resource utilization vector with variables (memory, data, threads, etc.).

The values of importance variables are predicted according to their influence on energy consumption. We found that the amount of data communications and the number of threads heavily influence the energy consumption. The decision tree model predicts for a given set of features (in the resource utilization vector) whether energy is high or low (0 or 1) as a binary classification problem. Recursive partitioning in a tree-structured model is used for classification. From Figure 16, we confirmed that the above features, such as the amount of data communications and number of threads, are the most influential factors in energy consumption by considering the Gini index of the random forest model.

7. Discussions

Orchestrator has been demonstrated as an effective behaviour predictor for the participating devices. The agent application installed on the smartphones reports logs to the orchestrator. It will be facilitated by the local validator and action triggers which will be regularly updated by the orchestrator on demand to reduce energy consumption in data logging, communication, and computation. Compared with smartphone based profiling, cloud-based orchestration can support crowdsourced profiling of energy consumption from multiple users. It also provides necessary resources to support advanced data mining and machine learning techniques. The analysis results from the cloud can be used to enhance local data filtering on the smartphones, which can further reduce energy consumption in the profiling process itself. In order to successfully deploy such orchestration service, we need to study and explore all the components and their interdependencies in detail.

One key contribution of this work is to reduce energy consumption in the profiling process itself by learning the trend and patterns in communications and operations of different mobile applications. Through learning the characteristics of the applications, the orchestrator can identify key applications and operations that consume most of the energy. It can make more accurate prediction and adjust the sample rates accordingly to minimize energy consumption in the profiling process itself. The learning results can also be used to support better operation system and application designs for energy saving.

Questions that we plan to further investigate include the following: (1) How to further reduce the energy consuming of profiler? (2) How to reduce data logs reporting and minimize energy consumption in data communication? (3) How to make orchestrator an epic predictor of device behaviours? (4) How to find optimal responsibilities of local agent by ensuring minimal computation and resources? (5) Is the current solution the best fit for mass open source contribution? (6) What are the most appropriate tools for energy-efficient cloud orchestration implementation? We plan to implement advanced features in the wisdom box and big data modules and to test the prototype iteratively. Other than smartphones, we would like to extend this framework for testing with different types of IoT devices.

We believe that the research questions regarding reducing energy consumption of the mobile devices will be most important for the future. Some initial ideas include exploring lightweight local data processing techniques on the mobile devices to reduce the amount of data logs to be reported to the cloud. We would also like to investigate machine learning methods based on crowdsourced data to profile the energy consumption of different applications both globally and individually.

8. Conclusions and Future Work

In this paper, we proposed a novel cloud orchestration framework for improving energy efficiency for smartphones in the IoT. The major advantage of this cloud orchestration is that it supports dynamic workflow and configuration of different processes and services. We have described the components of the orchestration and their interactions. Our architecture design is flexible, so that new components and advanced functions can be added to the system easily. The *big data*, *knowledge graph*, and *control box* are openly accessible, so that both single user and mass collaborators can participate and add new methods to the system.

We have conducted experiments using real resource utilization traces collected by four mobile users in a three-month period. The results demonstrated that our profiler can successfully characterize the energy consumption of different applications and identify the most power consuming applications. It can also give feedback to the energy profiler to reduce energy consumption in data logging. The amount of data logs can be reduced significantly through learning the key energy indicators and application characteristics. This iterative learning process can progressively reduce the communication and computation overheads in energy profiling. There is great potential that big data knowledge can be used for solving other problems as well, such as energy bug detection.

In the future, we would like to investigate advanced data mining and data filtering techniques to further reduce energy consumption in energy profiling. We will explore how data logging and communication can be optimized considering the application characteristics, usage pattern, and operation context.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was partially sponsored by the Swedish Governmental Agency Vinnova under Research Grant 2015-00347, STINT initiation grant for international collaboration IB2013-5237, and the Canadian Natural Sciences and Engineering Research Council. This work was also partially supported by the Shenzhen Engineering Laboratory for 3D Content Generating Technologies (no. [2017]476), Guangdong Technology Project (2016B010108010, 2016B010125003), National Basic Research Program of China (973 Program) (no. 2014CB744600), National Nature Science Foundation of China (61403365, 61402458, 61632014, and 61210010), and Program of International S&T Cooperation of MOST (no. 2013DFA11140). The authors would like to acknowledge the master thesis "Enabling Energy-Efficient Data Communication with Participatory Sensing and Mobile Cloud" written by P. Sathyamoorthy (2016), Uppsala University, Sweden (retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-274875).

References

[1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): a vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[2] J. Chase, *The Evolution of the Internet of Things*, Texas Instruments Incorporated, Dallas, TX, USA, 2013.

[3] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, 2015.

[4] S. Borenstein, "A Microeconomic Framework for Evaluating Energy Efficiency Rebound And Some Implications," National Bureau of Economic Research, 2013.

[5] International Energy Agency, More Data, Less Energy - Making Network Standby More Efficient in Billions of Connected Devices, © OECD/IEA, 2014 Licence: https://www.iea.org/tc/termsandconditions/.

[6] A. Carroll and G. Heiser, "An analysis of power consumption in a smartphone," in *Proceedings of the USENIX annual technical conference (USENIXATC'10)*, pp. 21-21, USENIX Association, Berkeley, CA, USA, 2010.

[7] A. Dzhagaryan, A. Milenković, M. Milosevic, and E. Jovanov, "An environment for automated measurement of energy consumed by mobile and embedded computing devices," *Measurement*, vol. 94, pp. 103–118, 2016.

[8] A. Tzanakaki, M. P. Anastasopoulos, S. Peng et al., "A converged network architecture for energy efficient mobile cloud computing," in *Proceedings of the International Conference on Optical Network Design and Modeling*, pp. 120–125, Stockholm, Sweden, May 2014.

[9] T. K. Kundu and K. Paul, "Improving Android performance and energy efficiency," in *Proceedings of the 24th International Conference on VLSI Design, VLSI Design 2011, Held Jointly with 10th International Conference on Embedded Systems*, pp. 256–261, Chennai, India, January 2011.

[10] P. Shu, F. Liu, H. Jin et al., "eTime: Energy-efficient transmission between cloud and mobile devices," in *Proceedings of the IEEE INFOCOM 2013 - IEEE Conference on Computer Communications*, pp. 195–199, Turin, Italy, April 2013.

[11] S. Nirjon, A. Nicoara, C.-H. Hsu, J. Singh, and J. Stankovic, "MultiNets: Policy oriented real-time switching of wireless interfaces on mobile devices," in *Proceedings of the 18th IEEE Real Time and Embedded Technology and Applications Symposium, RTAS 2012*, pp. 251–260, Beijing, China, April 2012.

[12] J. Tang, Z. Zhou, J. Niu, and Q. Wang, "An energy efficient hierarchical clustering index tree for facilitating time-correlated region queries in the Internet of Things," *Journal of Network and Computer Applications*, vol. 40, no. 1, pp. 1–11, 2014.

[13] A. Venčkauskas, N. Jusas, E. Kazanavičius, and V. Štuikys, "An Energy Efficient Protocol For The Internet Of Things," *Journal of Electrical Engineering*, vol. 66, no. 1, pp. 47–52, 2015.

[14] T. Zhang, X. Zhang, F. Liu, H. Leng, Q. Yu, and G. Liang, "ETrain: making wasted energy useful by utilizing heartbeats for mobile data transmissions," in *Proceedings of the 35th IEEE International Conference on Distributed Computing Systems, ICDCS 2015*, pp. 113–122, Columbus, OH, USA, July 2015.

[15] B. Steigerwald and A. Agrawal, "Developing Green Software," *Intel White Paper*, vol. 9, 2011.

[16] M. Altamimi, A. Abdrabou, K. Naik, and A. Nayak, "Energy cost models of smartphones for task offloading to the cloud," *IEEE*

Transactions on Emerging Topics in Computing, vol. 3, no. 3, pp. 384–398, 2015.

[17] Power Profiles for Android, July 2015, https://source.android.com/devices/tech/power/index.html.

[18] T. Dao, I. Singh, H. V. Madhyastha, S. V. Krishnamurthy, G. Cao, and P. Mohapatra, "TIDE: a user-centric tool for identifying energy hungry applications on smartphones," in *Proceedings of the 35th IEEE International Conference on Distributed Computing Systems, ICDCS 2015*, pp. 123–132, Columbus, OH, USA, July 2015.

[19] M. A. Hoque, M. Siekkinen, K. N. Khan, Y. Xiao, and S. Tarkoma, "Modeling, profiling, and debugging the energy consumption of mobile devices," *ACM Computing Surveys*, vol. 48, no. 3, 2015.

[20] A. Pathak, Y. C. Hu, and M. Zhang, "Where is the energy spent inside my app?: Fine grained energy accounting on smartphones with eprof," in *Proceedings of the 7th ACM European Conference on ComputerSystems, EuroSys'12*, pp. 29–42, Bern, Switzerland, April 2012.

[21] A. Nettstrater, M. Bauer, M. Boussard et al., "Internet of Things-Architecture IoT-A Deliverable D1. 3–Updated reference model for IoT v1.5," 2012, http://www.meet-iot.eu/deliverables-IOTA/D1_3.pdf.

[22] A. Gonzlez Garca, M. Alvarez Alvarez, J. Pascual Espada et al., "Introduction to devices orchestration in internet of things using SBPMN," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 1, no. 4, pp. 16–22, 2011.

[23] F. Liu, P. Shu, and J. C. S. Lui, "AppATP: An Energy Conserving Adaptive Mobile-Cloud Transmission Protocol," *IEEE Transactions on Computers*, vol. 64, no. 11, pp. 3051–3063, 2015.

[24] A. J. Oliner, A. P. Iyer, I. Stoica, E. Lagerspetz, and S. Tarkoma, "Carat: collaborative energy diagnosis for mobile devices," in *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems (SenSys '13)*, pp. 10:1–10:14, Rome, Italy, November 2013.

[25] L. Atzori, A. Iera, and G. Morabito, "The internet of things: a survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[26] P. Sathyamoorthy, E. C.-H. Ngai, X. Hu, and V. C. M. Leung, "Energy efficiency as an orchestration service for mobile internet of things," in *Proceedings of the 7th IEEE International Conference on Cloud Computing Technology and Science, CloudCom 2015*, pp. 155–162, Vancouver, BC, USA, December 2015.

[27] Android Studio, April 2017, "Logcat Command-line Tool" https://developer.android.com/studio/command-line/logcat.html.

[28] Android Studio, April 2017, "Dalvik Debug Monitor Server (DDMS)" https://developer.android.com/studio/profile/ddms.html.

[29] Qualcomm Technologies, Inc., April 2015, "Trepn Profiler" https://developer.qualcomm.com/mobile-development/increase-app-performance/trepn-profiler.

[30] GraphQL, "Introduction to GraphQL, April 2017, http://graphql.org/learn/.

[31] A. Pathak, Y. C. Hu, M. Zhang, P. Bahl, and Y.-M. Wang, "Fine-grained power modeling for smartphones using system call tracing," in *Proceedings of the 6th ACM EuroSys Conference on Computer Systems, EuroSys 2011*, pp. 153–167, April 2011.

[32] Emma Jane Hogbin, *ACPI: Advanced Configuration and Power Interface*, CreateSpace Independent Publishing Platform, 2015.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 4974165, 22 pages
https://doi.org/10.1155/2017/4974165

Research Article

Efficient Network Coding with Interference-Awareness and Neighbor States Updating in Wireless Networks

Xiaojiang Chen, Jingjing Zhao, Dan Xu, Shumin Cao, Haitao Li, Xianjia Meng, and Dingyi Fang

Northwest University, Xi'an, China

Correspondence should be addressed to Xianjia Meng; xianjiam@nwu.edu.cn and Dingyi Fang; dyf@nwu.edu.cn

Received 11 March 2017; Accepted 2 July 2017; Published 12 September 2017

Academic Editor: Bernard Cousin

Network coding is emerging as a promising technique that can provide significant improvements in the throughput of Internet of Things (IoT). Previous network coding schemes focus on several nodes, regardless of the topology and communication range in the whole network. Consequently, these schemes are greedy. Namely, all opportunities of combinations of packets in these nodes are exploited. We demonstrate that there is still room for whole network throughput improvement for these greedy design principles. Thus, in this paper, we propose a novel network coding scheme, ECS (Efficient Coding Scheme), which is designed to achieve a higher throughput improvement with lower computational complexity and buffer occupancy compared to current greedy schemes for wireless mesh networks. ECS utilizes the knowledge of the topologies to minimize interference and obtain more throughput. We also prove that the widely used expected transmission count metric (ETX) in opportunistic listening has an inherent error ratio that would lead to decoding failure. ECS therefore exploits a more reliable broadcast protocol to decrease the impact of this inherent error ratio in ETX. Simulation results show that ECS can greatly improve the performance of network coding and decrease buffer occupancy.

1. Introduction

Have you ever felt exhausted for the bad network while you were just immersed in your computer games like LoL? Throughput improvement in wireless mesh networks is also a major problem for social networking and e-commerce platforms, where servers suffer dramatically heavy traffic and users are dying for fast and reliable network services. Many industries are looking to address this problem and improve throughput. Applications that involve high throughput, such as Ix [1] and Prophecy [2], provide high throughput performance with proxied or operating systems. These systems, however, call for hardware or system update and may be very costly. Moreover, it is unpractical for users to update system or hardware continuously. Network coding that emerged as a promising technique that can provide significant improvements with low cost, on the other hand, has attracted much interest. The notion of network coding allows relay nodes to not only forward but also process (i.e., XOR) the incoming packets. Despite all the scenarios mentioned before, network coding is also a great opportunity for IoT as many sensors can transmit information at the same time. But there is a main question one may ask: how many packets should be encoded (XOR) to get better improvements? Most current schemes, however, encode as many packets as possible to get maximized throughput at relays, which does not fully exploit network utility and throughput in the whole network. The whole network, however, would benefit from this bandwidth-efficient technique, with the same bandwidth but much higher throughput gains.

The foreground of potential throughput gain has received much attention since it was first proposed in [3]. It is subsequently found to be particularly suitable for wireless networks, since the broadcast nature allows the transmitting packets to be overheard by all its neighbors. Prior work focuses on addressing two problems: (i) throughput gain bounds with various scheduling schemes and theoretical analysis on encoding part [4–6] and (ii) transport protocols on decoding

part [7–9]. However, most of the network coding schemes like [10, 11], are under the assumption that the buffer and computing resource are sufficient, which is unlikely in most wireless mesh networks. Consequently, these schemes are greedy, exploiting all opportunities of broadcasting combinations of packets in a single transmission. However, with the fading and interference in wireless networks, the results of these greedy schemes are not always satisfying, which sometimes even shows a decrease of throughput [6]. Moreover, while there is much literature on the throughput improvement and bounds [4–6], it is still unclear for every relay to obtain better performance with proper encoding threshold when topology changes in the network. Here, we use encoding threshold to refer to the maximum number of native packets that can be encoded at a relay. This unclearness, however, is exacerbated in wireless mesh networks, where interference is more severe. That would leave the better throughput inaccessible.

This paper introduces ECS, an interference-aware network coding scheme in a robust manner which works in wireless mesh networks. In line with common practice in network coding schemes, ECS transfers encoded packets and overhears neighbors' packets for decoding. The challenge, however, is to make decisions about the encoding threshold of packets to encode once for every relay node (i.e., encoding number) to get more throughput gains with less interference and make full use of nodes in the whole network.

Unlike existing approaches, which consider buffer and computing resource as arbitrarily large, ECS exploits an interference-awareness model for relays in wireless networks, which maximizes network utility and throughput gains by reducing interference. Particularly, encoding too many packets at one node is unpractical with its limited resources (both buffer and computation). Moreover, due to the large number of combinations of native packets, relays have to forward the encoded packet to a large swath of receivers, whose neighbors have to keep in NAV timer to avoid collision with congestion control. Such nodes that keep in NAV timer, however, are a waste of resource, since even though they are not in the communication range of senders, they still cannot transmit any data for a certain period of time. Therefore, throughput can be improved if one can deal with the low utilization of nodes in NAV timer. That is to say, if we can make good use of nodes in the network, we can better meet the demand for throughput increment in wireless mesh networks.

To illustrate ECS's approach, Figure 1(a) shows a toy example with 11 nodes leveraging greedy network coding scheme, where the center node broadcasts an encoded packet of four native packets to the other four receiving nodes, which improves the throughput with less transmissions. It is also noticeable that all the silent nodes have to keep silent to avoid wireless interference, namely, waiting for the end of the transmission because of the occupancy of the wireless medium, even though they are not within communication range of the sending nodes. The large swaths of silent nodes are a waste of network resource; therefore throughput can be improved by "awaking" them. So can we "awake" maximum number of silent nodes to get more throughput gains? Can we get more throughput gains? The answer is yes. To do this, we transform these silent nodes to senders or receivers, while the sending coverage remains in a high ratio. Here, sending coverage refers to the ratio of sending nodes to all nodes in the whole network. Consider the scenario in Figure 1(b), where transmitters turn to be 3 and silent nodes in contrast turn to be 2, decreasing by 66.7%. The current throughput is 6, 1.25 times of that in Figure 1(a). Note that the sending coverage in Figure 1(a) is 9.09%, while the sending coverage in optimized solution shown in Figure 1(b) rises to 27.3%. In this way, we maximize the network capacity with minimum number of nodes in NAV timer, that is, minimum interference. In addition, receivers in the encoding scheme shown in Figure 1(b) store fewer packets in their buffer, reducing $2/4 = 50\%$. With less buffer occupancy than in Figure 1(a), the scheme in Figure 1(b) yields better results. That is to say, the seemingly best solution to improve throughput in Figure 1(a) indeed sacrifices the potential throughput for a maximum number of encoded packets delivered in a single transmission. Hence, the performance of greedy schemes alone, like the transmissions from Figure 1(a), can be improved by enlarging the sending coverage and decreasing the number of idle nodes. And that is why it can decrease possible interference while transmitting. An efficient networking coding scheme therefore needs to consider a main question: how to enlarge the sending coverage and decrease the number of idle nodes at the same time?

To do so, we divide nodes into three colors according to the three statuses: sending, receiving, and silent. By doing so, we simplify the problem into a graph coloring problem, a classical algorithm that deals with assignment. In contrast to the illustrative example in Figure 1(a), however, the evaluation of throughput in real-world wireless mesh networks, instead of a single transmission, is the total throughput in the whole network. Current schemes, however, are the theoretical bounds or are adaptive for certain given topologies, which are unlikely in wireless mesh networks. In fact, the changes of topologies are inevitable. In designing a scheme that can enlarge sending coverage as well as considering the changes of topology, we modify the objective function to minimize nodes in NAV timer and keep sending nodes in a high ratio, differentiating from the classical graph coloring. For every relay node, it "colors" nodes into as many senders as possible with certain limits to avoid collision and with the same strategy "colors" the neighbors of these senders to receivers. However, this is time-consuming and too complicated as the network scales up. Thus, to jump out of the local optimum, we combine the classical graph coloring algorithm with simulated annealing (SA) to "cool" the results. For every result in coloring, we iterate and discard the former result if the current result has better performance; otherwise, we keep it in a random rate. The end of iteration mainly depends on the initial temperature T and current temperature T', which are discussed in Section 5.4. After these procedures mentioned above, relays know the upper bound of encoding numbers, that is, refer to the number of their surrounding senders. Here, we call this upper bound of encoding numbers as encoding number threshold. Given these certain encoding number thresholds for every relay, the network encodes proper number of packets, which efficiently decrease interference and make good use of nodes. Therefore, the

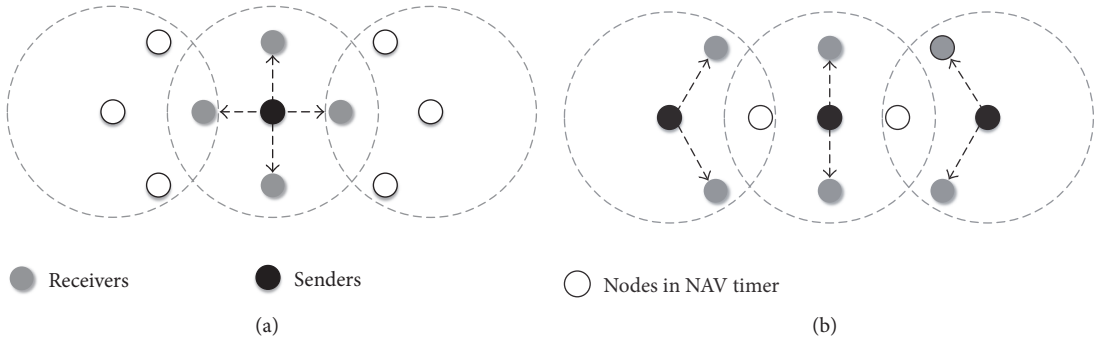


Figure 1: *Intuition underlying ECS's use of interference-aware model in coding packets*. (a) uses COPE as four packets encoded at the relay (black node), while (b) demonstrates our approach: two packets are, respectively, encoded at three relays, providing more throughput improvement.

whole network can obtain more throughput gains. We present the design in Section 5 and demonstrate in Section 5.3 the time complexity of the two.

Finally, while one could overhear packets and guess whether a neighbor has a particular packet leveraging opportunistic listening, network coding scheme today would require periodically computing ETX metric [12], which may make incorrect decisions in practice. Moreover, this is exacerbated when the network is of poor link quality. To solve this problem, relays can send request to get the information about what their neighbors store in the buffer. However, it is too costly to occupy the medium to exchange buffer information for each single transmission and therefore may reduce throughput. Hence, if we can find a low-cost approach to learn neighbor states and decrease the inherent error rate in ETX, the decoding procedure can be ensured. ECS carries decoding and encoding information while reserving the channel, namely, carries the total packets number encoded or buffered in RTS and CTS, which is in low cost and efficiently reduces the decoding failure.

Our simulations lead to the following findings:

(i) Receiving $\omega = [10, 300]$ packets per second, the throughput of ECS is 7 to 40 times higher than that of 802.11a/b/g and 3 times higher than that of greedy schemes, such as COPE scheme.

(ii) With same throughput gains, ECS stores fewer packets in the buffer than COPE, indicating that ECS has better efficiency and is of low cost.

Our contribution of this paper includes the following:

(i) This work presents a novel efficient scheme that scales whole network throughput by decreasing interference from overwhelmed encoding packets. To achieve this, we dynamically calculate an encoding number threshold for every relay node, according to communication range from topology. Finally, we show that our scheme can deliver throughput gains in mesh networks and any other networks that suffer great traffic.

(ii) We present a neighborhood listening state protocol, which reduces the unavoidable incorrectness when leveraging ETX by carrying coding information in RTS/CTS. Therefore, it decreases retransmissions and decoding failures in a low-cost approach.

2. Background

The attractiveness of network coding is marrying XOR and sending packets. Thus, it is a prerequisite for receivers to possess the ability of decoding the XOR-ed packets. To do so, receivers buffer packets that can be used to decode XOR-ed packets, which we call decoding resource. Differentiating the decoding resource, the network coding is classified into two categories. The first type of decoding resource only stores packets that have been sent by nodes, while the decoding resource of the second type includes the packets by opportunistic listening in addition.

The first category of network coding is demonstrated in Figure 2(a). In conventional cases, when Alice and Bob want to swap a pair of packets with the help of a router, the router has to forward the packets from two senders separately. Therefore, the number of transmissions needed is 4. After leveraging network coding, the router just needs to XOR the packets sent by Alice and Bob and broadcasts the XOR-ed packet. Alice and Bob can both hear the XOR-ed version and use their decoding resource to decode the XOR-ed packets. For example, Alice XORs the XOR-ed packet with her original sending packet; then she can get the message Bob wants to send to her. Consequently, the number of transmissions falls to 3.

The second category of network coding works in a more sophisticated network scenario. Take Figure 2(b) as an example. When these four persons want to exchange packets with network coding in a wheel topology, the senders not only send but also opportunistically listen to others. For example, let us assume that Alice can only have packet p_1 in its decoding resource in the first category, but it can also buffer packet p_4 and packet p_2 via overhearing to Sally and John. The relay then broadcasts packet p ($p = p_1 \oplus p_2 \oplus p_3 \oplus p_4$) after the receivers (but also the senders in this scenario) have enough decoding resource, and the receivers can decode the packet.

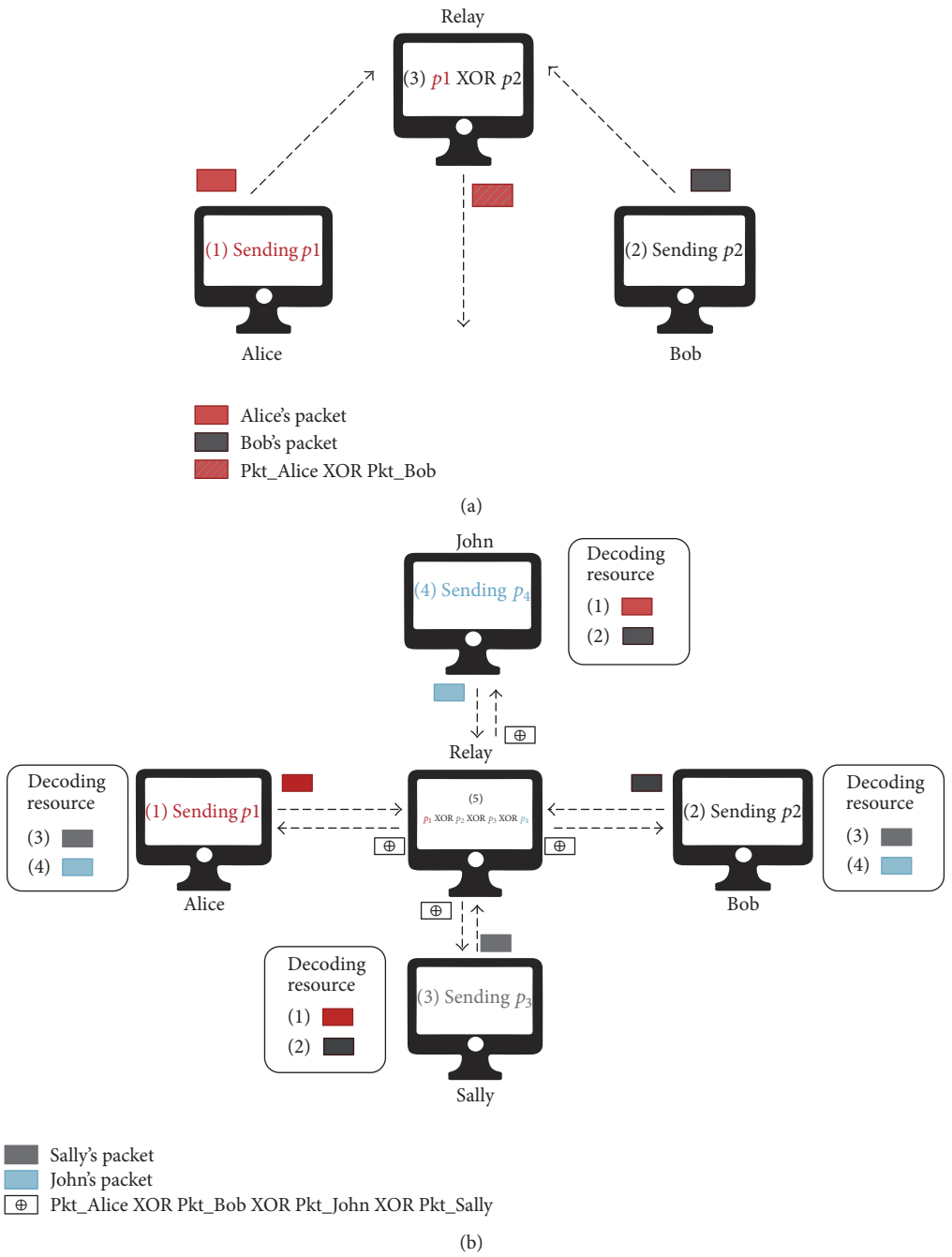


Figure 2: A brief illustration of network coding.

3. ECS Overview

ECS is an interference-aware scheme that can provide lower buffer occupancy and yield a higher throughput performance, much adaptive than most of the greedy network coding schemes including COPE [10]. To better improve the network throughout, ECS goes through the following steps:

(i) Depending on whether there are many encoding opportunities in the network, ECS first decides the start time of the network coding, using the technique in Section 4.

(ii) By modeling and giving solutions to the interference-aware model as described in Section 5, ECS calculates the maximum number of packets that a relay node can encode with the knowledge of topology.

(iii) Section 6 presents a neighbor status updating protocol. ECS enhances the accuracy on neighbor learning

scheme by adding encoding and decoding information in RTS and CTS while reserving the channel.

The next few sections elaborate on the above things, providing the technical details.

4. Encoding Time Decision

In order to deal with latency when there are few opportunities of encoding that can be exploited, we introduce encoding time decision, which turns on network coding at relays only when there are more encoding opportunities around them. Note that a key feature which arises in network coding is that when relays want to encode several packets in the transmission, they wait a period of time θ for the arrival of packets to encode, that is, encoding opportunities. In addition, all the receivers also have to wait for this encoding process and spend time decoding. In conclusion, the use of network coding increases latency in the network. This latency is inevitable in practice and is exacerbated in the cases where less opportunities of broadcasting combinations of packets can be exploited. The lack of opportunities leads to longer delays and provides fewer throughput gains. Hence, we need to find a tradeoff between latency and encoding opportunities and start to use network coding when the encoding opportunities exceed a certain threshold. Current schemes, however, leverage network coding all the time for throughput gains, which brings unwelcome latency especially when the throughput improvement is not tremendous. Our schemes, therefore, can efficiently reduce latency.

In order to decide the scenarios to start network coding, we first introduce two parameters that needs to be considered and then give some reasonable values after our simulations.

4.1. Definition

Inverse Packets. Let p_1 and p_2 be the packets relay R receives sequently. s_i and n_i denote the previous hop and next hop of p_i, respectively. t_i is the time when packet p_i has been received by the relay R. The neighbors set of m is denoted as $N(m)$. If p_1, p_2 are a pair of inverse packets of the relay R, then $n_1 \in N(s_2)$, $n_1 \notin N(s_1)$, $n_2 \in N(s_1)$, $n_2 \notin N(s_2)$, and $|t_1 - t_2| < \theta$. θ denotes the time period waiting for the arrival of inverse packets.

Take Figure 2(b) as an example. Let us assume that John wants to send the packet p_1 to Sally, while Sally also sends a packet p_2 to John. When both p_1 and p_2 arrive at the relay, the next hop of p_1 turns to Sally and the next hop for p_2 is John. Therefore, $n_1 = Sally$, $s_1 = John$, $n_2 = John$, $s_2 = Sally$. The neighbors set of John $N(s_1) = \{Alice, Bob\}$, and Sally has the same neighbors' set as John. Therefore, p_2 and p_1 satisfy the requirements: $n_1 \in N(s_2)$, $n_1 \notin N(s_1)$, $n_2 \in N(s_1)$, and $n_2 \notin N(s_2)$. t_1 and t_2 denote the times when packets p_1 and p_2 have been received by the relay R, respectively. If t_1 and t_2 satisfy $|t_1 - t_2| < \theta$, then p_1 and p_2 are inverse packets.

Data Flow Complexity (DFC). If a relay receives r native packets that contain i pairs of inverse packets in a unit of time, then DFC $= 2i/r$.

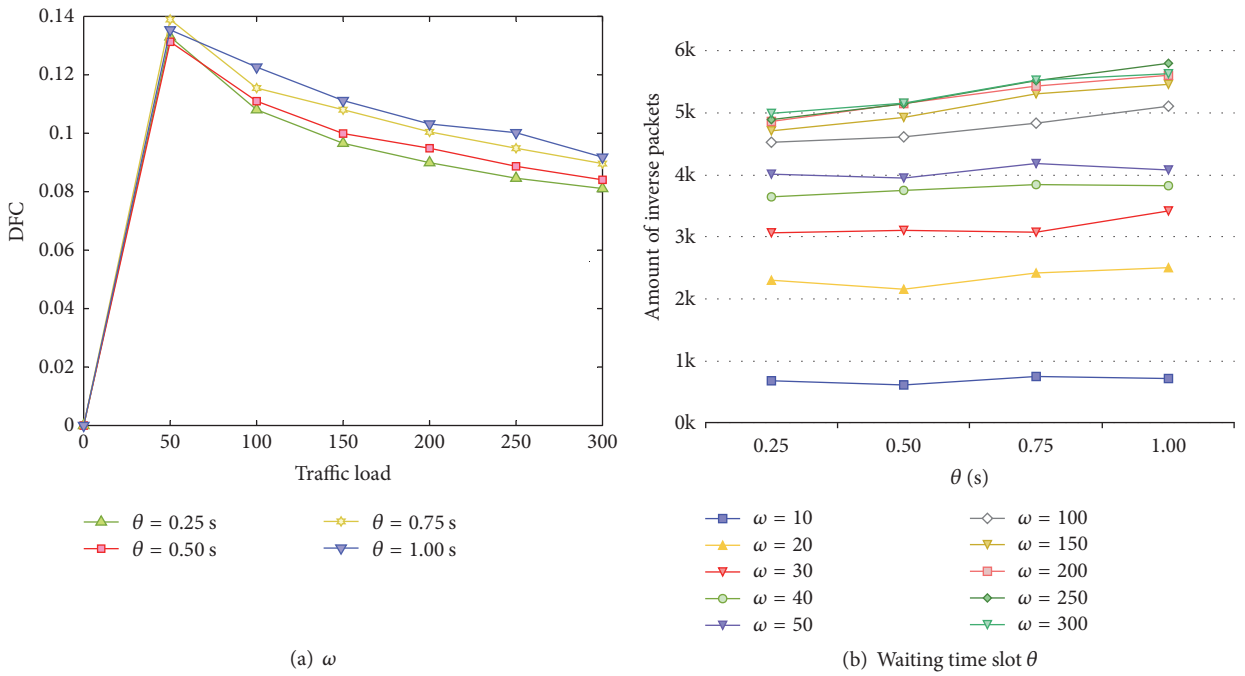
In summary, inverse packets are a pair of packets which can be encoded at a relay and decoded at receivers. The ratio of inverse packets to native packets, DFC, can therefore denote the encoding opportunities regardless of load. As shown in the definition, the calculation of DFC does not rely on topology or other network status; rather, it can be applied in any networks separately. The nodes would not start using network coding unless DFC exceeds a certain threshold ζ. Starting time decision is just like a sensor in voice-activated light in hall, which only turns on when it is noisy in the hall. As opposed to listening to the noise around, however, ECS exploits the DFC to identify the data complexity and the right time to use network coding.

4.2. Calibrating Parameters. In this part, we calibrate parameters ζ and θ through numerous simulations. It should be noted that the related parameters can be modified with various users' need. One important setting of our simulation is that we select sender and receiver randomly, and nodes are randomly distributed. Thus, we set up an ad hoc network with randomly distributed nodes.

4.2.1. Open Time Threshold ζ. With the same setup detailed in Section 7, we turn off network coding in the simulation and let the nodes send packets according to the set of ω (the amount of sending packets per second) and count the values of DFC according to θ with the 802.11 protocol. In the simulation $\omega = [10, 20, 30, 40, 50, \ldots, 300]$, there are 30 kinds of values in total, and in $\theta = [0.25, 0.5, 0.75, 1.00]$, there are 4 kinds of values in total. Therefore, every group of simulations (120 groups in all) repeats 1000 times. The results of simulation are shown in Figure 3.

Among the factors that can affect the DFC, θ in inverse packets definition and parameter ω are the most critical parameters, due to controlling network load. The impact of network traffic load (ω) on data flow complexity (DFC) is depicted in Figure 3(a). For small load, nodes are more likely to forward packets in the forwarding queue to the next hop. However, as few packets are stored in the forward queues, new arrival packets are hard to be matched into inverse packets, and that is why DFC equals 0.07 when $\omega = 10$. As the load increases ($\omega < 30$), the number of packets in forwarding queues increases, leading to a higher DFC rate and a larger amount of inverse packets. When the number of packets in forwarding queues exceeds a certain threshold, DFC levels off, reflecting the fact that inverse packets have reached their capacity and cannot combine additional inverse packets. Additionally, with the variation of θ, DFC shows only slight difference as load (ω) goes up, which reveals that ω has a greater impact on DFC than θ.

Figure 3(b) plots the amount of inverse packets with a group of ω when θ goes up. When ω is low, the change of the number of inverse packets is very slight. Hence, the impact of θ on inverse packets is correspondingly modest. As ω increases, forwarding queues have more packets due to the high load; thus the amount of inverse packets goes up. Surprisingly, these curves are very close, reflecting the fact that network has reached its capacity and cannot sustain additional load.



(a) ω (b) Waiting time slot θ

FIGURE 3: The number of sending packets ω and waiting time slot θ versus DFC.

The above simulation results reveal that when $\omega > 10$ (unit: packets per second), DFC is always beyond 0.07. From Figure 3(b), when $\omega = 10$, the amount of inverse packets is scarce, fewer than 1000 per minute. In these scenarios, packets can easily access the medium to transfer data due to the low load. The coding opportunities are also scarce due to the sparse packets in the forwarding queue. Thus, we suppose that very few throughput gains can be obtained with ECS in these scenarios. In summary, we take $\zeta = 0.07$. That is to say, if and only if $\zeta > 0.07$, the coding procedure will be started.

4.2.2. Waiting Period θ. In this part, we discuss θ, the waiting period for the arrival of inverse packets. Recall from inverse packets' definition that θ is the time waiting for inverse packets. We also set θ as the initial value for timer to overhear packets. In a same network scenario, the more time node waits for the inverse packets, the more packets it overhears and stores in its buffer. Thus we need to consider the buffer occupancy of nodes in the whole network, that is, how many packets nodes have overheard.

We start our flows again according to a Poisson process with ECS, that is, pick sender and receiver randomly and allocate certain memory (1 MB) for every node to store decoding resource. There are extra 10 MB for spare memory. That is to say, nodes will store decoding resource in spare memory after consuming 1 MB. The resource will be discarded if both memories run out. We classify nodes into three categories according to the usage of memory: overloaded nodes for those that use more than 1 MB memory; low-use nodes referring to those nodes whose memory usage is less than 0.25 MB; and the overhead of healthy nodes which is between 0.25 MB and 1 MB.

We repeat the simulation in varying load ω with ECS enabled. Figure 4 plots the distribution of the ratio of these three kinds of nodes. Figure 4(a) shows that the majority of nodes are unable to utilize the memory efficiently when the load is low, and the ratio remains flat with increasing θ. This illuminates weak correlation between θ and memory usage. Note that only those on the network trunk intersection fully utilize the memory. As the load increases, some low-use nodes turn into healthy nodes, while the memory usage of some healthy nodes ramps up into overloaded nodes. These transactions are also related to the location of nodes. For those healthy nodes in the trunk, they turn to overloaded nodes and then the low-use nodes in subtrunk become healthy nodes. The low-use nodes in the margin, however, are still in low use due to the lack of flow. As a result, nodes in low use decrease gradually, while overloaded nodes leap. But it should be emphasized that this does not mean that waiting period is related to the topology because the topology does not decide how many packets nodes send. As load surges even higher, nodes in low use decrease, almost equal to the amount of healthy nodes. Surprisingly, overloaded nodes decrease, reflecting that low-use nodes in relative margin networks turn into healthy nodes as load in subtrunk network goes up.

The reason that we make decisions on θ is to increase the ratio of healthy nodes and decrease the overloaded nodes, most obviously shown in Figure 4. As a result, we set $\theta = 0.3$ in our later simulation. Note that load also affects the number of these three categories nodes, which, however, is unpractical to control with the setting. Thus we use θ to control the ratio of these three kinds of nodes. According to this observation, the ratio of overloaded nodes is below 5%, while the number of healthy nodes is increasing when $\theta = 0.3$. Thus θ is set to be 0.3.

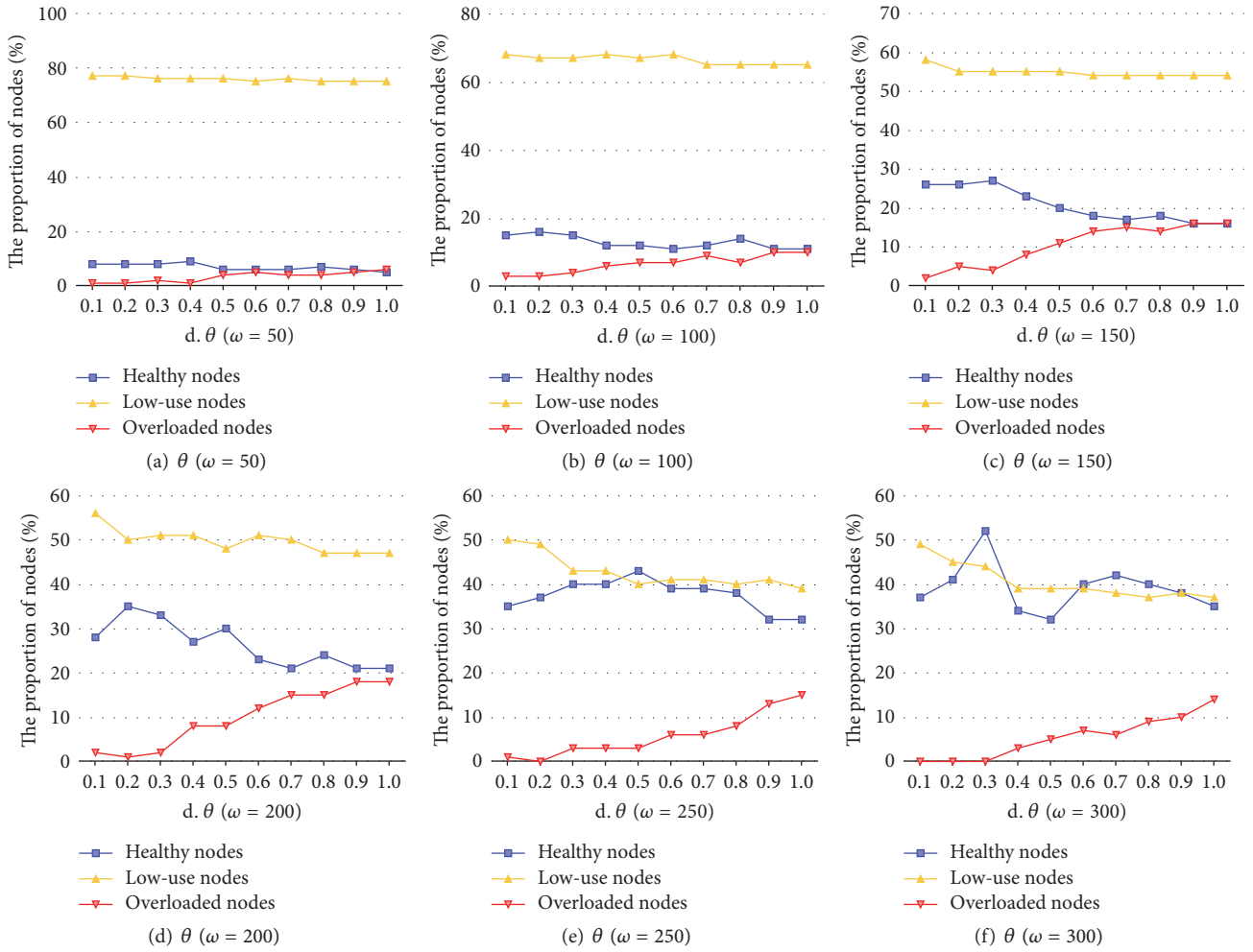


FIGURE 4: The proportion of all states (healthy, low-use, and overloaded) of nodes in various time periods θ.

5. Encoding Number Threshold Model

After acquiring the start point of coding scheme, ECS figures out the number of packets needed to be coded together in a single transmission to get more throughput gains. Many greedy schemes including COPE do not constrain this number, pursuing maximum throughput in a single transmission, which would encounter encoding and decoding failure and, most importantly, may reduce throughput when applied in the whole network. Exploiting not enough combinations of packets to broadcast, on the contrary, leaves much space for throughput gains due to the potential opportunities. ECS therefore needs to find a tradeoff between throughput gains and encoding numbers. Current schemes, however, are adaptive only for several general and unchanging topologies, which is unlikely in wireless mesh networks. In fact, the change of topology is inevitable. By maximizing utilization ratio every time when the topology changes to obtain throughput gains in the whole network, ECS makes use of topology and communication range, which yields a better result.

In order to get the encoding number threshold of packets, we divide nodes into three colors according to the three statuses: sending, receiving, and silent. By doing so, we simplify the problem into graph coloring, a classical algorithm to deal with assignment. Rather than coloring neighbor nodes with different colors to get a minimum number of colors in traditional coloring problem, we modify the objective function to maximize the difference between the ratio of total transmitters and the ratio of silent nodes to get optimal throughput gains in wireless mesh networks. Then we leverage SA to "cool" the results. In designing a scheme to jump out the local optimal results, we combine the classical graph coloring algorithm with simulated annealing (SA).

5.1. Formulation. As mentioned before, ECS aims at high usage rate and sending coverage. Therefore, the scheme is simplified into a kind of problem of minimizing nodes in NAV timer (silent nodes) while keeping sending nodes in a fairly high rate, which belongs to a particular channel assignment.

Coloring algorithm in connected graphs and conflict graphs works well for such channel assignment problems. Instead of coloring neighbors with a minimum number of different colors, the evaluation functions in ECS are depicted

as follows. Let an undirected graph $G = (V, E)$ represent a wireless mesh network: $|V| = n$ and $|E| = m$. Node v_i is the node in the network, while e_{ij} is the single-frequency link. The optional status for each node is in $C = \{1, 2, 3\}$ where 1 is for sending, 2 is for receiving, and 3 is for silent. Let $x_{ic} = 1$ be true to color node i with color c and 0 for not coloring with color c. And the model is shown as follows:

$$\begin{aligned}
\max \quad & f(x) = \frac{\left(\sum_{i=1}^{n} x_{i1} - \sum_{i=1}^{n} x_{i3}\right)}{n} \\
\text{s.t.} \quad & x_{ic} \in \{0, 1\}, \quad 1 \le i \le n, \ c \in C \\
& \sum_{k \in C} x_{ik} = 1, \quad i = 1, 2, 3, \ldots, n \\
& \sum_{e_{ij} \in E} x_{j1} = 0, \\
& x_{i2} = 1 \\
& \sum_{e_{ij} \in E} x_{j2} \in \{1\} \cup \{2n \mid n \in N_+\}, \\
& x_{i1} = 1.
\end{aligned} \tag{1}$$

5.2. Design Specification. In finding solution to this problem, we are inspired by simulated annealing algorithm. Simulated annealing algorithm is a combination of random algorithm and greedy algorithm. Greedy algorithm can get a swift solution but may easily be stuck in local optimum. Simulated annealing algorithm, on the contrary, can approach global optimal solution by discarding local optimal solutions with a certain probability.

Let T (sufficiently large) be the original temperature of annealing algorithm. The end of temperature of annealing algorithm is T' ($T' > 0$) and annealing rate is r ($0 < r < 1$). The evaluation function of annealing algorithm is the objective function in Section 5.1, whose answer is given by initial packet coding algorithm. Let f_i be the current local optimum of simulated annealing algorithm and let f_{i-1} be the previous one. The temperature increment is $\Delta t = f_{i-1}(x) - f_i(x)$. If $\Delta t \le 0$, accept f_i as the final result; otherwise, calculate $p = \exp(-\Delta t/T)$ and accept the result with the probability of p. If $T \le T'$, finish the algorithm while cooling the result with $T = rT$ for next round of iteration. The whole procedure that simulated annealing algorithm takes is illustrated in Algorithm 1.

The challenge in this procedure is determining the values of T, T', and r, which control the speed of this algorithm. The result we get at last will be far away from the optimal solution if the speed is too fast and will be a waste of resource on the contrary. Generally speaking, it is certain for r (e.g., 0.95) and alternative for T and T' which are discussed in detail in Section 5.4.

But what is the answer to $f(x)$? With the absence of $f(x)$, it is impossible to get the optimal solution. As mentioned before, we use coloring algorithm to design this greedy strategy. Typically, we choose the degree of vertexes considering the special features of network coding.

```
(1) FinalResult = f_0(x);
(2) while (T > T') do
(3)     Δt = FinalResult − f_i(x);
(4)     if Δt ≤ 0 then
(5)         FinalResult = f_i(x);
(6)     else
(7)         if random(0, 1) < exp(−Δt/T) then
(8)             FinalResult = f_i(x);
(9)         end if
(10)    end if
(11)    T = rT;
(12)    i + +;
(13) end while
```

Algorithm 1: Simulated annealing procedure.

Let $d_G(V_i)$ be the degree of vertex i in the undirected graph G. The set of neighborhood of i is s_i; that is, $s_i = \{v_j \mid e_{ij} \in E\}$. W_{ic} represents the set of neighbor of i that is colored in color c; that is, $W_{ic} = \{v_j \mid v_j \in S_i, \ x_{jc} = 1, \ c \in C\}$. The sum of the degree of i's neighbor is $d'_G(v_i)$; that is, $d'_G(v_i) = \sum d_G(v_j)$, $v_j \in S_i$.

Nodes execute the procedure as follows.

Step 1 (paint all the nodes into silent and put all of them in a temporary set B). Color the current node (the running node, i.e., *this* in Algorithm 2) and search for two nonadjacent neighbors of *this* and color them into receiving status. If these two neighbors do not exist, color the node that owns largest d'_G among neighbors of *this* into receiving status.

Step 2 (find sending nodes as many as possible). The node v_i cannot be painted to sending unless all the neighbors of v_i are silent. After that, find the node v_j with largest d'_G, and color v_i to receiving status. Then exclude the nodes that can only unicast with v_i. Find neighbors v_k of v_i which meet the following requirements: (a) v_k is in silent, (b) v_k is not adjacent to v_j, and (c) v_k has only one sending node v_i in its neighbor set. Receiving node v_k has the largest d'_G. If v_k does not exist, v_i and v_j can only unicast.

Step 3 (increase the broadcast amount of sending nodes). Clean temporary B and put all the sending nodes in it. Scan all the nodes in B; if the neighbor of node v_i has only one receiving node or the silent nodes are not more than 2, remove v_i from B; otherwise, find two neighbors v_j and v_k of v_i which meet the following requirements: (a) they are not adjacent to each other and (b) they both have and only have one sending node (i.e., v_i). If v_j and v_k do not exist, remove v_i; otherwise, color both nodes (v_j and v_k) into receiving status. End Step 3 until the set B is empty.

5.3. Computational Complexity Analysis. The key to network coding is the scheme for different purposes. However, some schemes are very sophisticated, as [13–16] are NP-complete. COPE has a much lower complexity, that is, $O(n^2)$. The

(1) $B = V$; $i = this$; $x_{i1} = 1$; $x_{i3} = 0$;
(2) **if** $\exists v_j, v_k\ (v_j v_k \in S_i, v_j \notin S_k)$ **then**
(3) $\quad x_{j2} = 1$; $x_{j3} = 0$; $x_{k2} = 1$; $x_{k3} = 0$;
(4) **else**
(5) $\quad x_{j2} = 1$; $x_{j3} = 0$; $(d'_G(v_i) = \max(d'_G(S_i)))$;
(6) **end if**
(7) **while** $B \neq \emptyset$ **do**
(8) $\quad$ Pick v_i from B randomly;
(9) $\quad$ **if** $sizeof(W_{i1} + W_{i2}) \neq 0 \parallel sizeof(W_{i3}) = 0 \parallel sizeof(S_i) = 0$ **then**
(10) $\quad\quad B = B - \{V_i\}$; **continue**
(11) $\quad$ **end if**
(12) $\quad$ **if** $\exists v_j\ (v_j \in S_i, d'_G(v_j) = \max(d'_G(S_i)), W_{j1} = 0)$ **then**
(13) $\quad\quad x_{i1} = 1$; $x_{i3} = 0$; $x_{j1} = 0$; $x_{j3} = 0$;
(14) $\quad\quad$ **if** $\exists v_k\ (v_k \in S_i, v_k \notin S_j, x_{k3} = 1, d'_G(v_k) = \max(d'_G(\{v_{kn}\})), w_{k1} = 1)$ **then**
(15) $\quad\quad\quad x_{k2} = 1$; $x_{k3} = 0$
(16) $\quad\quad$ **end if**
(17) $\quad$ **end if**
(18) **end while**
(19) $B = \{v_i \mid v_{i1} = 1\}$
(20) **while** $B \neq \emptyset$ **do**
(21) $\quad$ Pick v_i from B where $d_G(v_j) = \max(d_G(B))$
(22) $\quad$ **if** $sizeof(W_{i2}) = 1 \parallel sizeof(W_{i3}) < 2$ **then**
(23) $\quad\quad B = B - \{v_i\}$; **continue**;
(24) $\quad$ **end if**
(25) $\quad$ **if** $\exists v_j, v_k\ (v_j \notin S_k, v_j \in S_p, v_k \in S_p, v_p \in W_{i2}, x_{j_3} = 1, x_{k3} = 1)$ **then**
(26) $\quad\quad x_{j_2} = 1$; $x_{j_3} = 0$; $x_{k2} = 1$; $x_{k3} = 0$
(27) $\quad$ **else**
(28) $\quad\quad B = B - \{v_i\}$
(29) $\quad$ **end if**
(30) **end while**
(31) **return** $\sum x_{i1} - \sum x_{i3}/sizeof(V)$

Algorithm 2: Initial packet coding coloring.

complexity of our scheme, on the contrary, is much lower with $O(2n)$, which is more practical.

5.4. T and T'. We repeat the large scale simulation with different coloring decision, that is, unicast coloring, coloring with COPE, and coloring with ECS. All of the coloring decisions follow the same algorithm framework, and it is the solution to $f(x)$ that varies. Note that each node can only communicate with at most one neighbor once in unicast coloring and it can multicast to as many as reachable neighbors with greedy algorithm in COPE. In ECS, nodes make color decisions with the algorithm proposed in Section 5. The flows again arrive according to a Poisson process and pick sender and receiver randomly with a radio range of 175 m and the density of 1/194 m^2.

The number of nodes involved in communication shown in Figure 5 is much larger than unicast, revealing the advantages of network coding. However, the sending coverage of unicast is wider than COPE due to the MAC protocol based on CSMA/CA. ECS holds a more wider coverage than COPE and a slightly smaller number of communication nodes than COPE, which is obviously bigger than unicast.

Figure 5 plots the summary results in a bar graph. The sending ratio (the number of sending nodes/number of nodes) of unicast is beyond 13%. However, the utilization ratio ((sending nodes + receiving nodes)/nodes) is low, around 26%. COPE provides 50% improvements in utilization ratio. However, the sending ratio is lower than 9%. ECS holds approximate sending ratio with COPE, while maintaining a high sending coverage ratio. From the perspective of *this*, COPE constrains the encoding number: it broadcasts an encoding packet of four. ECS lowers this number to 2, avoiding affecting other nodes.

The next question is to determine T and T'. From the definition of the model, we can draw a hidden condition:

$$\sum_{i=1}^{n} \frac{x_{ip}}{n} + \sum_{i=1}^{n} \frac{x_{iq}}{n} + \sum_{i=1}^{n} \frac{x_{ir}}{n} = 1, \quad (p = 1,\ q = 2,\ r = 3). \tag{2}$$

Receiving nodes as the target nodes of communication apparently has a correlation with sending nodes in number. Temporarily let

$$\sum_{i=1}^{n} x_{iq} = c \sum_{i=1}^{n} x_{ip}, \quad (p = 1,\ q = 2), \tag{3}$$

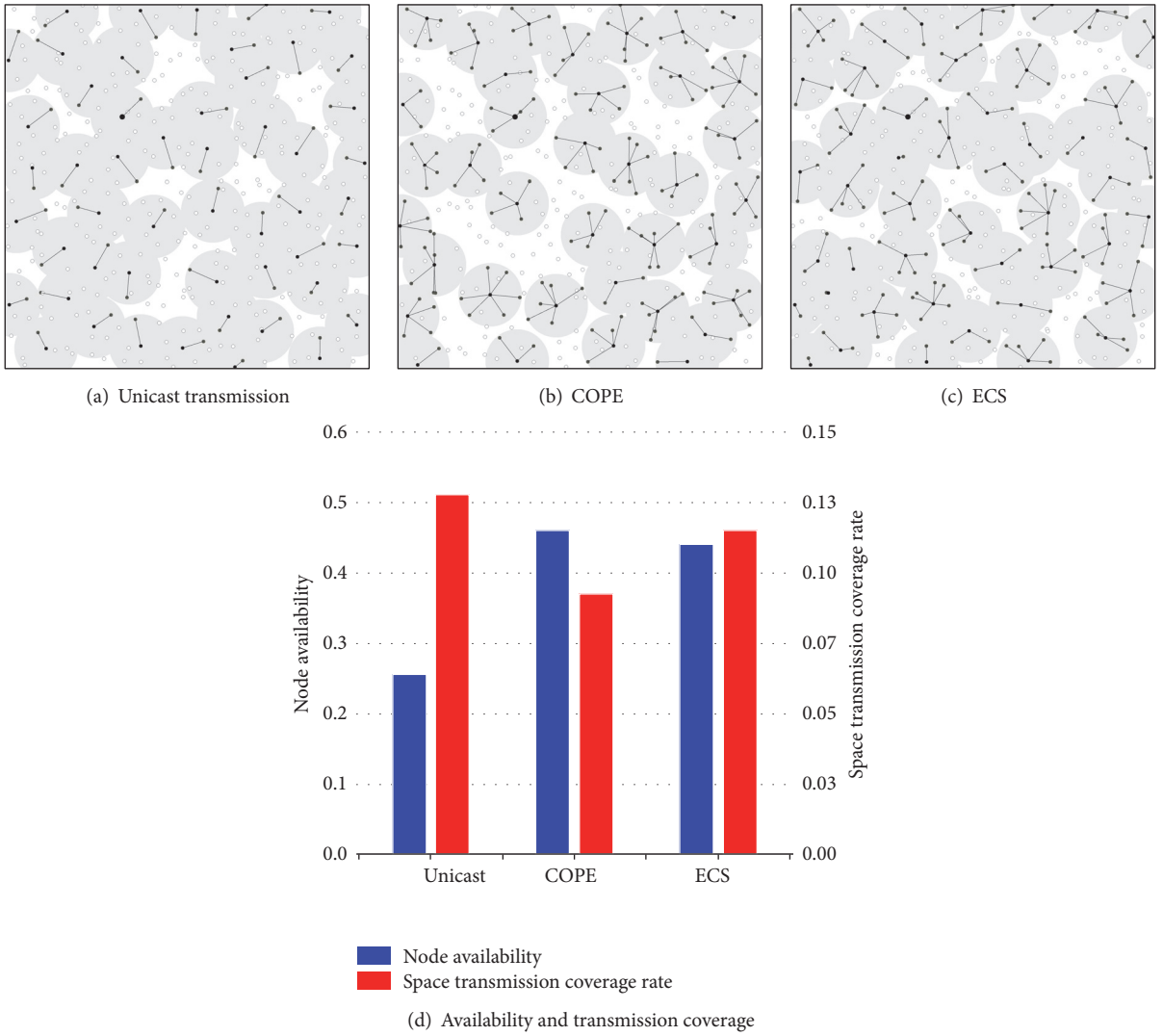


FIGURE 5: *The result of coloring*. Sending nodes are depicted in black, and receiving nodes are in light gray. White nodes are for nodes in NAV timer and the sending coverage is in light gray regions. The bigger black node is the algorithm execution node *this*.

where c is the constant coefficient affected by the density of deployment and $c > 1$, so (3) can be transformed into

$$(1+c)\sum_{i=1}^{n}\frac{x_{ip}}{n}+\sum_{i=1}^{n}\frac{x_{iq}}{n}=1,\quad (p=1,\ q=3). \tag{4}$$

Substituting (4) into the objective function, we can get

$$\max\quad f(x)=\frac{1}{1+c}-\frac{2+c}{n(1+c)}\sum_{i=1}^{n}x_{iq},\quad (q=3) \tag{5}$$

due to

$$\frac{2+c}{n(1+c)}\sum_{i=1}^{n}x_{iq}>0 \tag{6}$$

which is fixed, so the objective function is equivalent to

$$\min\quad g(x)=\sum_{i=1}^{n}\frac{x_{iq}}{n},\quad (q=3). \tag{7}$$

$G(x)$ and $f(x)$ are in the same order of magnitude, and both of them are influenced by T and T'. We simplify the objective function in this way.

We take $T \in [0.0.1, 1]$ of jump 0.01, $T' \in [0.0001, 0.01]$ of jump 0.0001, and 10000 groups of coloring simulations in total. For each T and T', we make color decisions for 100 times repeatedly and take the mean of coloring results $g(x)$ as a final result of this group simulation. Meanwhile, we count the time (unit: ms) of coloring 100 times as the total running time for this group simulation. Figure 6 plots simulation results. The effects of equivalent conditions $g(x)$ for T and T' are depicted in Figure 6(a), and the effect of T and T' impact on coloring duration is illuminated in Figure 6(b). Figures 6(c) and 6(d) plot the equivalent distribution figure of the two effects.

From Figures 6(c) and 6(d), $T = 1$ is a big enough value related to $g(x)$, and T' has greater impact on $g(x)$. When $T' < 0.001$, $g(x)$ is nearly close to the minimum. In addition, the common feature of the running time influenced by T

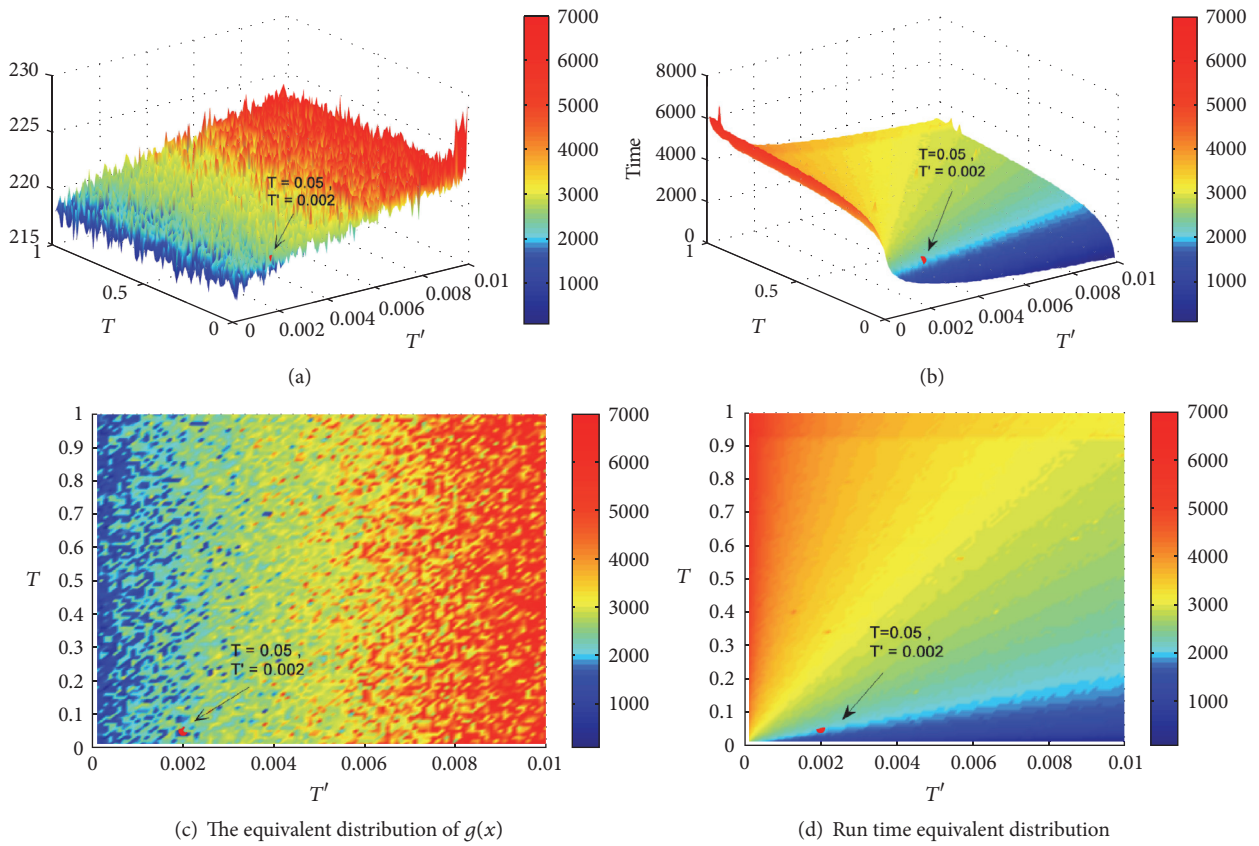


FIGURE 6: *The distribution in various T and T'*. (a) demonstrates how T and T' influence $g(x)$. (b) shows the relationship between these two parameters and the cost of the whole coloring algorithm.

and T' is apparent. Taking $g(x)$ and the running time into consideration, we finally take $T = 0.05$ and $T' = 0.002$ as the cooling speed.

6. Broadcast and Retransmissions

After making decisions about the encoding threshold, we need to reduce retransmissions due to the decoding failure. In this section, we describe how ECS learns neighbor state by altering the broadcast protocol.

Decoding failure is more likely to happen when the relay nodes make wrong guesses with opportunistic listening. In particular, we demonstrate in Appendix that there remains an inherent error ratio when leveraging ETX in opportunistic listening. Consequently, relay nodes have to reencode packets or send the packets one by one, which surely reduce the throughput. Transmitting control information alone is simple and reliable; however, it is too costly. Thus, it is crucial to find an approach to learn neighbors' states in a more reliable and low-cost way.

Our solution is to alter the broadcast protocol, which piggybacks on 802.11 broadcast and benefits from its control frame. Figure 7 shows the format of these three control frames in ECS. To build a neighbor state update protocol, we need to make a few designs. On the sending side, whenever the sender gets an opportunity to send, it adds packet information to the RTS frame. After receiving the RTS frame, the receiver checks the packet information in the received RTS frame and sends the CTS frame which adds its capability of decoding to the sender. When the CTS frame arrives at the sender, the node extracts the decoding capability in the CTS frame. Further procedure depends on whether the receiver can decode the encoded packet. If the receiver cannot decode the packet, senders need to reencode with appropriate packets; that is, the secondary encoding packet is the subset of the first packet. Then, the sender sends the RTS frame to the receiver and transmits the data frame. After successfully decoding the data with the reencoded packet, the receiver acks and ends this handshake. We add the times of handshake in CSMA/CA protocol. Therefore we can discard the nodes that cannot decode at receiver through the handshaking phase to decrease decoding failure.

6.1. Updating Neighbor State. Learning neighbor state is a major issue for network coding, which can be intelligently guessed by ETX metric. However, we prove in Appendix that this guess has inherent error rate. If the guess fails, the coded packet forwarded by transmit node would be undecodable by some next hop. Therefore, the transmit node needs to check the neighbors' ack and retransmit the undecodable packets. Such an approach works; however, it may decrease throughput in the whole network.

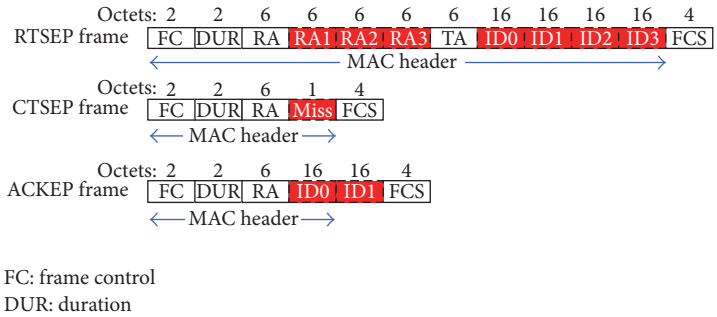


Figure 7: *The format of control frame*. The figure shows the format of control frame in 802.11 protocol and ECS. The field with white background color is what 802.11 protocol has presented. The highlighted fields are the added parts in ECS (e.g., as four encoded packets).

Table 1: Valid type and subtype combinations.

Type value (bit 3 | bit 2)	Type description	Subtype value (bit 7 | bit 4)	Subtype description
01	Control	0001	RTSEP (the set of 2)
01	Control	0010	RTSEP (the set of 4)
01	Control	0011	CTSEP
01	Control	0101	ACKEP

To visually understand the problem, let us consider a simple communication shown in Figure 9. For example, sender wants to communicate a packet to Receiver 1 and Receiver 2. And all their neighbors should stay in NAV state to avoid interference. If Receiver 2 cannot decode the packet, its neighbors (except the senders) may waste a single NAV cycle. We therefore design a protocol that alters the control frame in broadcast protocol to avoid such waste.

6.2. Control Frame Detail. In this section, we will introduce how ECS extends the control frame in CSMA/CA protocol to minimize the cost in transmitting data.

The first problem is to uniquely identify a packet in the whole network. Inspired by Message Digest Algorithm (MD5), we use the digest (16 B) of each data frame to be its ID number, which keeps the same data frame format in CSMA/CA protocol. Thus, our system could be compatible with the protocols in other layers. However, to reduce the query delay, nodes need to save the computed ID number of the packets when they are putting this packet into buffer.

As each packet can be identified in the whole network, the next to consider is how we use the packets' ID to update the neighbor state. The set of receivers and the encoded information are demonstrated in the first RTS frame sent by senders; that is, the RTS frame contains field RA and data frame ID. We define this control frame as RTSEP (Request to Send Encoding Packet). Figure 7 shows the RTSEP frame format of broadcasting four encoded packets. The receiver can compute the encoded packets number according to the Type and Subtype fields in the RTSEP frame. We enrich the Type and Subtype fields based on 802.11 protocol, described in Table 1. After the receiver obtains the RTSEP frame, it queries the index of itself in the receiver collection. According to the index, it can ensure the slot of sending CTS frame. The receiver should check the packets in its buffer based on the IDs in the RTSEP frame. If only one packet could not be found, that is to say, the presending packets can be decoded, then the receiver sends the CTS frame to the sender, which includes the index of those unfounded packets. We call this CTS frame CTSEP (Clear to Send Encoding Package). The format of CTSEP frame is illustrated in Figure 7. The Miss field is 1 byte in length destined for the result of querying ID number. Moreover, when neighbors of the receiver obtain the CTSEP frame, they check the Miss field in the control frame. If the Miss field has more than one bit which are set to value 1 in data type, the neighbor nodes regress from NAV state.

After the sender receives all the CTSEP frame, it reencodes the packets with the knowledge of CTSEP frame and removes the nodes that cannot decode the encoded packet. For example, Receiver 1 and Receiver 2 can encode the packet, while Receiver 3 and Receiver 4 cannot; that is, Receiver 3 and Receiver 4 are removed from the set of receivers. The sender then resends the RTSEP frame that updates the encoding packets information and reports the change of receiver collection to its neighbor nodes. After waiting for a SIFS slot, the sender sends the encoding data frame. Figure 8 visually shows the above process, and Receiver 3 and Receiver 4 are removed in the second handshake.

6.3. Control Frame Cost Analysis. Now we formally define a control frame cost as c, which is the total length of control frame in one communication. Let c_n be the cost of control frame through transmission by n set of packets: $n = 1, \ldots, k$. In particular, c_1 is the control frame cost of unicast. Then the control frame cost benefit is computed as

$$g_n = \frac{nc_1 - c_n}{nc_1}. \tag{8}$$

The duration of unicast is equal to one RTS frame (20 B per frame) plus one CTS frame (14 B per frame) and one ACK

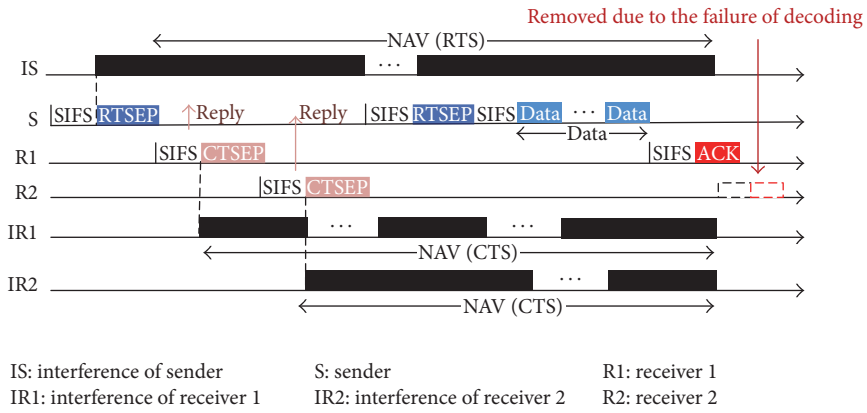


Figure 8: ECS broadcast protocol format.

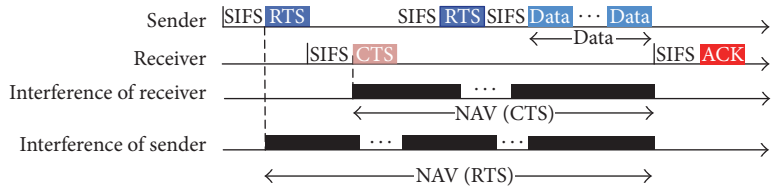


Figure 9: Broadcast protocol process.

frame (14 B per frame) intervals. Thus, the control frame cost of unicast $c_1 = 48$. When the set of packets is 2, the communication needs two RTSEP frames (58 B per frame), two CTSEP frames (15 B per frame), and two ACK frames (14 B per frame). So its control frame cost $c_2 = 174$, and its control frame cost benefit $g_2 = -0.81$. Next we consider the set of 4. The communication demands two RTSEP frames (102 B per frame), two CTSEP frames (15 B per frame), and two ACK frames (14 B per frame). In this case, the control frame cost $c_4 = 320$, and its control frame cost benefit is $g_4 = -0.67$. As we can see, the control frame cost benefit is negative in our Medium Access Control Protocol. This result means that we cost more communication resource in control frame.

However, in the above simple example, we only consider the control frame cost and ignore the total length of data frame. Note that the unicast and broadcast flows are totally different. Unicast flows cannot be encoded together with any other flow. Then, let k be the demand of the flows length and let c' be the cost of the set of packets. And we have

$$c'_n = c_n + k. \tag{9}$$

Similarly, let g' be the cost benefit of the set of packets, and we have

$$g'_n = \frac{nc'_1 - c'_n}{nc'_1}. \tag{10}$$

Take $n = 4$ as an example; (10) turns to k:

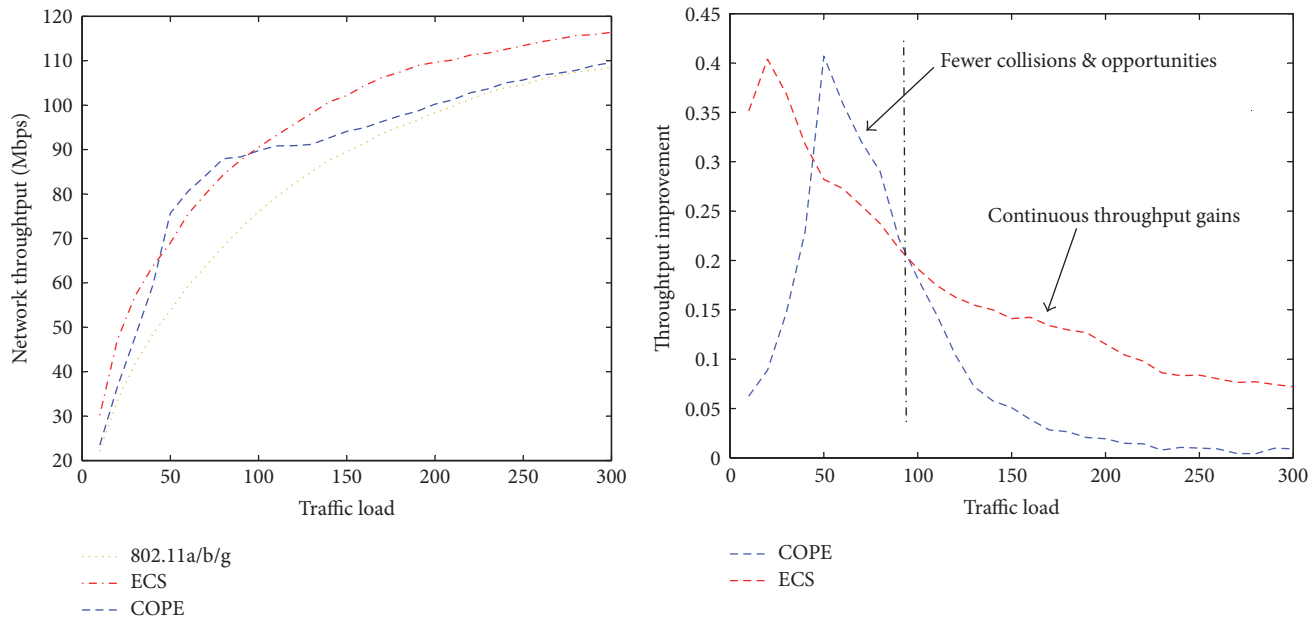
$$g'_4 = \frac{3k - 128}{4k + 192}. \tag{11}$$

When $k > 43$, the cost of the amount of information is positive in a coding scheme, approximating 0.75 finally.

7. Simulation and Result Analysis

Our simulation goes as follows. Flow packets are supposed to arrive according to a Poisson process. $\lambda = 100$ (unit: packets/per second). Senders and receivers are selected randomly using the uniform distribution law. We randomly distribute 86 nodes in a 1400 m × 600 m space, and the communication radius of nodes is 175 m. CSMA/CA protocol is used in MAC layer and minimum-hop routing protocol is used in network layer. The channel capacity is 54 Mbps. And the topology is set up randomly. The expectations of packet length are set to 1000 B, and each node generates ω packets every second. We control network load through ω. That is to say, the rate is fixed to rate = $86\omega/8$ bps. We repeat every group of simulation for 1000 times, lasting for 2 hours, and we use the expectations as the final results. We compare our scheme, ECS, with widely used 802.11a/b/g (i.e., no network coding) and scheme like COPE with greedy algorithm in throughput augment and buffer occupancy.

Note that all the throughput mentioned in this paper is information throughput, a brand new metric to measure the transmission capability of networks. Usually, we employ network throughput to do the job, which is defined as the amount of packets that the total nodes received correctly in a particular period of time. But this classical metric fails to reflect the advantages of network coding schemes (especially COPE). One of the reasons is that nodes regard the packets that cannot be decoded as incorrect packets, but most of these packets have to be delivered correctly. Our metric, information throughput, is defined as the amount of information content that the total nodes sent in a particular period of time. It is apparent that the two metrics are exactly the



(a) Network throughput

(b) Throughput improvement

Figure 10: Throughput and its improvement comparison with different network loads.

same when nodes communicate with each other only by unicast, regardless of delivery probability. Furthermore, the new metric can represent the tradeoff between bandwidth and delivery probability.

7.1. Network Load. The network load is driven by ω ($\omega \in \{10, 20, 30, 40, 50, 100, 150, 200, 250, 300\}$) with 802.11a/b/g, COPE scheme, and ECS scheme. As load increases, information throughput of three schemes increases in general. When $\omega \in (50, 100)$, the interference is scarce; thus CSMA/CA protocol can provide COPE with high-quality service. Consequently, the information throughput in this period is slightly higher than ECS, and 802.11a/b/g provides lowest throughput. As load increases, ECS maintains higher throughput gains than COPE, whose throughput gains level off, reflecting the fact that COPE cannot handle this high-level interference and collision with CSMA/CA. When $\omega >$ 200, interference surges, and since too many packets are encoded together, the performance of COPE deteriorates, approximating 802.11a/b/g. Figure 10 shows the throughput gains between ECS and COPE. Both ECS and COPE do have a suitable load to get maximum throughput gains. However, we try to design a scheme with correspondingly stable throughput gains. The above results reveal that ECS provides more than 7% improvements to 802.11a/b/g, the peak of which is 40%.

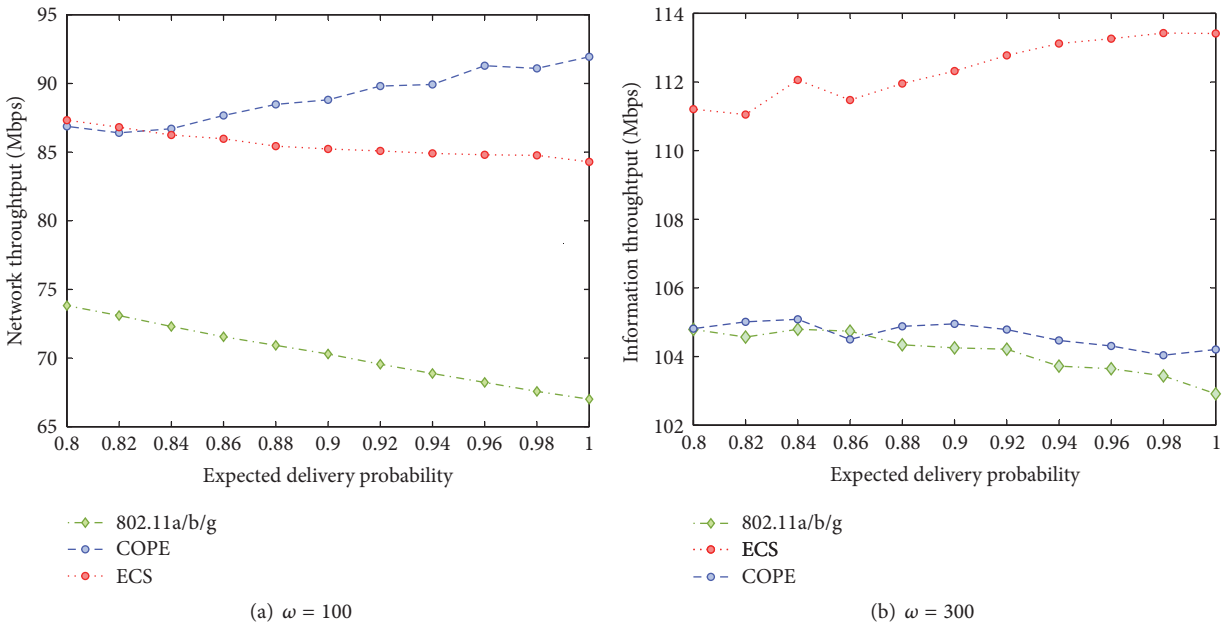
7.2. Link Quality. We repeat the simulation with link quality from 1 to 0.8 with the Gaussian distribution and $\omega = 100, 300$ (unit: packets per second). The flows again arrive according to a Poisson process under the assumption that network load has a slight impact on link quality.

Figure 11(a) shows that both ECS and COPE greatly improve information throughput with low load. Interestingly, with expected delivery probability decreasing, the information throughput decreases, varying from ECS and 802.11a/b/g. Note that the incorrect guess increases as expected delivery probability goes down. Receivers send request for retransmission as they cannot decode. Receivers, however, back off because they may regard this request as an unexpected interference.

Figure 11(b) plots the information throughput in high load, which is greatly shifted. ECS offers a greater improvement than COPE and 802.11a/b/g. Surprisingly, COPE has few improvements as it is close to 802.11a/b/g. Recall from Section 7.1 that the information throughput of COPE becomes flat due to the incorrect neighbor state guesses and encoding number. As link quality improves, the information throughput decreases, showing that the network has achieved its maximum capacity.

The above results reveal that COPE could not bring whole network information throughput in heavy load. It is because COPE substantially increases the utilization of nodes but decreases the sending coverage at the same time. ECS, on the contrary, makes a tradeoff between sending coverage and the utilization of nodes, which therefore yields higher throughput. Meanwhile, the model of learning neighbor state in ECS avoids incorrect decisions; thus retransmissions and long-term backoff greatly decrease. It should be noted that throughput in ECS also has a slight decrease as interference is sure to rise when multicasting encoding packets.

7.3. Buffer Occupancy. Many previous coding schemes assume that nodes' buffer is infinite; that is to say, infinite packets can be buffered. In practice, however, it is vital to



(a) $\omega = 100$ (b) $\omega = 300$

FIGURE 11: Throughput comparison with different link quality.

consider that buffer occupancy is limited. Compared with 802.11a/b/g, nodes maintain not only a forwarding queue but also a decoding resource queue. Data packets in the decoding resource queue may be used in decoding; otherwise, they would be discarded. As mentioned earlier in Section 2, there are two kinds of decoding resources: one is the data packets sent by nodes themselves and another one is the data packets overheard by opportunistic listening. The proportion of these two kinds of data packets is associated with node deployment location and topology. That is to say, for the nodes close to the edge of network, the queue of overhearing packets is small, and for the nodes on the trunk road of network, the vast majority of decoding resources are overheard packets. Figure 12 describes the buffer utilization of all nodes in COPE in the simulation, the horizontal axis is the node numbers, and the vertical axis is the number of data packets buffered by node when network is stable; broken line represents the number of data packets in the forwarding queue and the column represents the total amount of packets in the buffer. It can be seen from the figure that there is big difference between nodes; part of the nodes almost have no buffer decoding resources, and part of the nodes have more than 90% of the buffer to store decoding resources.

7.3.1. Buffer Occupancy and Jitter. We repeat the simulation with the load of 100 and 300, respectively, to summarize some feature values of buffer occupancy, including start values, end values, maximum values, and minimum values, as well as average values. We randomly pick 12 nodes and draw the K chart as shown in Figure 13. Buffer occupancy on ECS is much lower than COPE, since ECS releases invalid decoding resource. For higher load, the buffer occupancy of COPE doubles and even triples compared with medium load. Comparatively, ECS does not require much extra buffer. Also, the curve of ECS is much smoother than COPE, reflecting that ECS only jitters slightly and therefore is more stable and robust.

7.3.2. Cost of Throughput Gains. We repeat the simulation with load in low, medium, and heavy weight ($\omega \in \{60, 100, 200, 300\}$). The relationship between buffer occupancy and throughput gains is presented in Figure 14. Network throughput can be improved by wider frequency band with MIMO, frequency gains with advanced hardware, or our solution, network coding, which all increase cost. Distinguished from other approaches, network coding needs more computational time and storage. As we have shown in Section 5.3, ECS yields much lower computational complexity than COPE and separates the algorithm into two parts, resulting in running apart in spare time of nodes.

Almost every node exploiting ECS in Figure 14 lies in the left side, mirroring that with same throughput gains; ECS occupies fewer buffers. In addition, nodes of ECS gather together with rare variations, while nodes of COPE are more separated, indicating that ECS is more adaptive and robust. Surprisingly, some nodes (both in ECS and COPE) are distributed below zero. That is to say, not all the nodes can obtain throughput gains.

7.4. Deployment Density. It is significantly important to consider deployment density in wireless mesh network. For example, with intensive deployment, network is robust, but it also brings high redundancy and interference. Deploying loosely, network is in poor robustness, while the utilization of nodes is high. Thus, we evaluate the effect of deployment density on network coding scheme.

We evenly deploy n nodes in every rr area of one node randomly; therefore network coverage area is $r\sqrt{n} \cdot r\sqrt{n}$. We

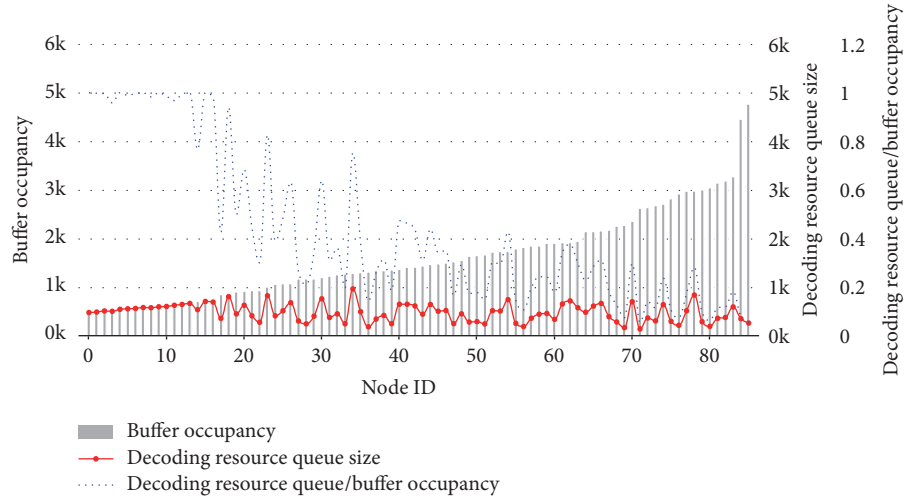
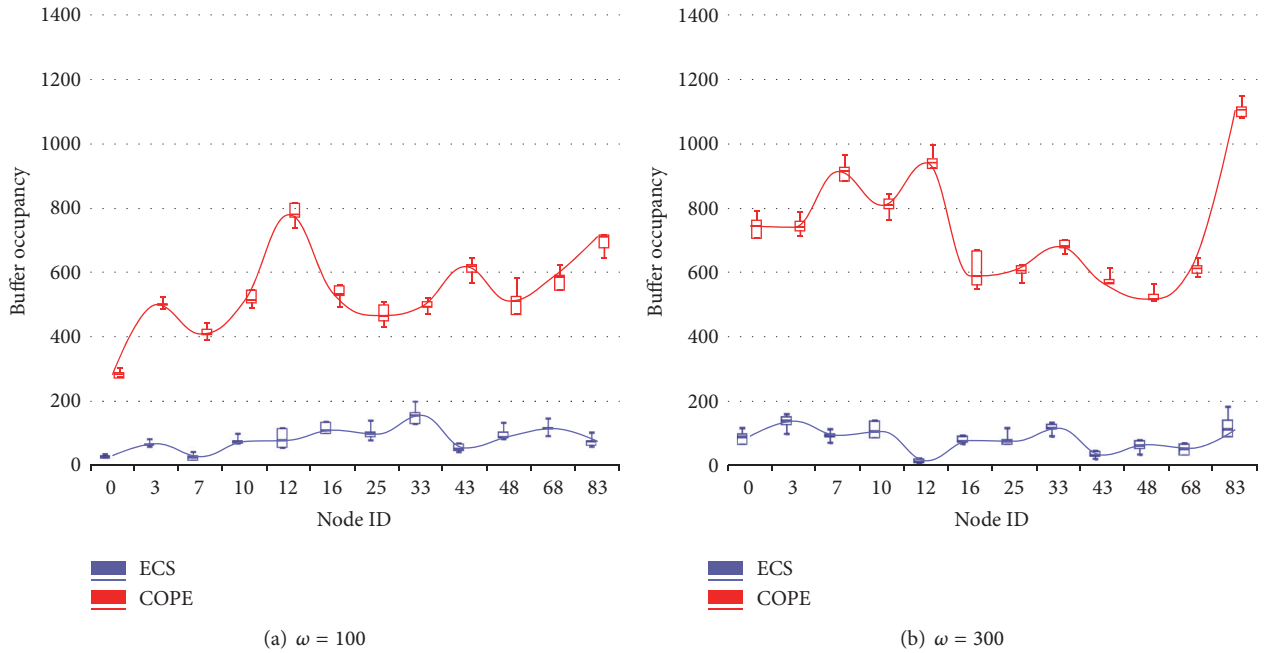


Figure 12: Buffer occupancy in COPE.



(a) $\omega = 100$ (b) $\omega = 300$

Figure 13: Buffer consumption and buffer jitter.

control deployment density by adjusting the parameter r. In order to ensure that the network is a connected graph, r should be less than $\sqrt{2}R/2$ (R is node communication radius). Meanwhile, in order to ensure that the network is totally disconnected graph, r should be greater than $\sqrt{2/n}R/2$. In the simulation, we set $n = 400$, $R = 175$, and $r \in [50, 125]$. Figures 15(a)–15(c), respectively, show the network topology when the particle size is 50, 90, and 120.

Under 16 kinds of deployment density, we use coloring algorithm mentioned in Section 5, respectively, on 802.11a/b/g, COPE, and ECS. Figures 16(a)–16(i) are the results of coloring simulation.

As we can see from Figure 16, no matter whether the deployment density is intensive or not, sending coverage rate of both ECS and 802.11a/b/g is very high, while COPE is relatively low. Leveraging network coding can bring high ratio of utilization of nodes, which is mainly because every sending node conducts multicast communication with many neighbors, just as shown in Figure 16(d).

Overall, the main approach in COPE for throughput increment is to improve node utilization, which, however, consumes too much sending coverage rate. The sending coverage ratio dries up when nodes are deployed intensively in COPE. ECS, by contrast, still keeps high sending coverage

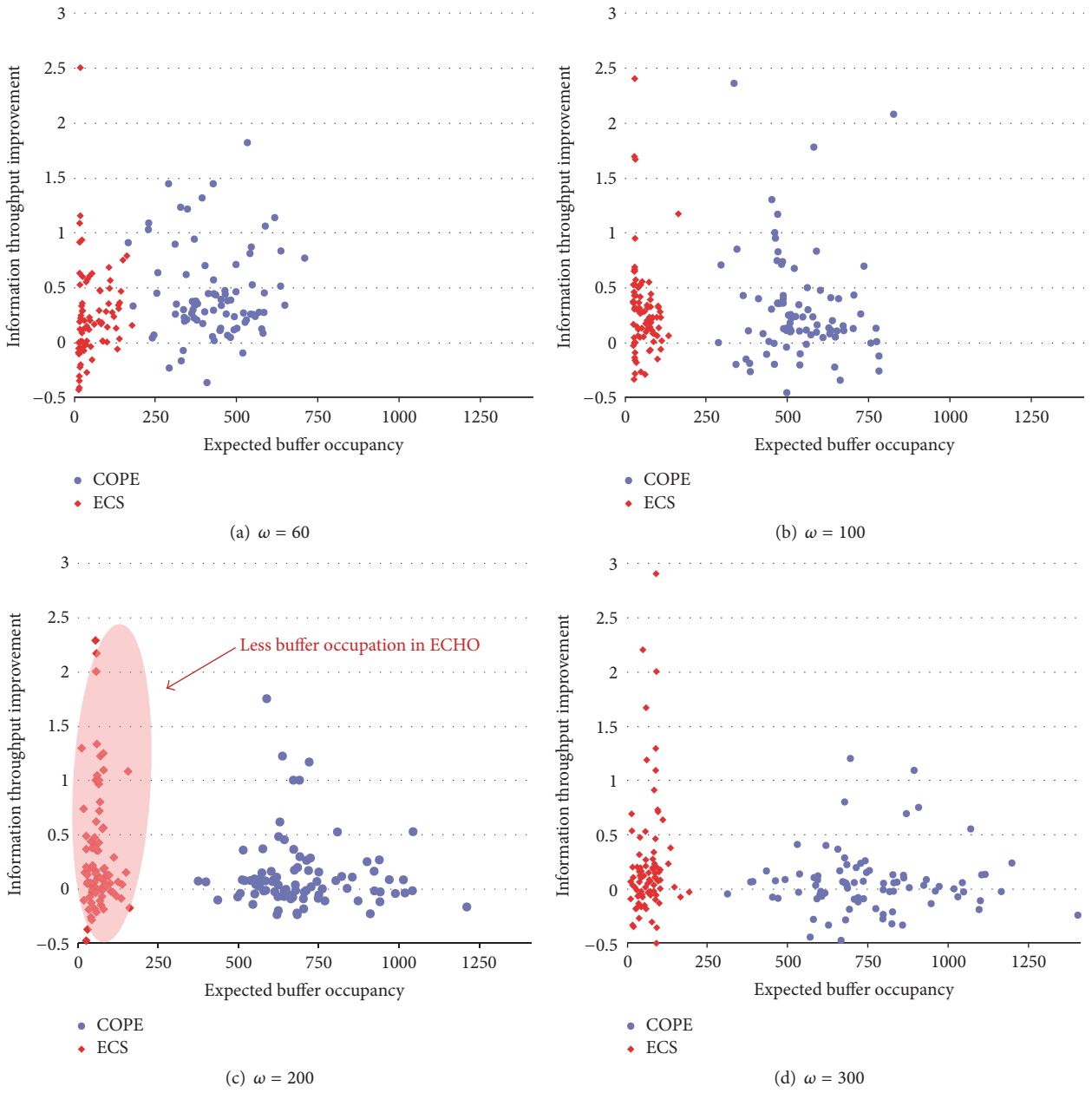


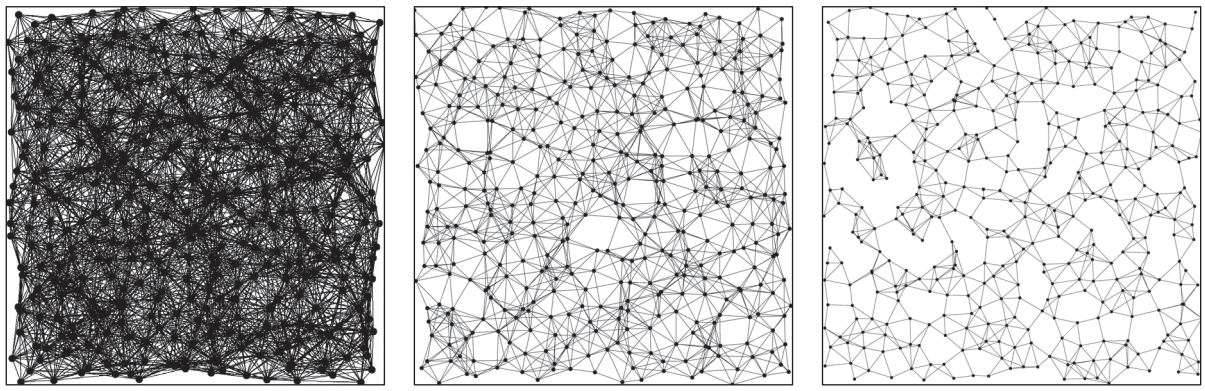
Figure 14: The overhead of throughput improvement.

ratio when nodes are deployed in high density, while the node utilization is almost the same as COPE. When the deployment density is very sparse, COPE and ECS coding are close to 802.11, because nodes barely have enough neighbors at that time; thus the coding opportunities are scarce. Figure 17 plots sending coverage ratio and node utilization under different deployment density.

No matter what deployment density is, it is apparent that ECS has similar sending coverage to 802.11a/b/g in Figures 17(a) and 17(b). At the same time, the node utilization in ECS is close to COPE. Although COPE has higher node utilization, the sending coverage ratio is much lower than ECS. Above all, there are more opportunities for network coding when the deployment is intensive; however, the interference can be very severe. But the advantage of network coding can barely take effect if the deployment is too scarce. Thus, it would be of great value in future study to find a tradeoff of encoding opportunities and deployment density.

8. Related Work

Due to the potential throughput benefit of network coding, it arose much attention when it was first proposed in [3]. Practical network coding schemes are aiming to increase throughput; what is vital is to design encoding algorithms and make decisions. For COPE [10], the first paper that puts



(a) $r = 50$ (b) $r = 90$ (c) $r = 120$

Figure 15: Topology.

network coding into practice has received much attention. Many follow-up researches [4, 6, 11, 17–19] are therefore addressing novel schemes to improve performance of COPE; some examples include the robust coding technique that covers the cases where COPE is oblivious of [19] and the novel scheme that allows relays forward coded packets [11]. However, the majority of them use greedy algorithm, which only focuses on the maximum packets in an encoding procedure. In fact, exploiting all opportunities of broadcasting encoded packets with greedy algorithm in a single transmission may reduce network throughput [6]. Encoding number [4] and optimality scheduling [6] aim to solve these problems, which are also the most closely related schemes to ours. However, these schemes apply for only given topologies, which is unlikely to happen in practice. In fact, it is noticeable that topologies change in wireless mesh networks is usual. Our scheme, calculating the maximum encoding number with the knowledge of topology and enlarging the sending coverage, is therefore more adaptive in mesh networks with large swaths of nodes. In particular, although [4] is also related to the upper bound of network coding, it only gives upper bound of COPE in some given topologies instead of coming up with a new approach to overcome the drawbacks of COPE. In [6], the schedule would be quite complicated when the network scales up. Every node in XOR-Sym allocates every session with a unique queue to store packets needed to be transmitted, which is very costly to maintain. Although it can achieve optimal results, it is quite complicated for scheduling algorithm to decide which sessions are allowed to transmit at a certain time. This would be aggravated when more nodes join the network. Moreover, topologies are more likely to be dynamic and arbitrary in wireless network. Therefore, fixed routing (the assumption of XOR-Sym) may not be guaranteed.

Another body of work has been looking for the improvements of network coding and decoding part based on protocols [8, 20, 21], all of which leverage opportunistic listening. Among these approaches, there are two main approaches to obtain decoding resource knowledge, namely, what native packets the neighbors should buffer to decode encoding packets: (i) transmit control information alone and (ii) via opportunistic listening. The first approach is simple and reliable. However, sacrificing throughput to transmit control information is costly, since network coding aims to augment throughput. Thus, many researches resort to the second approach, namely, exploiting ETX metric [12], to learn neighbor state, which is corrected by ROC check [22]. Typically, [18] works in lossy wireless networks with error-correcting capabilities. However, we prove in Appendix that it remains an inherent error rate, which leads to an incorrect guess and retransmission, therefore reducing throughput increment. Therefore, we elaborate on our approach, carrying coding/decoding information in RTS and CTS, which reduces communication cost, increases reliability, and consequently ensures the throughput gains.

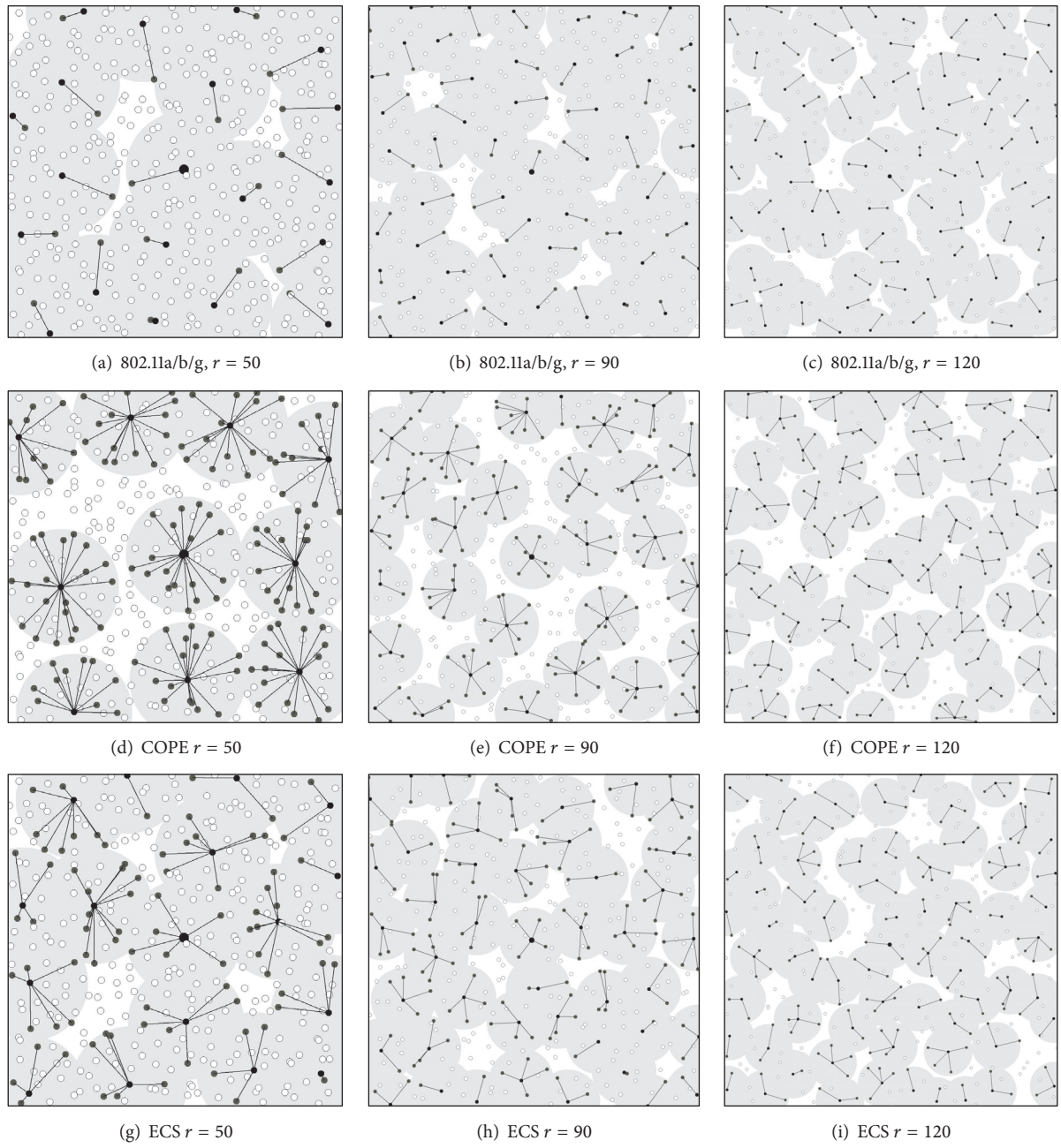
9. Discussion

In this section, we focus on designing an efficient and simple buffer management scheme, so that the similar throughput can be obtained with less buffer occupancy.

In COPE, opportunistically listened packets are stored in the buffer for a fixed time T (default 0.5 s). After a timeout, the packet would be removed. However, such a design is not applicable, since the buffer occupancy is affected by both network load and deployment topologies. The heavier load is, the more packets of neighbors can be opportunistically listened to. Therefore, if the scheme lacks buffer monitoring, the buffer in nodes may surge, which results in no space for forwarding packets. This will be a huge disaster for the network.

The buffer occupancy in summary can be divided into two categories just as we have mentioned in Section 2. We mainly aim to manage the overheard buffer, which is in large amounts in wireless mesh networks.

Our solution, simply speaking, is to efficiently manage decoding resource by releasing the invalid buffer as soon as possible. To better illuminate our idea, let us recall the figure in Figure 2; Alice sends packet p_1 to relay successfully and then moves packet p_1 from forwarding queue to decoding resources queue. At the same time, John and Sally both



(a) 802.11a/b/g, $r = 50$ (b) 802.11a/b/g, $r = 90$ (c) 802.11a/b/g, $r = 120$

(d) COPE $r = 50$ (e) COPE $r = 90$ (f) COPE $r = 120$

(g) ECS $r = 50$ (h) ECS $r = 90$ (i) ECS $r = 120$

Figure 16: The result of deployment density coloring.

receive packet p_1 by overhearing the transmission and store it in the respective decoding resources queue. In COPE, all of three persons remove packet p_1 from the resource queue in 0.5 seconds.

In this paper, relay searches for packet p_1's inverse packet in forwarding queue immediately after receiving packet p_1. If it is found, relay encodes packet p_1 firstly. Whether or not packet p_1 is encoded, relay will send ACK frame to Alice. If packet p_1 is encoded, relay sends ACKEP control frame; otherwise, it sends original ACK frame. Just as Section 6 demonstrates, ID1 stores packet p_1's identification and ID2 stores identification of packet p_1's inverse packet.

If all of them receive or overhear the ACKEP frame sent by relay, they know that packet p_1 has been coded and needs to be preserved. In this situation, packet p_1 is removed only in the following conditions: (i) node decodes successfully using packet p_1 and (ii) receiver gets data packet in ID2 field of ACKEP frame. What if packet p_1 cannot be encoded? For example, one of them announces to relay that he/she has no chance to overhear packet p_1; therefore relay can only unicast packet p_1 and inverse packet of packet p_1.

If Alice, Sally, and John monitor that relay has sent an unmodified ACK frame, they would be sure that packet p_1 is not encoded. In this case, they keep packet p_1 in a θ period

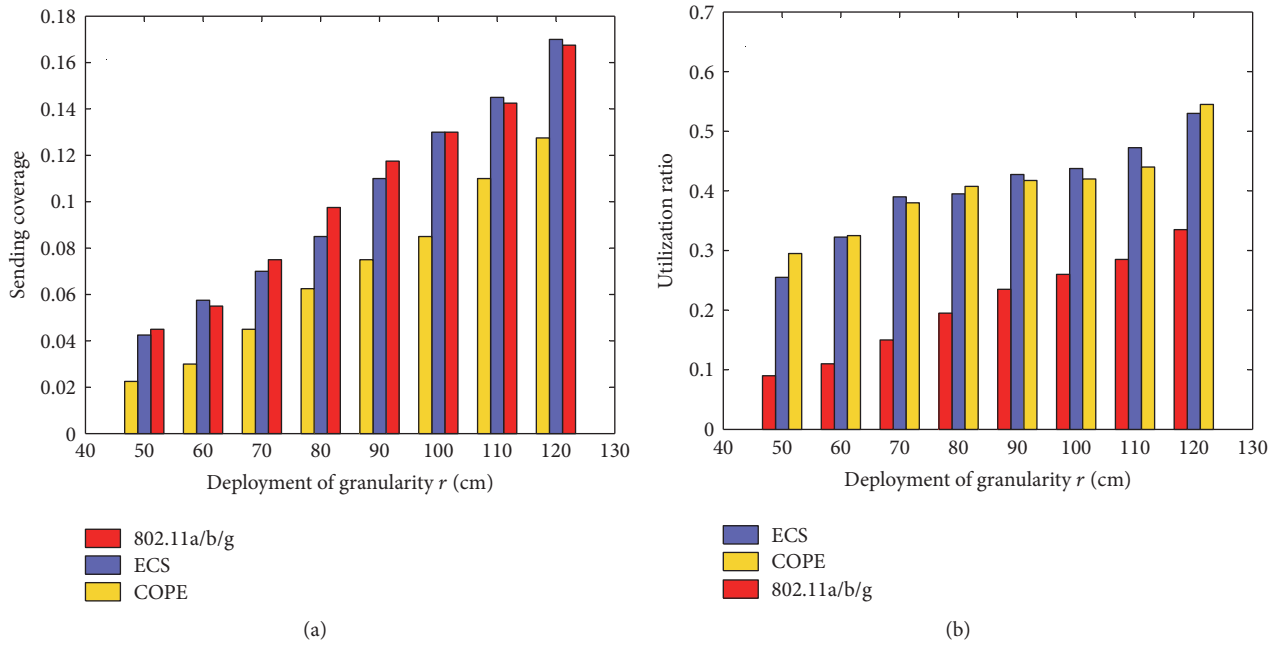


Figure 17: Deployment density with sending coverage.

of time. If they receive other ACKEP frames, whose frame of ID2 domain is packet p_1 in a θ period, it is shown that packet p_1 has been encoded with the packet whose ID number is ID1. Thus, receivers continue to keep packet p_1 until they decode the native packet p_1 or they receive packet p_1's inverse packet. If these nodes do not receive ACKEP that contains packet p_1's ID, packet p_1 would be directly removed from resource, so that packet p_1 is no longer encoded.

10. Conclusion

In this paper, we provide an efficient and low-occupancy interference-awareness model to decrease interference and make full use of nodes in mesh wireless networks. We also prove that the original ETX metric used in opportunistic listening has an inherent error ratio that would lead to decoding failure. Our solution of carrying decoding and encoding information in RTS and CTS is a low-cost way. Our model and findings can be useful for future wireless mesh networks study.

Appendix

Proof of Inherent Error

Let us consider the X-topology. Alice wants to send her packets to Bob and needs the relay to forward the packets. When Alice is sending packets, John overhears the packets. *link—ETX* is the expected number of transmissions. And the delivery ratio means that the transmission is successfully received and acknowledged. The relay nodes can measure the delivery ratio for each link, which is the reciprocal of $link—ETX_i$ in the communication. Assume that $link—ETX_{(aj)}$ from Alice to John is two; thus John has a 50% delivery ratio to overhear packet p_1 successfully. Relays consider whether packet p_1 and packet p_2 are coded together to improve throughout, calculating with $link—ETX_{(aj)}$ and $link—ETX_{(bs)}$. Let r_1 be the delivery ratio of packets from Alice to John and let r_2 be the delivery ratio of packets from Bob to Sally. Assume that there is no relation in the delivery for each link. Then, in COPE, John follows (A.1) to make decision:

$$r_1 r_2 \geq G, \tag{A.1}$$

where constant G is a threshold and the default value is 0.8 in COPE.

It works well when the links are in high quality. However, it remains a primary challenge to work in lossy wireless networks. Based on the ROC curve, r_1 is the delivery ratio of John overhearing packets that relay considers. $1 - r_1$ is the unsuccessful ratio. Then the accuracy of this forecasting is

$$r_1^2 + (1 - r_1)^2 = 2r_1^2 - 2r_1 + 1. \tag{A.2}$$

As we can see in Figure 18, the curve is an upward parabola. When the link is in very high quality or very poor quality, the forecast is highly reliable. But the accuracy in the middle is poor. When r_1 equals 0.5, which reaches its valley, the accuracy is only 0.5. That is to say, if the probability of overhearing packets of John is 0.5, the relay can make a correct guess in John's ratio by half. This is just the delivery ratio of only one packets forecast. To make a coding decision, we need at least two forecasts. In another scenario with wheel topology, we need to overhear for 8 times.

Let S denote the set of opportunistic listening which needs to be forecast, and let m be the size of it. Each $s_i \in S$ has one Alice and one listener Bob. Next, let p_i denote the delivery

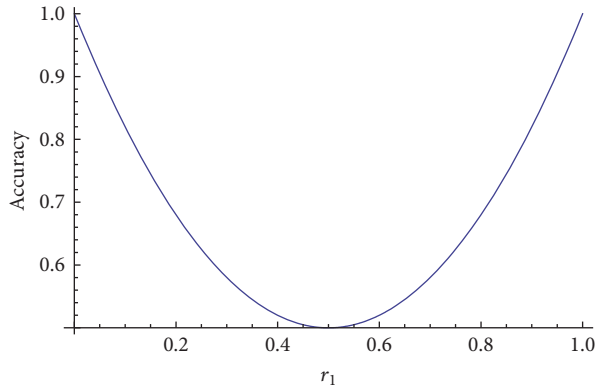


Figure 18: Accuracy of opportunistic listening in COPE.

ratio of link from Alice to Bob. In those network coding schemes such as COPE, if

$$\prod_{i=1}^{m} p_i \geq G, \tag{A.3}$$

the packets contained in S are encoded together, and the incorrect ratio is

$$1 - \frac{\prod_{i=1}^{m} p_i^2 + \sum_{i=1}^{m} \left[(1-p_i)^2 \prod_{j=1}^{i-1} p_j^2\right]}{\prod_{i=1}^{m} p_i}. \tag{A.4}$$

Proof. Consider transmit node forecast opportunistic listening for m times in turn. Let $\overleftarrow{p_i}$ be the correct ratio of forecast from s_1 to s_i. Similarly, $\overrightarrow{p_i}$ defines the correct ratio of forecast from s_{i+1} to s_m. Then, let p denote the correct ratio of all the forecast in S. There are four cases when the decision model examines each packet (see Table 2)

And we have

$$\overleftarrow{p_1} = {p_1}^2 + (1-p_1)^2. \tag{A.5}$$

However, if the first opportunistic listening does not meet the condition, it cannot forecast the following listening correctly. Thus,

$$p = (1-p_1)^2 + {p_1}^2 \overrightarrow{p_1}. \tag{A.6}$$

We define the result as fault if the forecast in S fails more than one time. Hence,

$$\begin{aligned} p &= (1-p_1)^2 + {p_1}^2 \left[(1-p_2)^2 + {p_2}^2 \overrightarrow{p_2}\right], \\ p &= (1-p_1)^2 + {p_1}^2 \left(1 - {p_2}^2\right) \\ &\quad + {p_1}^2 {p_2}^2 \left[(1-p_3)^2 + {p_3}^2 \overrightarrow{p_3}\right]. \end{aligned} \tag{A.7}$$

Obviously, $p = \overleftarrow{p_m}$ and $\overrightarrow{p_m} = 1$; thus

$$p = \prod_{i=1}^{m} p_i^2 + \sum_{i=1}^{m} \left[(1-p_i)^2 \prod_{j=1}^{i-1} p_j^2\right]. \tag{A.8}$$

Table 2

Probability	Reality	Prediction	Decision
${p_i}^2$	Success	Success	Correct
$(1-p_i)^2$	Failure	Failure	Correct
$p_i(1-p_i)$	Failure	Success	Incorrect
$(1-p_i)p_i$	Success	Failure	Incorrect

Therefore, the constant error rate in COPE learning neighbor state is

$$1 - \frac{\prod_{i=1}^{m} p_i^2 + \sum_{i=1}^{m} \left[(1-p_i)^2 \prod_{j=1}^{i-1} p_j^2\right]}{\prod_{i=1}^{m} p_i}. \tag{A.9}$$

□

Disclosure

This work was first published in *ACM MSCC15, 2015,* and the differences between this extended version and the published one are fully explained in Supplementary Material available online at https://doi.org/10.1155/2017/4974165.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was based on Projects 61572402, 61672428, 61672427, and 61170218 supported by NSFC and 2012JQ8049 supported by Natural Science Basic Research Plan in Shaanxi Province of China. They also want to thank three authors (Dr. Chen Liu, Xiaoyan Yin, and Tianzhang Xing) for their contributions in the conference version of this paper. Chen Liu helped them design the algorithm in the MAC layer, which is removed from this version. Xiaoyan Yin and Tianzhang Xing ran the simulations in this part. Although they are excluded from this version because of the removal of the algorithm, they still make great contribution to this paper.

References

[1] A. Belay, G. Prekas, A. Klimovic, S. Grossman, C. Kozyrakis, and E. Bugnion, "Ix: A protected dataplane operating system for high throughput and low latency," in *Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI '14)*, pp. 49–65, Broomfield, Colo, USA, 2014.

[2] S. Sen, W. Lloyd, and M. J. Freedman, "Prophecy: using history for high-throughput fault tolerance," in *Proceedings of the 7th USENIX Conference on Networked Systems Design And Implementation (NSDI '10)*, pp. 345–360, San Jose, Calif, USA, 2010.

[3] R. Ahlswede, N. Cai, S. R. Li, and R. W. Yeung, "Network information flow," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1204–1216, 2000.

[4] J. Le, J. C. S. Lui, and D. M. Chiu, "How many packets can we encode?—An analysis of practical wireless network coding," in *Proceedings of the 27th IEEE Communications Society Conference on Computer Communications (INFOCOM '08)*, pp. 1040–1048, Phoenix, Ariz, USA, April 2008.

[5] J. Liu, D. Goeckelt, and D. Towsley, "Bounds on the gain of network coding and broadcasting in wireless networks," in *Proceedings of the 26th IEEE International Conference on Computer Communications (IEEE INFOCOM '07)*, pp. 724–732, Barcelona, Spain, May 2007.

[6] P. Chaporkar and A. Proutiere, "Adaptive network coding and scheduling for maximizing throughput in wireless networks," in *Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking (MobiCom '07)*, pp. 135–146, Montreal, Canada, September 2007.

[7] T. Ho, M. Medard, R. Koetter et al., "A random linear network coding approach to multicast," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4413–4430, 2006.

[8] S. Katti, D. Katabi, H. Balakrishnan, and M. Medard, "Symbol-level network coding for wireless mesh networks," in *Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication (SIGCOMM '08)*, vol. 38, pp. 401–412, Seattle, Wash, USA, August 2008.

[9] P. Li, S. Guo, S. Yu, and A. V. Vasilakos, "CodePipe: an opportunistic feeding and routing protocol for reliable multicast with pipelined network coding," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM '12)*, pp. 100–108, Orlando, Fla, USA, March 2012.

[10] S. Katti, H. Rahul, W. Hu, D. Katabi, M. Médard, and J. Crowcroft, "XORs in the air: Practical wireless network coding," in *Proceedings of the 2006 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, vol. 36, pp. 243–254, Pisa, Italy, September 2006.

[11] S. Omiwade, R. Zheng, and C. Hua, "Butterflies in the mesh: lightweight localized wireless network coding," in *Proceedings of the 4th Workshop on Network Coding, Theory, and Applications (NetCod '08)*, pp. 1–6, Hong Kong, China, January 2008.

[12] D. S. J. D. Couto, D. Aguayo, J. Bicket, and R. Morris, "A high-throughput path metric for multi-hop wireless routing," *Wireless Networks*, vol. 11, no. 4, pp. 419–434, 2005.

[13] L. Chen, T. Ho, S. H. Low, M. Chiang, and J. C. Doyle, "Optimization based rate control for multicast with network coding," in *Proceedings of the 26th IEEE International Conference on Computer Communications (IEEE INFOCOM '07)*, pp. 1163–1171, Barcelona, Spain, May 2007.

[14] Y. Lin, B. Li, and B. Liang, "CodeOR: opportunistic routing in wireless mesh networks with segmented network coding," in *Proceedings of the 16th IEEE International Conference on Network Protocols, (ICNP '08)*, pp. 13–22, Orlando, Fla, USA, October 2008.

[15] A. Khreishah, C.-C. Wang, and N. B. Shroff, "Cross-layer optimization for wireless multihop networks with pairwise intersession network coding," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 5, pp. 606–621, 2009.

[16] A. Khreishah, I. M. Khalil, and J. Wu, "Distributed network coding-based opportunistic routing for multicast," in *Proceedings of the 13th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc '12)*, pp. 115–124, South Carolina, SC, USA, June 2012.

[17] F. Zhao and M. Médard, "On analyzing and improving COPE performance," in *Proceedings of the Information Theory and Applications Workshop (ITA '10)*, pp. 317–322, San Diego, Calif, USA, February 2010.

[18] H. Seferoglu, A. Markopoulou, and K. K. Ramakrishnan, "I^2NC: intra- and inter-session network coding for unicast flows in wireless networks," in *Proceedings of the IEEE INFOCOM 2011*, pp. 1035–1043, Shanghai, China, April 2011.

[19] Q. Dong, J. Wu, W. Hu, and J. Crowcroft, "Practical network coding in wireless networks," in *Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking (MobiCom '07)*, pp. 306–309, Montréal, Québec, Canada, September 2007.

[20] X. Zhang and B. Li, "Optimized multipath network coding in lossy wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 5, pp. 622–634, 2009.

[21] S. Chachulski, M. Jennings, S. Katti, and D. Katabi, "Trading structure for randomness in wireless opportunistic routing," in *Proceedings of the ACM SIGCOMM Conference on Computer Communications*, vol. 37, pp. 169–180, Kyoto, Japan, August 2007.

[22] J. R. Beck and E. K. Shultz, "The use of relative operating characteristic (ROC) curves in test performance evaluation," *Archives of Pathology and Laboratory Medicine*, vol. 110, no. 1, pp. 13–20, 1986.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 1268515, 11 pages
https://doi.org/10.1155/2017/1268515

Research Article

Dealing with Insufficient Location Fingerprints in Wi-Fi Based Indoor Location Fingerprinting

Kai Dong,[1] Zhen Ling,[1] Xiangyu Xia,[1] Haibo Ye,[2] Wenjia Wu,[1] and Ming Yang[1]

[1]*School of Computer Science and Engineering, Southeast University, Nanjing, China*
[2]*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China*

Correspondence should be addressed to Kai Dong; dk@seu.edu.cn

Received 28 April 2017; Accepted 18 June 2017; Published 9 August 2017

Academic Editor: Zhe Yang

The development of the Internet of Things has accelerated research in the indoor location fingerprinting technique, which provides value-added localization services for existing WLAN infrastructures without the need for any specialized hardware. The deployment of a fingerprinting based localization system requires an extremely large amount of measurements on received signal strength information to generate a location fingerprint database. Nonetheless, this requirement can rarely be satisfied in most indoor environments. In this paper, we target one but common situation when the collected measurements on received signal strength information are insufficient, and show limitations of existing location fingerprinting methods in dealing with inadequate location fingerprints. We also introduce a novel method to reduce noise in measuring the received signal strength based on the maximum likelihood estimation, and compute locations from inadequate location fingerprints by using the stochastic gradient descent algorithm. Our experiment results show that our proposed method can achieve better localization performance even when only a small quantity of RSS measurements is available. Especially when the number of observations at each location is small, our proposed method has evident superiority in localization accuracy.

1. Introduction

With the development of the Internet of Things (IoT) and the popularization of mobile devices such as smart phones, a variety of mobile applications have changed people's lifestyles tremendously. These applications enable users to access a plenty of services at any time in any place and often use their location information in order to provide them with personalized experiences.

The global positioning system (GPS) can achieve meter-level accuracy in outdoor environments. However, GPS works poorly inside buildings due to the signal attenuation caused by roofs, walls, and other objects. During the past decades, a variety of indoor positioning systems (IPS) have been introduced. Since wireless information access is now widely available, many of these approaches tap into wireless signals for estimating locations.

In the last couple of years, the location fingerprinting (LF) technique using existing wireless local area network (WLAN) infrastructure has been suggested for indoor areas. Location fingerprinting estimates the target location by matching online measurements of received signal strength (RSS) with the closest offline features (i.e., the location fingerprints) composed of location coordinates and respective RSS values. It is relatively simple to deploy, compared to the other wireless indoor positioning techniques using Bluetooth beacons [1] or RFID tags [2], which can achieve higher location accuracy.

To deploy a traditional LF-based indoor positioning system, the positioning server should generate the location fingerprints by performing site survey of the RSS information from multiple access points (APs). With these fingerprints, the positioning server is able to localize a mobile device based on its RSS measurements. The site survey is extremely time-consuming and labor-intensive, which raises the cost of initiating an LP-based localization service. Furthermore, the positioning server should periodically reperform this site survey to update the fingerprints so as to control errors in the changeable Wi-Fi environment, which also raises the cost of maintaining the localization service.

Benefiting from cloud computing and big data techniques, the LF system may also be deployed without an offline site survey. Following [3], much work has been done to enable the collection of fingerprints based on crowdsourced solutions. In these solutions, users are required to continuously upload their RSS measurements to the positioning server as the training data. In the meanwhile, an additional incentive mechanism is required to guarantee the number of volunteering users.

In the situation when the crowdsourced fingerprints are insufficient to deploy an LF system, the offline site survey is still required to refine the fingerprint database. Suppose an indoor environment as an example; there are some locations which have never been occupied by any volunteering users. Thus, no RSS fingerprints of these locations are generated. To ensure the functionality of the location service, the service provider may still need to perform an offline site survey at these locations.

Regardless of whether the RSS measurements are collected via a traditional site survey or a crowdsourced approach, it is a widely existing fact that sufficient RSS measurements cannot be collected (periodically for maintaining fingerprints) in most indoor environments. Ways of collecting RSS measurements are not the focus of this paper. We are interested in the following several issues which may be interesting and useful but, however, rarely studied by existing researches:

(i) At a given location, how many RSS measurements are required to generate an accurate online location or offline fingerprint?

(ii) Most importantly, when the collected RSS measurements are insufficient to generate an accurate location fingerprint database, how do we perform localization in this situation?

Although the answers to these questions may vary in different indoor environments, the readers should take into account the instructive significance of deeply analyzing these issues in a certain indoor application scenario. In this paper, we propose a novel localization method which reduces noise in measuring the received signal strength based on the maximum likelihood estimation and estimates locations from inadequate location fingerprints by using the stochastic gradient descent algorithm. We also use an open dataset to evaluate our proposed method by comparing it with the most commonly used location fingerprinting methods and investigate the number of RSS measurements required to deploy an LF system. The results show that our proposed method can achieve better localization accuracy when only a small quantity of RSS measurements is available.

This paper is organized as follows. Section 2 surveys existing location fingerprinting methods. Section 3 introduces two major problems that arise from insufficient RSS measurements in deploying a location fingerprinting system. Section 4 describes our basic idea in solving these problems, and the detailed solution is presented in Section 5. We evaluate our proposed method in Section 6. At last, we conclude this paper in Section 7.

2. Related Work

To reduce the cost of deploying an indoor localization system, many researches leverage existing Wi-Fi infrastructures and introduce location fingerprinting based on the RSS measurements of the Wi-Fi signals. The deployment of location fingerprinting systems is often divided into two phases: an offline phase, in which a site survey of the RSS from multiple APs is collected, and an online phase, in which a location can be computed based on the currently observed RSS measurements by using a matching algorithm.

2.1. Collecting Online and Offline RSS Measurements. At least four key factors can decide the accuracy of an LF technique.

The first is the density of the offline observing locations where RSS measurements are collected to generate fingerprints: a higher accuracy in LF requires higher intensity of the observing locations, which leads to heavier workload in collecting and updating the fingerprints.

The second is the quantity of available information, including the number of RSS observations used to generate fingerprints and the number of dimensions (observed APs) in each RSS observation. Existing approaches use channel state information [4, 5] or environmental information such as light [6], sound [7, 8], temperature, humidity, magnetic, or pressure data to improve location accuracy. Both of these factors deal with the sufficiency of RSS measurements.

2.2. Reducing Noises and Generating Fingerprints. The third key factor deciding the accuracy of localization is the algorithm used to reduce noises. It can be used in both the offline and the online phases.

The most common way in denoising RSS measurements is to observe multiple times at the same location and average multiple observations so that noises can be reduced. With multiple observations at the same location, one can also make sure that all observable APs are observed. The tricky part is how to deal with situations when some APs are missed from some (but not all) observations. A common but also naive approach is to simply set the RSS to unobserved APs to −100 dBm. Some other approaches assume that only APs far away from the observing location can be missed (we will show that this assumption is wrong) and make a threshold (e.g., −80 dBm) to consider only RSS measurements larger than this threshold. There are also approaches that use a complex algorithm to reduce noises [9–11]; however, most of these approaches require a large quantity of RSS observations at the same location. Some approaches use a lightweight machine learning method to generate limited location fingerprints [12], or variations of fingerprints such as RSS differences between every pair of APs [13], or do not need to generate location fingerprints [14]; however, they suffer from relatively low localization accuracy.

By only reducing the measurement errors, it is still difficult to achieve a high localization accuracy. With sufficient RSS measurements, localization accuracy is mainly decided by the fourth factor.

2.3. Matching Algorithm. The fourth key factor deciding the accuracy of localization is the matching algorithm used in the online phase, which outputs the final location by comparing the online RSS observations with the location fingerprints. By now, most LF systems mainly use, but are not limited to, the following types of matching algorithms.

2.3.1. Probabilistic Method. The probabilistic method treats the matching problem as a classical classification problem. It computes the probabilities that the online observing location belongs to every offline candidate location and finally performs matching from the candidate location based on the probabilities. The result of the localization can be either the candidate location with the highest probability or an averaged value calculated from every candidate location weighted by its corresponding probability.

2.3.2. k-Nearest Neighbors. Based on the context information collected at the observing location, k nearest neighbors are defined as the k offline candidate locations which have the most similar context information. The locations of the k-nearest neighbors (KNN) contribute to the result of the localization by direct averaging or weighted averaging in weighted KNN (WKNN). It must be taken into consideration that the context information can be of various kinds (e.g., wireless signal strength, brightness, temperature, and humidity), and the metric quantifying the distance between the vectors of context information should be carefully designed. In the situation where only wireless signal strength is used, the Euclidean distance in the wireless signal strength space is often used as the metric. Locations which have smaller distance with the observing location are the k-nearest neighbors, and the distances can be used to compute the weights in WKNN.

2.3.3. Other Machine Learning Methods. Existing machine learning methods can be used in matching the online location to those offline locations. A neural network can be created in the offline phase, which takes as input the context information collected at the observing online location, takes as object the location of the fingerprint, learns the weight matrix for each dimension of the context information, and finally outputs the localization result. The support vector machine can be used in small sampled, nonlinear, and high dimensional pattern recognition. The matching and localization can be accomplished by treating the location fingerprint information of candidate locations as support vectors and by performing classification and regression analysis on the context information collected at the target observing location. Other machine learning methods may also be used in location fingerprinting.

2.4. Localization without Site Survey. The site survey in the offline phase can be extremely time-consuming and labor-intensive. Recently, many researches introduced crowdsourcing based systems [3, 15–20] which require the users to continuously observe their RSS measurements and upload the data to the positioning server. These approaches do not require the site survey to be performed, and they do not require the map of the floorplan. However, additional incentive mechanisms are required to attract enough participation, since the one who uploads his observed RSS measurements cannot obtain any benefits like positioning accuracy but will definitely take the privacy risk and the transmission cost. In our previous work [21], we propose a novel indoor navigation mechanism for shopping mall environments, which requires only few shop owners as RSS information contributors. Compared with our previous work, this work improves the method by adjusting it to more general indoor location fingerprinting scenarios and also evaluates our proposed method by comparing it with existing location fingerprinting techniques. Furthermore, we do not focus on ways of collecting RSS measurements. We are only interested in the quantity of the RSS measurements, regardless of whether they are collected via a traditional site survey or a crowdsourced approach.

3. Problem Definition

There are so many situations in our real life when we are asking or being asked a question like "How can I go to?" or "Where is?" For instance, a consumer may want to find a certain shop in a shopping mall, or a patient may want to find the correct consulting room in a hospital. Nowadays, most indoor environments like the aforementioned shopping malls or hospitals always have WLAN infrastructures; however, localization in these environments is still unavailable. The key reason deals with the cost in building and maintaining the fingerprint database. Existing techniques highly rely on an assumption that sufficient RSS measurements can be collected, either by a site survey, which is extremely time-consuming and labor-intensive, or by a crowdsourced approach, which requires too many collaborative contributors.

Here is an example showing how much time one should spend in collecting "sufficient" RSS measurements. Consider a very tiny shopping mall with a total area of only 5,000 m^2 that includes all the floors. The offline observations are collected every 1 m^2, and at every observing location, at least 10 observations have to be collected. After each observation, a time interval of, for example, about 3 seconds is spent so as to obtain a next observation. Suppose one spends no time moving from one observing location to another, and the observations can never fail. We can compute that he should spend at least 150,000 seconds (i.e., 41.67 hours) to perform a site survey. If the fingerprint database needs to be updated every day, then we need at least 5 long-term employees, each of whom works for 8 hours a day with no weekend and must not rest during working. Remember that that is only for tiny shopping malls. For large shopping malls, the workload can be incredibly heavy. Perhaps this is the reason why building owners always choose to deploy infrastructures to provide localization, not the "infrastructure-free" location fingerprinting.

So, our problem is, when the RSS measurements collected are not sufficient, how do we perform localization? At least the following two problems should be addressed.

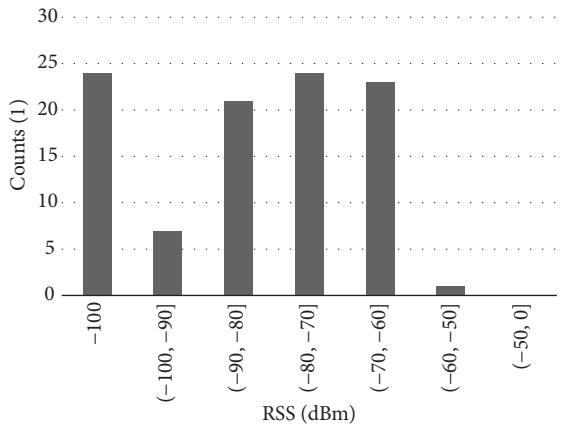


Figure 1: RSS collected in 100 times to the same AP.

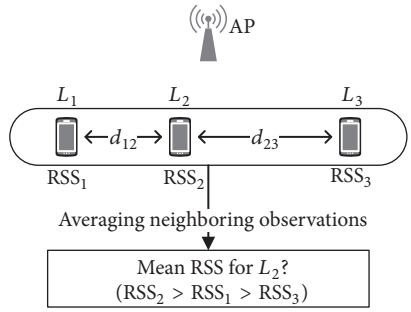


Figure 2: Denoising RSS value by averaging different observations from neighboring locations is not a good idea.

3.1. Measurement Noises. One problem arising from insufficient RSS measurements deals with noises in the RSS measurements. RSS values can change greatly in different observations even at the same location to the same AP, as shown in Figure 1. In this example, the standard deviation is 13.66. Without denoising the RSS measurements, no accurate fingerprints can be generated and no accurate localization can be performed.

One may think of an intuitive solution by averaging different observations at neighboring locations. This idea is not always correct as illustrated in Figure 2. Suppose three observing locations L_1, L_2, and L_3 are in a line, and L_2 lies in between L_1 and L_3. The simple but incorrect solution denoises the observation at L_2 by weighted-averaging the observations at L_1 and L_3, and the weights can be computed from the distances d_{12} and d_{23}. However, this denoising method is not always correct (if not always incorrect), since it relies on a totally wrong hypothesis that the RSS to different locations in a 2D or 3D space can be modeled by a linear function. In Figure 2, suppose the AP is located closer to L_2; we can find that the RSS to this AP observed at L_2 should be larger than those observed at L_1 and L_3. So, the denoising method will definitely reduce the value of RSS_2.

3.2. Missed APs. Another problem deals with dimensional mismatches between different RSS observations in the signal space. When the Wi-Fi scan operations are performed frequently, many APs can be missed in the RSS observations. As a result, even RSS observations from very nearby locations may observe different APs. Since the distance between different RSS observations is computed in a high dimensional signal space where each AP is a dimension, the missed APs will cause dimensional mismatches. If the RSS measurements are not sufficient, dimensional mismatches can always occur. The dimensional mismatches can cause localization failures and errors, and we call this problem the missed AP problem.

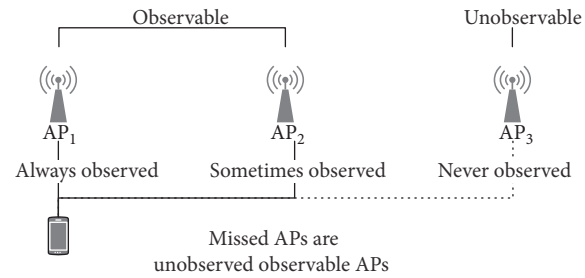


Figure 3: Relationship between (un)observed, (un)observable, and missed APs.

We use an open dataset to show how frequently an AP can be missed in an arbitrary observation. The dataset is the Mannheim/compass dataset [22] which contains Wi-Fi observations of different locations. Our experiments described in Section 6 are also based on this dataset. For a given RSS observation, the APs can be classified into the following three categories as shown in Figure 3:

(i) *Observed APs*. Those are observed in the record. Reversely, the unobserved APs are those not observed in the record.

(ii) *Unobservable APs*. Those cannot be observed at the observing location. An AP is unobservable if no records at this location ever observed this AP. The unobservable APs must be unobserved APs, but unobserved APs may be observable.

(iii) *Missed APs*. Those are observable but unobserved in this record.

The proportions of the observed APs, the unobservable APs, and the missed APs are shown in Figure 4. One interesting finding is that the probability of missing an AP is not obviously related to the averaged RSS value. According to this finding, it is not reasonable to treat RSS to a missed AP (i.e., an unobserved but observable AP) as −100 dBm, since −100 dBm means the AP is unobservable.

4. Basic Idea

In the following, we present how our proposed method deals with the missed APs and the measurement noises.

4.1. Dealing with Missed APs. Consider an indoor environment as illustrated in Figure 5. There are 4 APs denoted as AP_1, AP_2, AP_3, and AP_4 and 4 observing locations denoted

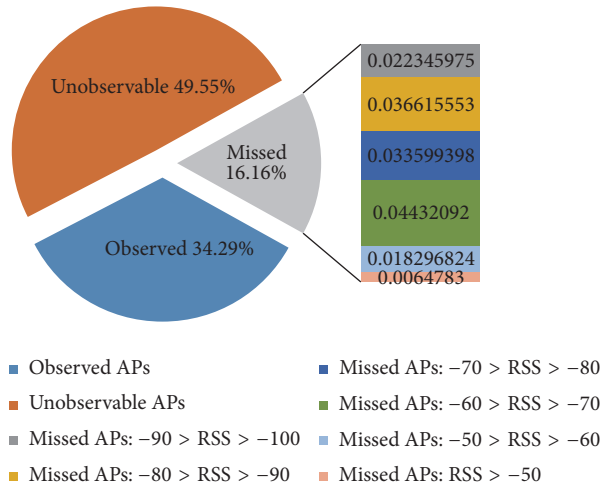


Figure 4: Proportion of the observed APs, the unobservable APs, and the missed APs. Here, "RSS" is not an observed value, but a theoretically computed value by averaging the measurements in which this AP is observed.

as L_1, L_2, L_3, and L_4. At each observing location, we perform an RSS observation; and within each observation, a specific AP is missed. We can further suppose that the location of an arbitrary observation is unknown and needs to be localized. In this situation, we find that it is difficult to perform traditional localization, since these observations observe different APs. And if we discard the dimensions of the missed APs to avoid dimensional mismatches, we will find that no dimensions are left in the signal space. As a result, none of these observations can be used for traditional localization techniques.

Our idea to solve the missed AP problem is straightforward. The observations with missed APs really cannot be directly used for location fingerprinting; however, the information within the observations is valuable, since it tells the relationship between the relative locations of all APs and observations. Back to our example in Figure 5, AP_3 is missed in the observation at L_2. Our idea is to compute RSS_3 for L_2 based on the values of other RSS measurements at other locations. The RSS to all other missed APs can also be computed in a similar way. If we can compute a theoretical RSS value for each of the missed APs in all observations, the dimensional mismatches can be avoided and localization can be performed. The detailed algorithm is presented in Section 5.

4.2. Reducing Measurement Noises. Our idea in denoising the RSS measurements is to some extent related to our solution to the missed AP problem. As shown in Figure 5, after we compute RSS_4 for L_1, RSS_3 for L_2, RSS_2 for L_3, and RSS_1 for L_4, finally we have 4 observations each of which contains RSS measurements to all the 4 APs. Now, the localization can be performed by matching the RSS observation distance in the 4D signal space to the locational distance in the 2D physical space. However, the output of this localization process is far from accurate, since every RSS measurement in every observation is noisy. Without sufficient RSS measurements at the same location, traditional localization techniques cannot reduce the noises effectively.

Again, we make use of all RSS measurements to compute the relationship between the relative locations of all APs and observations. Back to our example in Figure 5, this relationship can be primarily computed based on the primary localization result. With this relationship, RSS_1 for L_3 and RSS_1 for L_4 can be used to modify RSS_1 for L_1, and every other RSS measurement can also be denoised by carefully computing the weights of the measurements to the same AP, no matter at the same location or other locations. Then, the newly denoised RSS measurements can be used to improve the accuracy of the previous localization and thus will output more accurate locations. Our proposed method iterates denoising the RSS measurements and refining the locations until convergence. The detailed algorithm is presented in Section 5.

5. Designing Details

We now introduce our proposed localization method in detail. In the Notations, we summarize the main notations introduced throughout this article.

Suppose there are N APs within an area, denoted as

$$\mathrm{AP} = \{\mathrm{AP}_1, \mathrm{AP}_2, \ldots, \mathrm{AP}_N\}; \tag{1}$$

the RSS information collected at one location can be described as an N-dimensional vector:

$$\mathrm{RSS} = \left\langle \mathrm{RSS}_{\mathrm{AP}_1}, \mathrm{RSS}_{\mathrm{AP}_2}, \ldots, \mathrm{RSS}_{\mathrm{AP}_N} \right\rangle, \tag{2}$$

where each dimension corresponds to the RSS information of an AP. If an AP is not observed, the RSS value is $-\infty$ and is marked as 0.

For situations when the RSS observations are insufficient, we suppose the location fingerprinting algorithm takes as input M offline RSS observations at different locations and one online RSS observation and outputs the online observing location. The RSS information can be collected as a sparse matrix, denoted as

$$S = \left[\mathrm{RSS}_1, \mathrm{RSS}_2, \ldots, \mathrm{RSS}_M, \mathrm{RSS}_{M+1}\right]. \tag{3}$$

We denoise RSS information in this sparse matrix based on the following two assumptions:

(i) The value of RSS (in dBm) follows the Gaussian distribution:

$$\mathrm{RSS} \sim \mathrm{Gaussian}\left(\widetilde{\mathrm{RSS}}, \sigma^2\right). \tag{4}$$

(ii) The signal propagation path loss varies exponentially with distance:

$$\mathrm{PL} = \mathrm{PL}_0 + 10n\log_{10}\left(\frac{d}{d_0}\right) + X_g, \tag{5}$$

where PL_0 is the path loss at unit distance d_0, n is the propagation path loss exponent, and X_g is a Gaussian random variable with 0 mean.

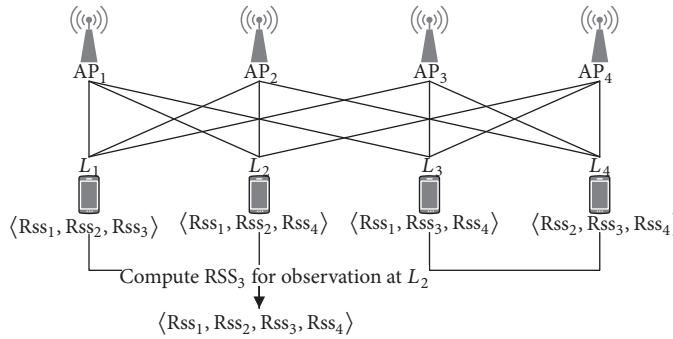


Figure 5: Dealing with the missed AP problem.

Let $\mathrm{RSS}_{0,j}$ denote the observed value of RSS of the jth AP at $d_0 = 1$ m; we can obtain the relationship between the value of the RSS and the distance by

$$\mathrm{RSS}_{i,j} = -10n \lg d_{i,j} + \mathrm{RSS}_{0,j}. \tag{6}$$

Moreover, the relationship among the location of the ith observing location (L_i), the location of the jth AP (A_j), and the distance between L_i and A_j can be formulated as

$$d_{i,j} = \left\| L_i - A_j \right\|_2. \tag{7}$$

The above assumptions are also made in our previous work in [21] and many other approaches. With these assumptions, we can compute the RSS values for the missed APs and fill in the blank items in the sparse matrix of RSS by using the maximum likelihood estimate with probability density function:

$$f_{\mathrm{RSS}}\left(\mathrm{RSS}_{i,j}\right) - \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{\left(\mathrm{RSS}_{i,j} - \widetilde{\mathrm{RSS}}_{i,j}\right)^2}{2\sigma^2}\right\}. \tag{8}$$

Under the independent and identical distribution hypothesis on L, A, and RSS_0, the maximum probability of RSS is observed as

$$p\left(L, A, \mathrm{RSS}_0 \mid \mathrm{RSS}\right) = \prod_{i,j} f_{\mathrm{RSS}}\left(\mathrm{RSS}_{i,j}\right). \tag{9}$$

This is equivalent to minimizing

$$F\left(S \parallel \widetilde{S}\right) = \sum_{i,j}\left(\mathrm{RSS}_{i,j} - \widetilde{\mathrm{RSS}}_{i,j}\right)^2 \tag{10}$$

$$= \sum_{i,j}\left(\mathrm{RSS}_{i,j} - \mathrm{RSS}_{0,j} + 10n \lg \left\| L_i - A_j \right\|_2\right)^2. \tag{11}$$

We define the estimation error as

$$\begin{aligned} E_{i,j} &= \mathrm{RSS}_{i,j} - \widetilde{\mathrm{RSS}}_{i,j} \\ &= \left(\mathrm{RSS}_{i,j} - \mathrm{RSS}_{0,j} + 10n \lg \left\| L_i - A_j \right\|_2\right). \end{aligned} \tag{12}$$

The process of fitting can be achieved by using the stochastic gradient descent method:

$$\begin{aligned} L_i &\coloneqq L_i - \alpha \cdot \frac{\partial F}{\partial L} = L_i - 2\alpha \cdot \frac{10n \cdot \left(L_i - A_j\right) E_{i,j}}{\ln 10 \cdot \left\| L_i - A_j \right\|_2}, \\ A_j &\coloneqq A_j - \alpha \cdot \frac{\partial F}{\partial A} \\ &= A_j + 2\alpha \cdot \frac{10n \cdot \left(L_i - A_j\right) E_{i,j}}{\ln 10 \cdot \left\| L_i - A_j \right\|_2}, \\ \mathrm{RSS}_{0,j} &\coloneqq \mathrm{RSS}_{0,j} - \alpha \cdot \frac{\partial F}{\partial \mathrm{RSS}_0} = \mathrm{RSS}_{0,j} + 2\alpha E_{i,j}, \end{aligned} \tag{13}$$

where α is the length of step.

This fitting process is hoping to compute a large number of unknown data (i.e., the RSS values to the missed APs) from only a little amount of given data. As a result, the convergence of this fitting process is generally describing the random error or noise instead of the underlying relationship between the RSS and the location information. To address this problem, a typical solution is to use regularization, which modifies the objective function as

$$\begin{aligned} \min_{L_i, A_j, \mathrm{RSS}_{0,j}} &\sum_{i,j}\left(\mathrm{RSS}_{i,j} - \mathrm{RSS}_{0,j} + 10n \lg \left\| L_i - A_j \right\|_2\right)^2 \\ &+ R\left(w\right), \end{aligned} \tag{14}$$

where w is the weight vector and $R(w)$ is the regularization term. Take $L2$ regularization as an example; $R(w)$ can be defined as

$$R\left(w\right) = \lambda \cdot \left[\left\| L_i \right\|^2 + \left\| A_j \right\|^2 + \left(\mathrm{RSS}_{0,j} - \overline{\mathrm{RSS}}_{0,j}\right)^2\right], \tag{15}$$

where λ is a free parameter, which needs to be adjusted by methods like cross-validation. And in our experiment, we find that, for most APs, $\overline{\mathrm{RSS}}_{0,j} = -36$ dbm. It is worth noting that it is usually difficult to use the cross-validation method, so an early exit strategy can also be used.

6. Evaluation

In this section, we evaluate our proposed localization method and compare it with some most commonly used location fingerprinting methods.

6.1. Benchmark. In the following, we detail the benchmark used in our experiments.

6.1.1. Dataset. We use an open dataset, the Mannheim/compass dataset [22], to perform our experiments. It records traces of signal strength of 802.11 APs and contains data in both an offline training phase and an online positioning phase, in an area of about 35 meters in width and 60 meters in length. The offline fingerprinting data contains 14,300 measurement records for 130 locations (110 records each), and the online positioning data contains 5,060 measurement records for 46 locations (110 records each).

6.1.2. Compared Methods. We choose three most commonly used location fingerprinting methods for comparison. All the three methods generate the same location fingerprint database by simply averaging observations at the same location.

(i) The *weighted K-nearest neighbor (WKNN) method* [23] is a deterministic method which computes the estimate location by weighted-averaging the fingerprint locations:

$$\widehat{L_x} = \sum_{i=1}^{M} \frac{w_i}{\sum_{j=1}^{M} w_j} L_i, \tag{16}$$

where w_i represents the weight of the fingerprint location L_i. It can be computed by

$$w_i = \frac{1}{\left\| O_x - O_i \right\|_2}. \tag{17}$$

Here, the Euclidean norm (2-norm) is used. WKNN keeps K biggest weights and sets the others to zero.

(ii) The *K-nearest neighbor (KNN) method* [24] is a simplified version of WKNN which sets the K biggest weights to $1/K$ and others to zero.

(iii) The *histogram method* [25] is a probabilistic method, which computes the probability that an RSS observation O_x can be observed at the location L_x by using Bayes' rule:

$$p\left(L_x \mid O_x\right) = \frac{p\left(O_x \mid L_x\right) p\left(L_x\right)}{p\left(O_x\right)}, \tag{18}$$

where $p(O_x)$ is a normalized constant and $p(L_x)$ and $p(O_x \mid L_x)$ can be computed as follows:

$$p\left(L_x\right) = \frac{\sum_{i=1}^{M} \chi B_i\left(L_x\right)}{\sum_{j=1}^{M} \left|B_j\right|},$$
$$p\left(O_x \mid L_x\right) = \sum_{i=1}^{M} p\left(O_x \mid i\right) \chi B_i\left(L_x\right), \tag{19}$$

where $|B_i|$ is the volume of B_i and

$$\chi B_i\left(x\right) = \begin{cases} 1, & x \in B_i \\ 0, & x \notin B_i, \end{cases}$$
$$p\left(O_x \mid i\right) \approx \prod_{j=1}^{N} H_{v_{ij}}\left(O_{xj} - O_{ij}\right), \tag{20}$$

where $v_i = O_x - O_i$ and $H_{v_{ij}}(L_x)$ is the normalized centralized histogram.

6.1.3. Experiment Settings. We use a program to randomly choose the RSS observations based on two parameters. y represents the number of observations selected in each observing location (for both offline and online); z represents the size of a cell where one offline observing location is selected (i.e., the volume of the observing location). For example, when $y = 8$ and $z = 36 \times 2.25 = 81\,\mathrm{m}^2$, this means that, for every $81\,\mathrm{m}^2$ area, there should be no more or no less than one offline observing location, and at this location, 8 RSS observations are selected. An online observing location is then randomly selected, and again at this location, 8 RSS observations are selected. Using these online and offline data, the location is estimated by using each of the three comparing methods and also our proposed method. We let $y \in \{1, 2, 4, 8, 16, 32, 64\}$ and $z \in \{1, 4, 9, 16, 25, 36\} \times 2.25\,\mathrm{m}^2$ (the distance between two nearby blue dots in Figure 6 is 1.5 m in real world, so $2.25\,\mathrm{m}^2$ is the minimum value for z). The experiment is performed 100 times for each pair of $\langle y, z \rangle$.

6.1.4. Comparing Metrics. Two metrics are used for comparing the performance of different localization methods: the mean localization error and the mean squared localization error.

6.2. Results and Analysis. The results are shown in Figures 7 and 8. With different settings on y and z, we illustrate the mean localization error in Figures 7(a)–7(f) and the mean squared localization error in Figures 8(a)–8(f), for all the compared localization methods.

It is worth noting that the experiment results provide a direct answer to the questions we listed in the Introduction. From Figure 7(a), we can see that KNN and WKNN can output accurate locations (i.e., the mean localization error is about 2 m) when we can observe at least 4 RSS measurements within an area of $2.5\,\mathrm{m}^2$ in size. From Figure 7(b), we can see that 16 RSS measurements are required for a $10\,\mathrm{m}^2$ area. This means that, on average, one should observe at least 1.6 RSS measurements per square meter to achieve accurate localization. This is the answer to the question "How many RSS measurements are required to compute an accurate location?" with our experiment settings.

The results also show that our proposed method may be an answer to the question "How do we locate accurately with insufficient RSS measurements?" We can see that, whatever the values of y and z, our propose method achieves smaller mean localization errors and smaller mean

Figure 6: Floorplan of the testing area in the dataset used for evaluation. The red dots show the locations where online RSS observations are collected, and the blue dots show the offline observing locations. Yellow dots show the locations of the APs, which we do not assume to be known in our experiments.

squared localization errors, especially when z (the number of observing locations) is relatively large and y (the number of observations at each location) is relatively small. This is reasonable since traditional methods can denoise RSS measurements at the same location, so our proposed method does not have evident superiority with a large y and a small z. However, with a small y and a large z, our proposed method can (while the compared methods cannot) address the missed AP problem and denoise the RSS measurements at different observing locations.

The performances of KNN and WKNN are nearly the same, and the performance of the histogram method is not as good as other methods. It is always with a large mean localization error and a large mean squared error. Besides, the histogram method can fail to estimate a location when the RSS measurements are not sufficient. The failure rate is as shown in Figure 9.

7. Conclusion

This paper investigates the problem of localization arising from insufficient RSS measurements, that is, the missed AP problem and the RSS measurement noise problem. Traditional location fingerprinting methods rely on a large quantity of RSS observations at the same location to finally observe all the APs so that no APs can be missed from the location fingerprints and to denoise RSS measurements by averaging RSS observations at the same location. We propose a novel localization method which uses the maximum likelihood estimation and the stochastic gradient descent to estimate locations in case the RSS measurements are insufficient to generate accurate location fingerprints. The results show that our proposed method can achieve better localization accuracy than most commonly used location fingerprinting methods like the KNN, WKNN, and histogram methods. Especially when the number of observations at each location

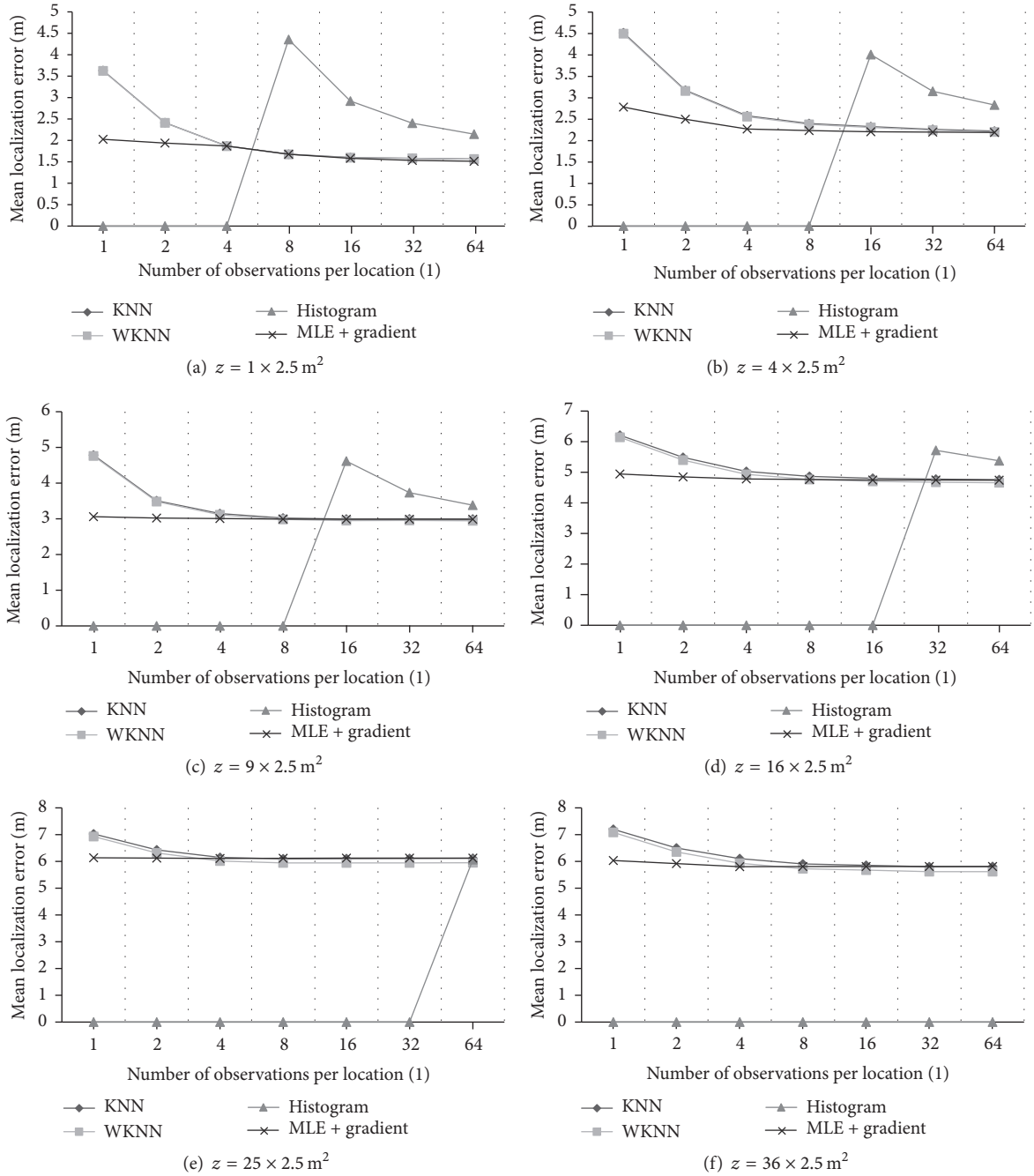


FIGURE 7: Mean localization error to the number of observations per location y, for the KNN, WKNN, histogram method, and our proposed method, with different settings on the volume of observing location z.

is relatively small, our proposed method has evident superiority.

Notations

$RSS_{i,j}$: RSS measurement at the jth location to the jth AP
L_i: Coordinates of the jth observing location
A_j: Coordinates of the jth AP's location
$d_{i,j}$: Distance between L_i and A_j
S: The set of all RSS measurements
E: The estimation error
O: Observation of RSS measurements
M: The number of observing locations
N: The number of APs
y: The number of observations per location
z: The volume of each observing location.

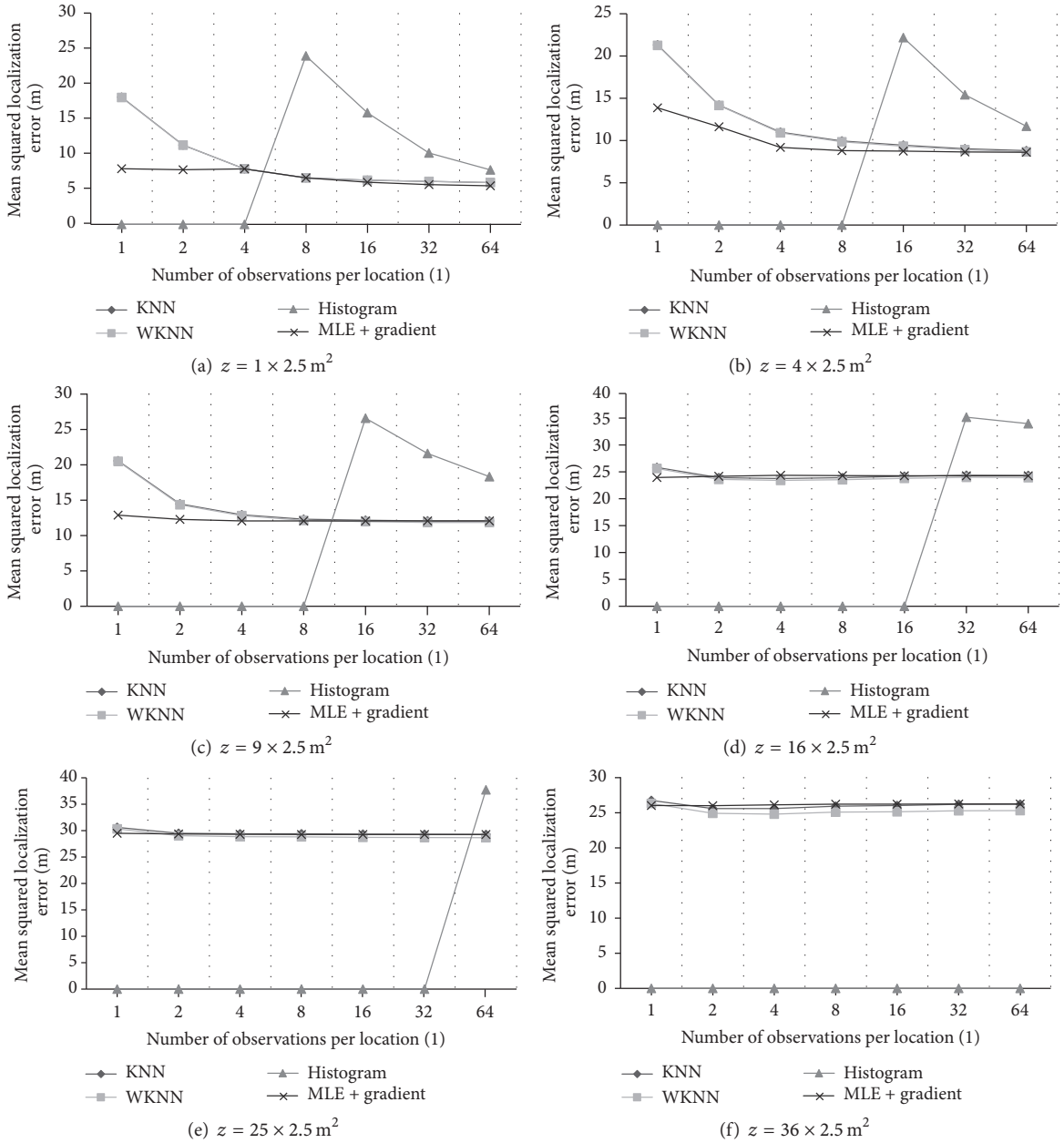


Figure 8: Mean squared localization error to the number of observations per location y, for the KNN, WKNN, histogram method, and our proposed method, with different settings on the volume of observing location z.

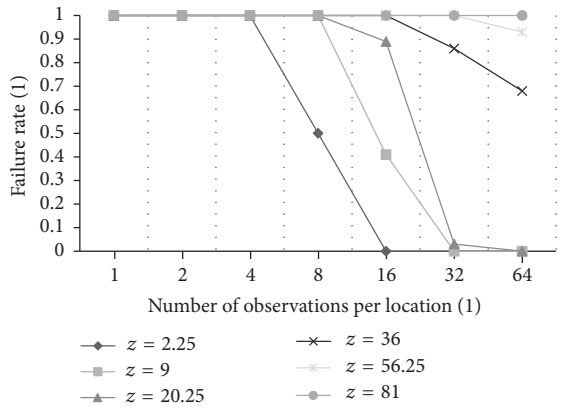


Figure 9: Failure rate to the number of the observations per location y, for the histogram method, with different observing location volumes z.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 61602111, 61320106007, 61402104, 61502100, 61532013, 61572130, and 61632008, by the Jiangsu Provincial Natural Science Foundation of China under Grants BK20150628, BK20140648, and BK20150637, and by the Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

[1] R. Faragher and R. Harle, "Location fingerprinting with bluetooth low energy beacons," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 11, pp. 2418–2428, 2015.

[2] P. Yang, W. Wu, M. Moniri, and C. C. Chibelushi, "Efficient object localization using sparsely distributed passive RFID tags," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 12, pp. 5914–5924, 2013.

[3] K. Chintalapudi, A. P. Iyer, and V. N. Padmanabhan, "Indoor localization without the pain," in *Proceedings of the 16th Annual Conference on Mobile Computing and Networking (MobiCom '10)*, pp. 173–184, September 2010.

[4] K. Wu, J. Xiao, Y. Yi, D. Chen, X. Luo, and L. M. Ni, "CSI-based indoor localization," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 7, pp. 1300–1309, 2013.

[5] Z. Yang, Z. Zhou, and Y. Liu, "From RSSI to CSI: indoor localization via channel response," *ACM Computing Surveys*, vol. 46, no. 2, article 25, 2013.

[6] F. Yang, Q. Zhai, G. Chen, A. C. Champion, J. Zhu, and D. Xuan, "Flash-Loc: Flashing mobile phones for accurate indoor localization," in *Proceedings of the 35th Annual IEEE International Conference on Computer Communications, IEEE INFOCOM 2016*, San Francisco, CA, USA, April 2016.

[7] W. Wang, A. X. Liuy, and K. Sun, "Device-free gesture tracking using acoustic signals," in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking, MobiCom 2016*, pp. 82–94, New York, NY, USA, October 2016.

[8] W. Mao, J. He, and L. Qiu, "CAT: High-precision acoustic motion tracking," in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking, MobiCom 2016*, pp. 69–81, New York, NY, USA, October 2016.

[9] S. He, S.-H. Gary Chan, L. Yu, and N. Liu, "Fusing noisy fingerprints with distance bounds for indoor localization," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM '15)*, pp. 2506–2514, IEEE, Hong Kong, April 2015.

[10] Y. Wen, X. Tian, X. Wang, and S. Lu, "Fundamental limits of RSS fingerprinting based indoor localization," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM '15)*, pp. 2479–2487, Hong Kong, May 2015.

[11] P. Barsocchi, S. Lenzi, S. Chessa, and F. Furfari, "Automatic virtual calibration of range-based indoor localization systems," *Wireless Communications and Mobile Computing*, vol. 12, no. 17, pp. 1546–1557, 2012.

[12] L. Li, W. Yang, M. Z. Alam Bhuiyan, and G. Wang, "Unsupervised learning of indoor localization based on received signal strength," *Wireless Communications and Mobile Computing*, vol. 16, no. 15, pp. 2225–2237, 2016.

[13] J. Jun, S. Chakraborty, L. He, Y. Gu, and D. P. Agrawal, "Robust and undemanding wifi-fingerprint based indoor localization with independent access points," in *Proceedings of the Microsoft Indoor Localization Competition (IPSN)*, pp. 13–17, Seattle, WA, USA, 2015.

[14] Q. Zhang, Z. Zhou, W. Xu et al., "Fingerprint-free tracking with dynamic enhanced field division," in *Proceedings of the 34th IEEE Annual Conference on Computer Communications and Networks, IEEE INFOCOM 2015*, pp. 2785–2793, Kowloon, Hong Kong, May 2015.

[15] A. Rai, K. K. Chintalapudi, V. N. Padmanabhan, and R. Sen, "Zee: Zero-effort crowdsourcing for indoor localization," in *Proceedings of the 18th annual international conference on Mobile computing and networking (Mobicom '12)*, pp. 293–304, August 2012.

[16] H. Wang, S. Sen, A. Elgohary, M. Farid, M. Youssef, and R. R. Choudhury, "No need to war-drive: unsupervised indoor localization," in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services (MobiSys '12)*, pp. 197–210, ACM, June 2012.

[17] Z. Yang, C. Wu, and Y. Liu, "Locating in fingerprint space: wireless indoor localization with little human intervention," in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking (Mobicom '12)*, pp. 269–280, August 2012.

[18] A. Purohit, Z. Sun, S. Pan, and P. Zhang, "SugarTrail: Indoor navigation in retail environments without surveys and maps," in *Proceedings of the 2013 10th Annual IEEE Communications Society Conference on Sensing and Communication in Wireless Networks, SECON 2013*, pp. 300–308, New Orleans, LA, USA, June 2013.

[19] K. Sheng, Z. Gu, X. Mao et al., "The collocation of measurement points in large open indoor environment," in *Proceedings of the 34th IEEE Annual Conference on Computer Communications and Networks, IEEE INFOCOM 2015*, pp. 2488–2496, Kowloon, Hong Kong, May 2015.

[20] C. Wu, Z. Yang, C. Xiao, C. Yang, Y. Liu, and M. Liu, "Static power of mobile devices: Self-updating radio maps for wireless indoor localization," in *Proceedings of the 34th IEEE Annual Conference on Computer Communications and Networks, IEEE INFOCOM 2015*, pp. 2497–2505, Kowloon, Hong Kong, May 2015.

[21] K. Dong, W. Wu, H. Ye, M. Yang, Z. Ling, and W. Yu, "Canoe: an autonomous infrastructure-free indoor navigation system," *Sensors*, vol. 17, article 996, no. 5, 2017.

[22] T. King, S. Kopf, T. Haenselmann, C. Lubberger, and W. Effelsberg, CRAWDAD data set mannheim/compass (v. 2008-04-11), Downloaded from http://crawdad.org/mannheim/compass/, Apr. 2008.

[23] B. Li, J. Salter, A. G. Dempster, and C. Rizos, "Indoor positioning techniques based on wireless lan," in *Proceedings in the LAN, first IEEE international conference on wireless broadband and ultra wideband communications, Citeseer*, 2006.

[24] P. Bahl and V. N. Padmanabhan, "Radar: an in-building rf-based user location and tracking system," in *Proceedings in the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2000)*, vol. 2, pp. 775–784, 2000.

[25] T. Roos, P. Myllymäki, H. Tirri, P. Misikangas, and J. Sievänen, "A Probabilistic Approach to WLAN User Location Estimation," *International Journal of Wireless Information Networks*, vol. 9, no. 3, pp. 155–164, 2002.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 6345316, 11 pages
https://doi.org/10.1155/2017/6345316

Research Article

Compressed RSS Measurement for Communication and Sensing in the Internet of Things

Yanchao Zhao,[1,2,3] Wenzhong Li,[2] Jie Wu,[4] Sanglu Lu,[2,3] and Bing Chen[1,3]

[1]*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China*
[2]*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China*
[3]*Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, China*
[4]*Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA*

Correspondence should be addressed to Yanchao Zhao; yczhao@nuaa.edu.cn

Received 29 April 2017; Accepted 6 July 2017; Published 7 August 2017

Academic Editor: Feng Wang

The receiving signal strength (RSS) is crucial for the Internet of Things (IoT), as it is the key foundation for communication resource allocation, localization, interference management, sensing, and so on. Aside from its significance, the measurement process could be tedious, time consuming, inaccurate, and involving human operations. The state-of-the-art works usually applied the fashion of "measure a few, predict many," which use measurement calibrated models to generate the RSS for the whole networks. However, this kind of methods still cannot provide accurate results in a short duration with low measurement cost. In addition, they also require careful scheduling of the measurement which is vulnerable to measurement conflict. In this paper, we propose a compressive sensing- (CS-) based RSS measurement solution, which is conflict-tolerant, time-efficient, and accuracy-guaranteed without any model-calibrate operation. The CS-based solution takes advantage of compressive sensing theory to enable simultaneous measurement in the same channel, which reduces the time cost to the level of $\mathcal{O}(\log N)$ (where N is the network size) and works well for sparse networks. Extensive experiments based on real data trace are conducted to show the efficiency of the proposed solutions.

1. Introduction

With the ubiquitous wireless networking devices, we envision realizing the Internet of Things, which requires the wireless networks to develop with higher spectrum utilization, less transmission delay, and lower energy consumption. This trend gives birth to many emerging technologies (e.g., OFDMA, network coding, and cognitive radio [1–3]). To realize the IoT, a key job is how to optimize the allocation of existing wireless communication resources (e.g., link scheduling, channel allocation, and power allocation [4, 5]), improve the communication bandwidth, and reduce the communication power consumption. Aside from the traditional usage of data transmission, recent research efforts even extend the usage of the wireless signal in IoT to perform localization [6] and activity recognition [7]. All of them, although used differently, rely on the receiving signal strength (RSS) and its variant. Specifically, for communication resource allocations, it requires the RSS to build the interference models. Nevertheless, for localization in wireless networks, it requires the RSS to compute the distance or to generate the fingerprint of certain location. Thus, the accuracy of the measured RSS will finally affect the optimization results as well as the localization accuracy. The efficiency of the measurement process will also affect the applicability and efficiency of these wireless applications.

Towards the efficient and accurate RSS measurement and its application in resource allocation and localization, most existing works [8–10] mainly focus on how to derive the metric (e.g., SINR and RSS fingerprint) by some signal propagation models (e.g., the path loss model [10]) with fixed empirical parameters. However, such propagation models and the corresponding empirical parameters cannot characterize the complex, time-varying channel conditions

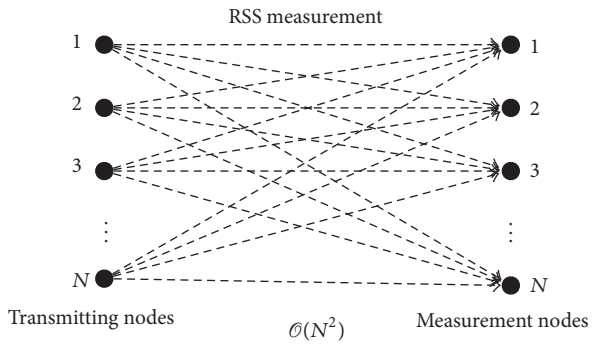


Figure 1: Traditional exhaustive RSS measurement.

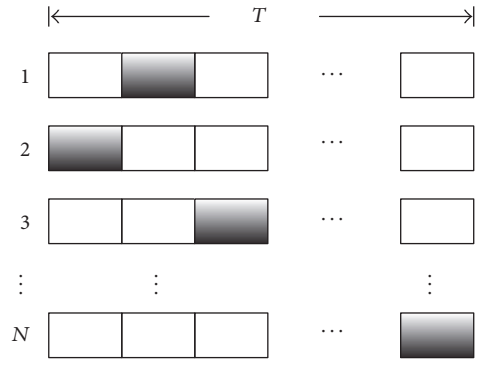


Figure 2: Nonconflict RSS measurement.

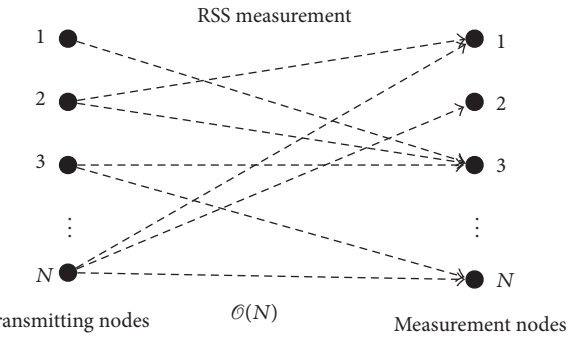


Figure 3: Compressive sensing-based RSS measurement.

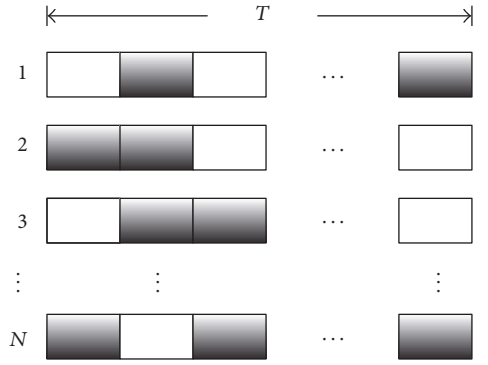


Figure 4: Simultaneous RSS measurement.

accurately, which in turn compromise the optimization and localization results.

To obtain accurate RSS values, exhausting measurement on all wireless links will incur unacceptable time cost. First, the RSS collection process, for example, the SINR-optimization process, requires the RSS values between every pair of nodes in the network, so the number of RSS values to be measured grows quadratically with the network size. As is illustrated in Figure 1, the left row and right row of the nodes are the same set of nodes in the network, while the links between the nodes represent the measurement conducted between different nodes. The RSS measurements should be conducted in every pair of nodes, thus leading to a measurement cost in level of $\mathcal{O}(N^2)$. Meanwhile, to measure the RSS accurately, it usually requires that the receiving signal is decodable, such that the transmitting node could be identified with the source information recoded in the packet. As is illustrated in Figure 2, no measurement should be performed simultaneously in the process of RSS measurement, because the conflict will lead to two major drawbacks. First, the conflict of two signals could cause the failure of packet decoding, so that the sources of the signals could not be identified. Secondly, even the packet could be decoded when the capture effect exists; the measured RSS will be the overlapping one of both signals; thus, the measured RSS will be far away from the intended one. Due to the above reasons, traditional RSS measurements, as illustrated in Figure 2, where the grey grid stands for the slot used for measurements, were performed in a nonconflicting way. This kind of measurement method makes the time consumption of whole measurement process up to the level of $\mathcal{O}(N^2)$.

Both the measurement cost and the time consumption are unacceptable for a network with large size of nodes. To tackle this, we propose a compressive sensing-based solution, whose basic idea could be summarized as two aspects: *partial measurement* and *simultaneous measurement*.

The idea of partial measurement is illustrated in Figure 3. As in the figure, only a subset of node pairs is selected to measure the RSS. The number of measured links could be as few as $\mathcal{O}(N)$ or even decreased to the level of $\mathcal{O}(\log N)$, which is achieved by the method proposed in this paper. Based on the measured RSS, the other unknown RSS values could be estimated with certain accuracy. In summary, the partial measurement method is based on the concept of "measure a few, predict many."

The idea of simultaneous measurement is illustrated in Figure 4, where T is the total time used, while the grids colored with grey imply that slots are used for measurement. It clearly presents the advantage that the total measurement time could be deducted. However, as aforementioned, the simultaneous measurement is forbidden, as we cannot tell two overlapped signals apart. In turn, we also cannot get the values of RSS and do not know where is the signal coming from. However, by utilizing the nature of linear additivity of RSS, we show how to distinguish the source of the RSS without assuming that the signal could be decoded and tell the values of the RSS.

The problems associated with the partial measurement and simultaneous measurement could be jointly solved in our proposed compressive sensing-based solution, which mainly introduce the idea of compressive sensing and lead to the measurement cost and time cost to the level of $\mathcal{O}(\log N)$. The CS-based solution allows simultaneous measurements on a single channel. This is mainly empowered by the development of compressive sensing theory [11–13]. The efficiency is achieved by the result of the CS theory that not enough overlapping measurements (N dimensions require N sets of measurements) can recover the original matrix (in our case it refers to the RSS matrix). It further reduces the time cost to the level of $\mathcal{O}(\log N)$, which highly outperforms the traditional solutions. However, it only achieves acceptable accuracy in sparse networks. The accuracy is guaranteed by controlling the number of measurements and the sparsity of the RSS matrix. A number of theories are established to study the relationship between the accuracy and the number of measurements [13]. A measurement matrix generation method and a sparsity control method are proposed to address the specific issues of CS-based measurement problem.

The contributions of this paper are summarized as follows:

(i) We reveal the important problem of accurate and efficient RSS measurement for communication and sensing in IoT.

(ii) We modeled the RSS measurement process as a linear system and propose a basic framework to perform simultaneous measurement of RSS instead of the traditional nonconflict measurement.

(iii) We further present a compressive sensing- (CS-) based solution to achieve partial measurement. It can achieve the time efficiency of $\mathcal{O}(\log N)$ with accuracy control. Both centralized and distributed solutions are proposed.

(iv) We introduce a LDPC-based measurement matrix, which only generates a small number of measurements. It greatly reduces the energy consumption for IoT.

(v) We conduct extensive experiments using real communication traces collected from a wireless mesh network testbed, which show the efficiency of the proposed solutions.

The rest of this paper is organized as follows. Section 5 introduces related work. The system model and problem definition are presented in Section 2.1. The CS-based solution is presented in Section 3. The numerical results are illustrated in Section 4. Finally, the paper is concluded in Section 6.

2. Preliminaries

In this section, we explore the first step to achieve partial measurement and simultaneous measurement, which is to model the efficient RSS measurement problem as a linear system. Before proposing the modeling, we first propose the network model and some important metrics to evaluate our solution.

2.1. Network Model. We consider a synchronized, time-slotted wireless network consisting of N nodes denoted by $\mathcal{N}$. A set of channels, denoted by $\mathcal{M}$, is available to each node in $\mathcal{N}$. We denote P as the sending power of each node operating over any channel and by p_{ij}^m as the RSS of a signal from node $i \in \mathcal{N}$ over channel $m \in \mathcal{M}$ received at node $j \in \mathcal{N}$.

Our main task is to *obtain all the RSS values over each pair of nodes and each channel*: that is,

$$\left\{ p_{ij}^m \mid i, j \in \mathcal{N},\ i \neq j,\ m \in \mathcal{M} \right\}. \tag{1}$$

A measurement scheme could be evaluated via the following metrics:

(i) Time cost: it is the total time slots to accomplish the measurement process.

(ii) Overhead: it is the total number of measurements in all nodes and channels.

(iii) Accuracy: we defined two levels of accuracy, which are link-wise accuracy and network-wide accuracy, receptively. Regarding link-wise accuracy, the result of measurement should be within a certain level of confidence $1 - \alpha/2$. Regarding the network-wide accuracy, it implies that β portion of $\{p_{ij}^m\}$ is accurate. The CS-based solution only achieves network-wide accuracy.

It is worth mentioning that the relation between link-wise accuracy and network-wide accuracy is not exclusive. One can achieve both of them when restrictions are put in the frequency in single link measurement and also optimization accuracy is put in global optimization results. However, such dual restriction will lead to unacceptable overhead. In fact, one can achieve the link-wise accuracy through a measurement method [14], which is independent with our compressive sensing method. Our method could provide a promise on the network-wide accuracy that β portion of $\{p_{ij}^m\}$ is accurate but leave link-wise accuracy for the implementation.

The accuracy is assured by an adequate number of measurements, while the overhead and time cost metric require as few measurements as possible. Thus, our target is to design the solution that achieves good tradeoffs among these metrics.

2.2. A Linear System Formulation. Basically, the measurement process can be modeled as a linear system. By applying the prevalent compressive sensing [12] theory on this linear system with proper specification, we derive an efficient measurement with low overhead and time cost.

Before introducing the solution with partial measurement and simultaneous measurement, we first formulate our problem in the form of a linear system.

According to the SINR model, the RSS is approximately linear additive. This property implies that when several nodes in the network send signals in the same slot, the RSS of a

certain node is the sum of the RSSs from all of the sending nodes. Formally, in one time slot, we have $r_j = \sum_i \phi_i p_{ij}$, where r_j stands for the RSS measured in node j and ϕ_i is a binary variable standing for whether node i should send a measurement signal in this time slot. When we extend this formulation into the scenario of multiple time slots and ensemble them into matrix form, we have

$$\mathbf{R} = \mathbf{\Phi P}. \tag{2}$$

Here, $\mathbf{P} = [p_{ij} \mid 0 \le i, j \le N]$ is called a RSS matrix, whose element p_{ij} represents the RSS from node i to node j. The matrix $\mathbf{R} \in \mathbb{R}^{N \times T}$ stands for the measurement result. The element in the ith row and the jth column, denoted as r_{ij}, stands for the RSS measured in node i at time slot j. According to the SINR model, r_{ij} is the sum of the value of the RSS that node i received. The matrix $\mathbf{\Phi} \in \{0, 1\}^{T \times N}$, called measurement matrix, stands for the measurement schedule for each node and each time slot, where $\phi_{tj} = 1$ indicates that node j sends a measurement signal in slot t.

From this linear system perspective, our problem could be stated as follows.

Definition 1 (efficient RSS measurement problem). Given a network of N nodes, try to get the RSS matrix $\mathbf{P} = [p_{ij}, \forall i, j \in \{0, 1, \ldots, N\}]$ through a planned measurement process $\mathbf{R} = \mathbf{\Phi P}$ with minimum $|\text{row}(\mathbf{\Phi})|$.

Note that, if we choose $\mathbf{\Phi}$ as an $N \times N$ identity matrix, the RSS matrix could be easily recovered. However, as the column number of $\mathbf{\Phi}$ stands for the measurement slots, we need to generate a matrix $\mathbf{\Phi}$ with $T \ll N$. Thus, an $N \times N$ identity matrix is unacceptable, especially for the networks with a large number of nodes. The linear system has a unique solution only if $\text{rank}(\mathbf{\Phi}) = N$. Thus, a matrix $\mathbf{\Phi}$ with $T \ll N$ is not enough for solving $\mathbf{R} = \mathbf{\Phi P}$ with a unique solution. However, this formula could be resolved with the tools provided by compressive sensing theory, as long as $\mathbf{P}$ is sparse enough. The accuracy of recovered $\mathbf{P}$ is assured with high probability.

2.3. Fundamentals of Compressive Sensing. Before presenting our solutions, we briefly introduce the compressive sensing theory. Compressive sensing (or sampling) (CS) [12, 13] is a notion generated from the field of signal processing. In the conventional paradigm, natural signals are first acquired at the Nyquist-Shannon sampling rate and then compressed for efficient storage or transmission. CS shifts this paradigm by combining the two processes into a single compressive sampling process, greatly reducing the complexities in data acquisition. The most important idea in CS theory is that a small amount of random linear projections of sparse or compressible signals have contained sufficient information for signal reconstruction and processing.

In other words, signals can be accurately rebuilt based on the following conditions:

(1) The a priori knowledge of sparsity or compressibility of signals is known.

(2) A small number of global linear measurements are provided.

In purpose of integrity and consistency, we present the following definitions.

Definition 2 (sparse signal). Let $d = (d_1, d_2, \ldots, d_N)^\top$ be an N-dimensional signal. We say d is a K-sparse signal if there are only K ($K \ll N$) nonzero entries in d. Further, we say d is a K-sparse signal in x domain, if there exists a set of orthonormal basis, denoted as $\Psi = [\psi_1, \psi_2, \ldots, \psi_N]$, $\psi_i \in \mathbb{R}^N$,, in which d can be represented by a K-sparse vector x:

$$d = \sum_{i=0}^{n} \psi_i x, \quad \text{or } d = \mathbf{\Psi} x. \tag{3}$$

Compressive sensing theory states that an N-dimensional signal s, which is K-sparse in the domain of $\mathbf{\Psi}$ ($s = \mathbf{\Psi} x$, x is a K-sparse vector), can be efficiently represented by T ($T < N$) linearly measurements. Specifically, let $\mathbf{\Phi}$ be a $T \times N$ ($T < N$) matrix; then, the measurements of s can be obtained by $y = \mathbf{\Phi} s, s = \mathbf{\Psi} x$, where y is the measurement results. Matrix $\mathbf{\Phi}$ is referred to as measurement matrix and the matrix $\mathbf{\Psi}$ is referred to as the representing basis. The key questions are whether it is possible and how to recover the N-dimensional signal s from the T-dimensional measurements y. Candes and Tao [13] have shown that when $K \le (1/2)T$, and $\mathbf{\Phi}$ follows the restricted isometry property (RIP) [11], the exact recovery of d can be achieved through solving a linear optimization problem:

$$\min_{x \in \mathbb{R}^N} \quad \|x\|_{l_1} \qquad \text{s.t.} \quad y = \mathbf{\Phi} s, \; s = \mathbf{\Psi} x. \tag{4}$$

The l_p norm of vector x is defined as $\|x\|_{l_p} = (\sum_{p=1}^{N} |x_i|^p)^{1/p}$. It is known that l_1-minimization problem can be solved with linear programming (LP) techniques [11]. The l_1/l_0 equivalence relies on the incoherence property [15] between $\mathbf{\Phi}$ and $\mathbf{\Psi}$ or the restricted isometry property (RIP) [11] of matrix $\mathbf{\Phi} = \mathbf{\Phi\Psi}$.

It has been established that Gaussian matrix $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$, whose entries are independently and identically distributed realizations of certain zero mean random variables with variance $1/T$, satisfies the RIP with high probability when $T \ge C(K \log(N/K))$, where C is a constant [16].

3. Compressive Sensing-Based Solution

As aforementioned, efficient RSS measurement relies on the process of partial measurement and simultaneous measurement. The former could be achieved with solving the linear system modeling, while the latter relies on how we recover the RSS matrix with only a few time slot measurements. Our basic idea is applying the compressive sensing theory to the linear system.

The process of partial measurement and simultaneous measurement is mainly enabled by the careful design of

measurement matrix in the compressive sensing, which owns the ability to recover the full signal from partially measured signal information and also has the ability to distinguish the overlapped signal when they are linearly combined. We will discuss it specifically when we propose the design of measurement matrix.

3.1. Solution Framework. The success of the solution depends on two crucial components. The first one is the generation of measurement matrix with good RIP. According to the theorem in [17], good RIP refers to that δ_k (restricted isometry constant) is smaller than $\sqrt{2}-1$. Further, we should also find a representation basis, in the space of which the RSS matrix could be represented in the form of K-sparse matrix. Note that we deal with a matrix rather than a vector. In the context of matrices, low rank is analogous to sparsity because the spectrum formed by the singular values of a low-rank matrix is sparse. Thus, in our problem, the K-sparse matrix means the rank of matrix is K.

Assume that the RSS matrix $\mathbf{P}$ is K-sparse in a certain domain, which is formally stated as

$$\mathbf{P} = \mathbf{\Psi}\mathbf{P}'. \tag{5}$$

According to (2), the measurement result can be expressed by the product of a matrix and a RSS matrix. Thus, we can rewrite $\mathbf{R}$ by

$$\mathbf{R} = \mathbf{\Phi}\mathbf{\Psi}\mathbf{P}', \tag{6}$$

where $\mathbf{\Phi}$ is the measurement matrix to be decided. Based on these formalizations, we now proceed to the determination of measurement matrix and representation basis.

3.1.1. Measurement Matrix. In our measurement, the measurement matrix is a binary matrix, with each row as the sending plan of all nodes in one time slot. The number of rows represents the number of time slots used to perform measurements. In addition, because the representation basis is usually orthonormal matrix, the RIP of matrix $\mathbf{\Phi\Psi}$ is the same as $\mathbf{\Phi}$. Our target is to find a binary matrix, which has good RIP and a small number of rows and fulfill the partial measurement and simultaneous measurement requirements.

To design the measurement matrix satisfying the partial measurement requirement, we only have to make sure the measurement matrix is in form of $\{\mathbf{\Phi}_{MN},\ M < N\}$, where M is the row number. In such way, the measurement time cost is smaller than the size of the signal (network size in our case). Thus, this measurement matrix will be satisfied with the partial measurement requirement.

To design the measurement matrix satisfying the simultaneous measurement requirement, we only have to make sure the measurement matrix is in form of $\{\mathbf{\Phi}_{MN},\ \sum_i \phi_{ij} \geq 1\}$. This implies that each row of the measurement matrix has multiple nonzero entries, which also means there are more than one node performing measurement at the same time. It is worth mentioning that the number of nonzero entries in each row has direct relation with the least number of rows of measurement matrix to ensure an accurate recovery of the RSS matrix. Specifically, it has to make sure the measurement matrix satisfy the RIP. Thus, by considering both the partial measurement and simultaneous measurement requirements, how to design a good matrix with RIP becomes our major concern. In the following, we present our design.

To begin with, we derive the size of rows for the measurement matrix. This will inherently control the network-wide accuracy. According to [13], T must be larger than $K\log(N/K)$ to provide a recovery accuracy of $1-2e^{-T\delta/8}$, where δ is the restricted isometric constant.

Then, we derive the elements of measurement matrix. As aforementioned, the measurement matrix in CS is usually drawing from a random matrix whose entries are i.i.d. Gaussian variables complying to $\mathcal{N} \sim (0, 1/T)$ [13]. However, due to the randomness in structure and the uncertainty on RIP, these random matrices are prohibited in real applications. In our application, a binary measurement matrix is required. A simple way is to generate a binary matrix with entries complying to Bernoulli distribution with success probability p. According to the proof in [16], this kind of matrix bears good RIP *w.h.p.*. However, this kind of matrix may have a large portion of nonzero entries, which, in other words, introduces large measurement overheads. To reduce the overhead of measurement, we could apply the result from the work [18]. It provides a binary matrix generated by LPDC (Low Density Parity-Check) whose definition is as follows.

Definition 3 (LDPC matrix [18]). A binary matrix $A(T, N, d) \in \{0,1\}^{T\times N}$ consists of $2 \leq d \leq T-2$ nonzero entries per column and Nd/T nonzero entries per row. In structure, any two columns are allowed to share at most one same nonzero position.

According to [18], this matrix has good RIP and low density; thus, it greatly fulfills our requirements.

The measurement overhead could be further reduced by a matrix manipulating trick. We split the measurement matrix $\mathbf{\Phi}$ into two parts, $[\mathbf{\Phi_1} \mid \mathbf{\Phi_2}]$. Here $\mathbf{\Phi_1}$ is a $T\times T$ matrix and $\mathbf{\Phi_2}$ is an $(N-T)\times T$ matrix. In this form, we have the following theorem.

Theorem 4. *If matrix $\mathbf{\Phi_1}$ is unit matrix, and $\mathbf{\Phi_2}$ has good RIP, then $[\mathbf{\Phi_1} \mid \mathbf{\Phi_2}]$ also has good RIP.*

Proof. When $\mathbf{\Phi_2}$ comply with following equations, we say it has a good RIP:

$$(1-\delta)\|x\|_{l_2}^2 \leq \left\|\mathbf{\Phi_2}x\right\|_{l_2}^2 \leq (1+\delta)\|x\|_{l_2}^2. \tag{7}$$

Then, $[\mathbf{\Phi_1} \mid \mathbf{\Phi_2}]$ is as follows:

$$\begin{aligned}\left\|[\mathbf{\Phi_1} \mid \mathbf{\Phi_2}]\,x\right\|_{l_2}^2 &= \left\|[\mathbf{\Phi_1} \mid \mathbf{\Phi_2}]\left[x_1^T \mid x_2^T\right]^T\right\|_{l_2}^2 \\ &= \left\|[\mathbf{\Phi_1}x_1 + \mathbf{\Phi_2}x_2]\right\|_{l_2}^2 \\ &= \left\|\mathbf{\Phi_1}x_1\right\|_{l_2}^2 + \left\|\mathbf{\Phi_2}x_2\right\|_{l_2}^2 \\ &\quad + 2\left\langle\mathbf{\Phi_1}x_1, \mathbf{\Phi_2}x_2\right\rangle.\end{aligned} \tag{8}$$

Because $\mathbf{\Phi_1}$ is unit matrix, we have

$$\langle \mathbf{\Phi_1} x_1, \mathbf{\Phi_2} x_2 \rangle = \langle x_1, \mathbf{\Phi_2} x_2 \rangle = x_1^T \mathbf{\Phi_2} x_2. \tag{9}$$

Because $\mathbf{\Phi_2}$ has good RIP, then we have

$$\left| x_1^T \mathbf{\Phi_2} x_2 \right| \leq \delta \left\| x_1 \right\|_{l_2}^2 \left\| x_2 \right\|_{l_2}^2 \leq \delta \left\| x \right\|_{l_2}^2 \tag{10}$$

$$-\delta \left\| x \right\|_{l_2}^2 \leq 2 x_1^T \mathbf{\Phi_2} x_2 \leq \delta \left\| x \right\|_{l_2}^2 \tag{11}$$

$$\begin{aligned} \left\| x_1 \right\|_{l_2}^2 + (1+\delta) \left\| x_2 \right\|_{l_2}^2 &= \left\| x \right\|_{l_2}^2 + \delta \left\| x_2 \right\|_{l_2}^2 \\ &\leq (1+\delta) \left\| x \right\|_{l_2}^2 \end{aligned} \tag{12}$$

$$\begin{aligned} \left\| x_1 \right\|_{l_2}^2 + (1-\delta) \left\| x_2 \right\|_{l_2}^2 &= \left\| x \right\|_{l_2}^2 + \delta \left\| x_2 \right\|_{l_2}^2 \\ &\geq (1-\delta) \left\| x \right\|_{l_2}^2 . \end{aligned} \tag{13}$$

Combing (7), (11), (12), and (13), we have

$$(1-2\delta) \left\| x \right\|_{l_2}^2 \leq \left\| \left[\mathbf{\Phi_1} \mid \mathbf{\Phi_2} \right] x \right\|_{l_2}^2 \leq (1+2\delta) \left\| x \right\|_{l_2}^2 . \tag{14}$$

Then, $[\mathbf{\Phi_1} \mid \mathbf{\Phi_2}]$ has good RIP. □

According to this theorem, we find that if we choose an orthonormal matrix as $\mathbf{\Phi_1}$ and randomly generate $\mathbf{\Phi_2}$ with good RIP, then, whole matrix $\mathbf{\Phi}$ has the same RIP with $\mathbf{\Phi_2}$. This is due to the orthonormal columns not affecting the RIP of full random matrix. Based on this property, we can choose and identify matrix $\mathbf{I}$ as $\mathbf{\Phi_1}$, which only has one nonzero element for each column.

3.1.2. Representation Basis. In this section, we describe how to control the sparsity of RSS matrix. Note that, in the RSS matrix, most of the elements are in fact close to but not equal to zero. This situation requires us to carefully drop some elements to make the matrix sparse. With all these considerations, we apply singular value decomposition here.

Simply stated, a $N \times N$ matrix could be decomposed such that

$$\mathbf{P} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top, \tag{15}$$

where $\mathbf{U}$ and $\mathbf{V}$ are $N \times N$ unitary matrices (i.e., $\mathbf{U}\mathbf{U}^\top = \mathbf{U}^\top \mathbf{U} = \mathbf{I}$), and $\mathbf{\Sigma}$ is an $N \times N$ diagonal matrix containing the singular values. The rank of a matrix is the number of linearly independent rows or columns, which equals the number of nonzero singular values of $\mathbf{\Sigma}$.

To determine the representation basis $\mathbf{U}$ and $\mathbf{V}$, we could use the path loss model to compute an approximate matrix $\widehat{\mathbf{P}}$. A singular value decomposition performed to $\widehat{\mathbf{P}}$ could help us to get $\mathbf{U}$ and $\mathbf{V}$.

As we learned from several signal propagation models, the RSSs increase linearly with the sending power. The sending power, however, impacts the interference range of the node. In a CS-based solution, we prefer a low-rank RSS matrix. A higher sending power will result in a larger interference and, in turn, make more entries in RSS matrix not close to zero. Meanwhile, if the sending power is tuned too small, the RSSs (which are close to zero) tend to be too vulnerable when encountering noises and recovery errors. Thus, when applying the result derived from too small sending power to the ordinary scenarios, the error ratio will be amplified. A proper sending power is needed for a better performance of the CS-based solutions. According to the experiment result in Figure 7, the sending power should be set to tune the average interference ratio (defined in Section 4.1) smaller than 0.322.

3.1.3. Recovery of RSS Matrix. Different from traditional work with CS, where the recovery target is a vector, our work aims to recover a matrix with CS. According to [19], the perfect recovery equals solving the following problem:

$$\begin{aligned} \min \quad & \text{rank}\,(\mathbf{P}) \\ \text{s.t.} \quad & \mathbf{R} = \mathbf{\Phi P}. \end{aligned} \tag{16}$$

However, the problem is rather hard to solve directly. Thus, in [19], the authors provided an equivalent form of the problem:

$$\begin{aligned} \min \quad & \left\| \mathbf{P} \right\|_* \\ \text{s.t.} \quad & \mathbf{R} = \mathbf{\Phi P}. \end{aligned} \tag{17}$$

By introducing SVD in (15), we have

$$\begin{aligned} \min \quad & \left\| \mathbf{P}' \right\|_* \\ \text{s.t.} \quad & \mathbf{RV} = \mathbf{\Phi U P}'. \end{aligned} \tag{18}$$

Here, $\left\| \mathbf{P}' \right\|_*$ is the nuclear norm [19]. Thus, the problem is transformed to a convex programming problem and could be solved efficiently. Because $\mathbf{P}'$ is in its most low-rank form with the smallest nuclear norm, $\mathbf{U}\mathbf{P}'\mathbf{V}^\top$ will be the original RSS matrix w.h.p.

In summary, the CS-based solution achieves $\mathcal{O}(\log N)$ time efficiency. We summarize the CS-based solution in Algorithm 1.

3.2. Accuracy Control. The accuracy control consists of two parts, namely, controlling the row number of measurement matrix and controlling the sparsity of representation basis.

Regarding the row number of the measurement matrix, it inherently controls the network-wide accuracy. According to [13], T must be larger than $K \log(N/K)$ to provide a recovery accuracy of $1 - 2e^{-T\delta/8}$, where δ is the restricted isometric constant. Here, the recovery accuracy is related to the measurement time T. Thus, with this, we can control the tradeoff between the time consumption and the accuracy.

Regarding the sparsity of the representation basis, the accuracy of the recovery will decrease as we drop some of the small singular values of the original RSS matrix. Thus, to increase the recovery accuracy, the representation basis should comply with the sparsity K. The mathematical relation between K and the accuracy will be our future work.

3.3. Dealing with Background Noise. Apart from using the CS solution as a single solution, it could also be used as an

Require: Position of each node, Number of Nodes
Ensure: RSS matrix $\mathbf{P_m}$ for each channel m
(1) Use one node to choose a channel m with less or constant noise.
(2) Utilize the pathloss model to compute an estimation of RSS matrix $\widehat{\mathbf{P}}$;
(3) Decompose $\widehat{\mathbf{P}}$ in to form of $\mathbf{U\Sigma V}^\top$, using the singular value decomposition method.
(4) Get the sparsity of $\widehat{\mathbf{P}}$ by omitting the singular value smaller than δ;
(5) Generate the measurement matrix $\mathbf{\Phi} = [\mathbf{I} \mid \mathbf{\Phi_2}]$ following the LDPC matrix with the predetermined time slot $T = K\log(N/K)$;
(6) Distribute the measurement plan to each node i with ith column of $\mathbf{\Phi}$;
(7) Each node sends a test signal following the measurement plan.
(8) The RSS in each node i as r_i is aggregating to the central server. Denote as $\mathbf{R} = \{r_1^\top, r_2^\top, \ldots, r_N^\top\}$;
(9) Solve the problem $\mathbf{RV} = \mathbf{\Phi UP}'$ with minimizing the $\|\mathbf{P}'\|_*$ by linear programming;
(10) Get the RSS matrix $\mathbf{P_m} = \mathbf{UP}'\mathbf{V}^\top$

Algorithm 1: CS-based solution.

extension to mitigate the harm due to background noises. In practice, the RSS is not all from the nodes inside network; the background noise is usually sporadic and affects almost all nodes.

The measured result could be divided into two parts: the noise matrix and the RSS matrix. Formally put,

$$\mathbf{P} = \mathbf{P_r} + \mathbf{P_n}. \tag{19}$$

We could arrange these two matrices as a new one $\mathbf{P}' = [\mathbf{P_r} \mid \mathbf{P_n}]^\top$ and form a new measurement matrix as $\mathbf{\Phi}' = [\mathbf{\Phi} \mid \mathbf{I}]$, which is a $T \times 2N$ matrix. Here I is $T \times N$ identity matrix. Then a CS solution in Algorithm 1 could be performed in the linear system of

$$\mathbf{R} = \mathbf{\Phi}'\mathbf{P}'. \tag{20}$$

Note that the background noise usually affects a large area of network. The nodes in network could measure the average background noise for each part and derive a low-rank matrix P_n. In this way, this extension could have the noise and RSS matrix distinguished in a rather low cost.

3.4. Distributed Scheme. The CS solution could be easily transformed into a distributed solution. As each node could easily collect the RSS from the nodes in the networks, assuming the measurement matrix is known, the measurement process in the node i could be formalized as

$$y_i = \mathbf{\Phi} x_i, \tag{21}$$

where y_i is a T vector and x_i is an N vector.

Thus, this is a classical compressive sensing form, as long as x_i is sparse. To make sure the algorithm is completely distributed, each node should generate the measurement matrix and the representation matrix independently (Algorithm 2).

For the measurement matrix, we can still use the LDPC matrix. As the measurement matrix is a global measurement schedule, it must be synchronized. The random generation could be synchronized for all nodes if they use the same random seed. This seed could be the global clock or the others that have already been synchronized.

For the representation basis, each node could have their own basis without affecting the recovery results. We apply the standard representation basis here. Specifically, we use Discrete Cosine Transform Basis (DCT) [20] here. The RSS vector generally becomes sparse after the DCT transformation.

With the synchronized measurement matrix $\mathbf{\Phi}$ and identical representation basis $\mathbf{\Psi}_i$ we can easily solve l_1 minimization problem:

$$\min_{x_i \in \mathbb{R}^N} \quad \left\| y_i - \mathbf{\Phi\Psi}_i x_i \right\|_{l_1}. \tag{22}$$

3.5. Implementation Issues. To implement our algorithm to the real deployed wireless networks, we have to consider some implementation issues, especially those with the distributed algorithms. Consequently, in this subsection, we discuss two major issues, time synchronization and measurement matrix generation.

3.5.1. Time Synchronization. Time synchronization has been constantly drawing research attentions ever since the distributed systems to wireless sensor networks and the mac protocol in wireless networks. Our algorithm mainly relies on the network synchronization in two folds. The first one is that our measurements are performed in a slotted fashion, such that the clock must be synchronized among all the nodes. The second one is that the measurement matrices are generated by the same random function with the same seed. This seed is usually the time clock. Thus, the synchronization is critical for our algorithm. The synchronization frequency and the time cost is the major overhead in real implementation. These costs could be alleviated by selecting suitable synchronization scheme. One suitable solution for our situation is the diffusion algorithm proposed in literature [21], which incurs low energy cost and low synchronization delay.

Another critical concern in the synchronization is how to discover the inconsistency of the measurement matrix between each node. This situation is mainly caused by the clock shift. In our algorithm the inconsistency could be easily discovered by neighbor exchange of the measurement matrix or certain form of checksum, for example, the sum of all elements in measurement matrix.

Require: Number of Nodes, Synchronized global
Ensure: RSS matrix $\mathbf{P_m}$ for each channel m
(1) Use one node to choose a channel m with less or constant noise.
(2) Utilize the pathloss model to compute an estimation of RSS matrix $\widehat{\mathbf{P}}$;
(3) Get the sparsity K of $\widehat{\mathbf{P}}$ by omitting the singular value smaller than δ;
(4) Each node generates the measurement matrix $\mathbf{\Phi} = [\mathbf{I} \mid \mathbf{\Phi_2}]$ following the LDPC matrix with the predetermined time slot $T = K\log(N/K)$ under the random function with the same seed as the global time.
(5) Each node sends a test signal following the measurement plan.
(6) Each node i generate their own representation basis $\mathbf{\Psi}_i$ using DCT.
(7) Each node solve

$$\min_{x_i \in \mathbb{R}^N} \quad \|y_i - \mathbf{\Phi\Psi}_i x_i\|_{l_1}, \tag{$*$}$$

and get the RSS in node i.
(8) Aggregate the x_i from each node and get the final RSS matrix.

Algorithm 2: Distributed CS-based solution.

3.5.2. Measurement Matrix Generation. As aforementioned, the measurement matrix is generated complying with LDPC matrix. Traditional scheme is to generate a binary matrix with entries complying to Bernoulli distribution with success probability p. Both schemes have to resolve the challenge of how to transmit the parameters, for example, p in Bernoulli distribution or d in LDPC matrix. Such information synchronization process could be performed with the clock synchronization concurrently by combining the transmission of the clock and the parameters.

4. Evaluation

In this section, we analyze the performance of the proposed solutions with experiments. We first present the experimental methodology and simulation settings; then, we discuss the numerical results.

4.1. Simulation Settings. Our simulations are based on the data collected from the SWIM platform [22]. It consists of 10 wireless nodes, which are capable of running in 802.11a/b/g mode. We collected the data of the RSSI (Receiving Signal Strength Index) of the beacons from each AP. Specifically, we activated one node at a time, while each AP was tuned to 11 different channels sequentially. Then, we walked to 25 different locations (including the locations of 10 AP), to collect the 50 different beacon messages from one AP in each channel. The RSSI, AP ID, and channel ID were recorded.

We generated several experimental scenarios from this data set. The experimental scenarios consist of 5, 10, 15, 20 different nodes with the RSS between them. We also set the total operating spectrum to approximately 2.4 GHz, with 11 channels of 20 MHz, which are the general settings in IEEE 802.11g. 200 scenarios are generated to perform a statistical performance comparison evaluation. The throughput of the whole network is computed using the algorithm in [23].

The benchmarks used to quantify our solutions are how the RSS metrics obtained via the solutions impact the performance of the throughput optimization. The accuracy is quantified using the MPE (Mean Percentage Error), which is formally defined as $(1/N)((p'_{ij} - p_{ij})/p_{ij})$. Here, p'_{ij} is the estimated RSS for (i, j).

Regarding the CS-based solution, we mainly examine how this method improves the performance of SINR-based optimization. We also provide a comparison of the performance between the CS solution and the solution proposed in [14] versus network density. Regarding the network density, we prefer a benchmark that directly connects the network density and the sparsity of RSS matrix. A metric called *average interference ratio* is used, which stands for how many portions of the network are interfered by a single node. Assuming that $\mathcal{S}_i$ stands for the node set interfered by node i, we get the formal definition of average interference ratio as $\sum_i \|\mathcal{S}_i\|/N^2$. In the implementation, we let the row number T of measurement matrix be equal to $K\log(N/K)$. We also examine the performance change of the introduction of the LDPC code generated measurement matrix.

4.2. Experimenting Results. First, we compare the fundamental performance of CS-based solution and model-based solution [14] in terms of their improvement to the SINR-based throughput optimization algorithm. The result in Figure 5 shows that these two solutions performed almost in the same level. They are all close to the optimal results with exhaustive measurement.

We also examine how the recovery is affected by the introduction of LDPC-based measurement matrix. As aforementioned, the LDPC code has better RIP comparing to the Gaussian matrix. It also has the advantage of less measurement cost, as there are limited nonzero elements per row. The latter advantage could be easily examined with mathematical computation. Thus, we do not provide the numerical result here. The experiment examines the recovery advantage of LDPC-based measurement matrix as shown in Figure 6, where the line "CS-w-LDPC" is indeed better than another one. This also implies that better recovery accuracy could help to improve the optimization results.

We further compare these two solutions in different network densities. As mentioned above, dense networks give rise to a RSS matrix with higher rank, which in turn will

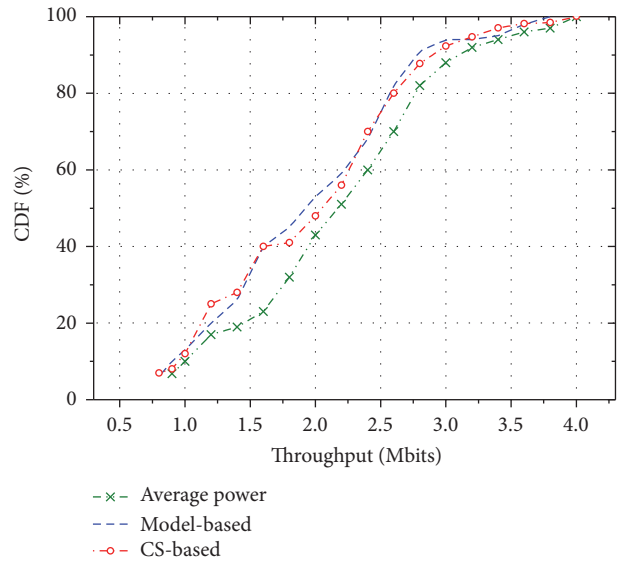


Figure 5: The CDF of throughput under model-based and CS-based solutions.

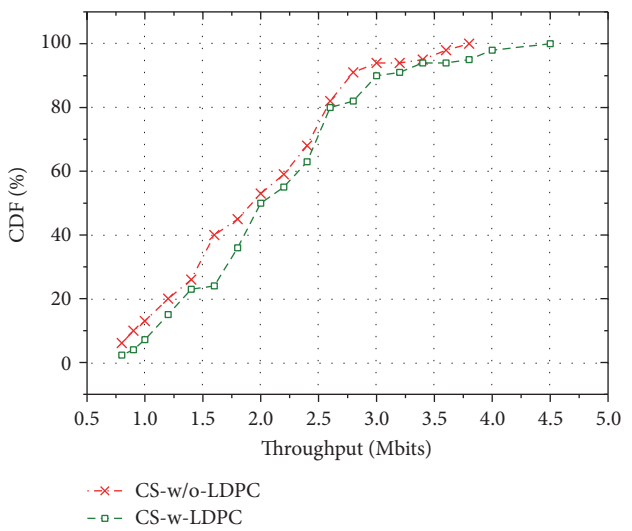


Figure 6: The recovery advantage of LDPC-based measurement matrix.

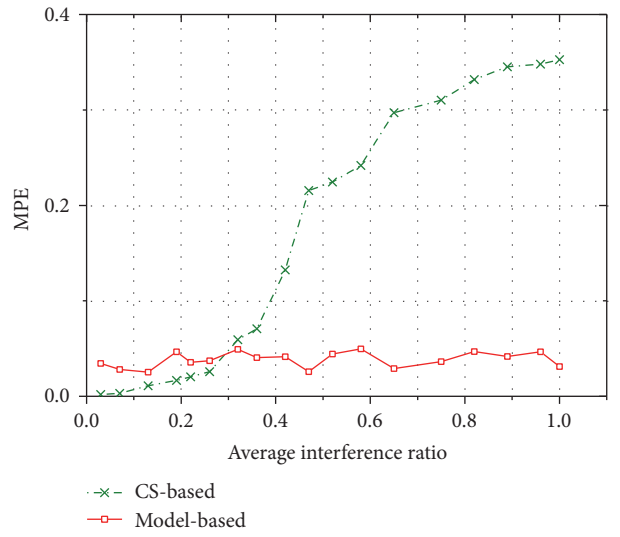


Figure 7: The comparison of model-based and CS-based solutions in different network densities.

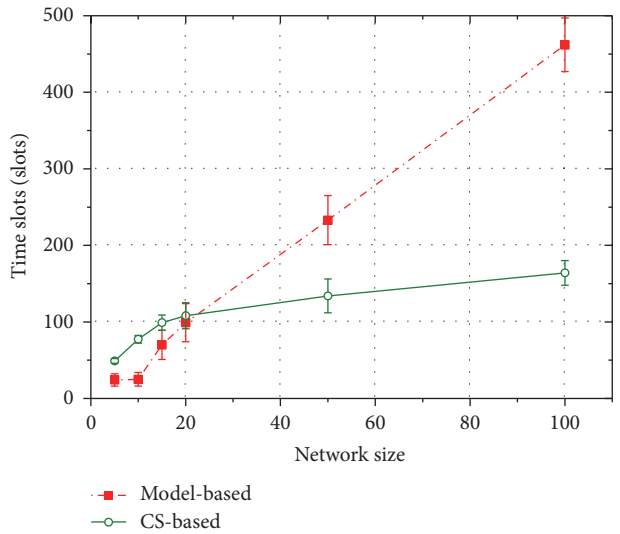


Figure 8: The total time slots used by model-based and CS-based solution versus different network sizes.

compromise the accuracy. This is illustrated using a MSE versus nodes density graph. In Figure 7, the MPE of the CS-based solution increases with the average interference ratio, while the model-based solution performs stably at different network densities. When the average interference ratio is smaller than 0.3 (which also means 70% of the RSS matrices are zero entries), the CS-based solution outperforms the model-based solution. When the whole network is a single-hop network, the MPE of the CS-based solution increases to as high as 36%. Thus, the model-based solution is more accurate in the networks whose sending power could not be tuned to reach less than 30% average interference ratio.

Finally, we examine the time cost of CS-based solution in different network size. The result is shown in Figure 8. Note that, to illustrate the advantage of the CS-based solution in a large network size, we generate several 50-nodes and 100-nodes scenarios. The total channel number is set to 11. It is clear that the finishing time of the model-based solution grows linearly with network size (except the first two points where the network sizes are small than number of channels), while the CS-based solution grows in logarithm order. The total measurement process for networks with less than 100 nodes takes less than 200 time slots. Today's 802.11 on-self products claim 200~500 μs per slot. Thus, every round of measurement with accuracy control takes less than one second.

In summary, CS-based solution performed superiorly in the scenario of low density and large network size in terms of time efficiency.

5. Related Work

5.1. RSS Estimation. In [24], the authors considered a network where the path loss between a few low-cost sensors was measured and stored for future use. They proposed an algorithm that employs interpolation techniques to estimate the path loss between a sensor and any arbitrary point in the network. In [25], a PLE estimator based on the method of least squares was introduced in the design of an efficient handover algorithm. Estimation based on a known internode distance probability scheduling was discussed in [26]. The authors assumed that the distance distribution between two neighboring nodes, i and j, is known or can be determined easily. Regarding the accuracy control method, in [27], the authors provided a framework to control the accuracy for the measure of the SINR-PRR relation. Our previous work [14] focused on using the path loss model to improve SINR-based optimization in wireless networks.

These works endeavour to make more accurate signal attenuation models in a single transmission. But, none of them focus on making a group measurement on all potential links and all channels, which are crucial for the throughput optimization in wireless networks.

5.2. Applications of Compressive Sensing. Compressive sensing theory was firstly introduced to recover sparse signal with less sampling [11, 13], which was latterly found to be useful in compressive data gathering in sensor networks. Compared with the conventional paradigm, CS-based data compression shifts most computations from the encoder to the decoder, which makes it a perfect fit for in-network data processing in wireless sensor networks. The data gathering problem studied immediate data transmission from sensor nodes to a distant base station after data collection. In a single-hop network, compressive wireless sensing (CWS) [28] was shown to be able to reduce the latency of data gathering by delivering linear projections of sensor readings through synchronized amplitude-modulated analog transmissions. Luo et al. [29] explored the compressive data transmission and decoding method, which provided a constant energy consumption scheme. Compressive sensing was also introduced to solve the traffic matrix derivation and interpolation problem [30] and path reconstruction in WSNs [31]. There are some other applications that use the compressive sensing in dealing with cloud media [32]. It mainly considered the changeable property of the wireless network between the media cloud and users and proposed a novel significance-evaluation method for video frames based on CS. Compressive sensing is also applied in the vehicular infotainment system [33]. Our previous paper [34] has explored the efficient RSS measurement algorithm in general wireless networks.

Our method is different from the above method in that our method is used in a distributed fashion and we proposed a new measurement matrix based on LDPC and its variant, which could reduce the measurement cost and time cost greater than the traditional way. Comparing to our previous work, we further consider the feature of the IoT systems and propose distributed algorithm with improved energy efficient measurement matrix.

6. Conclusion

The efficiency and accuracy of the RSS measurement in the wireless networks are of great importance for throughput optimization, localization, and wireless sensing in the Internet of Things. Traditional efficient RSS measurement adopt a "measure a few, predict many" fashion with calibrating the parameters in the propagation models. However, we claim that these kinds of methods are not good enough as they miss the chance of simultaneous measurement and controllable partial measurement, which are all achieved with the compressive sensing-based solution we proposed. With CS-based solution, the whole measurement process could be finished in time of $\mathcal{O}(\log N)$, rather than $\mathcal{O}(N^2)$ in the traditional way. Furthermore, an accuracy control tool is also provided to make balance between the accuracy and efficiency in a quantitative way. Experiments with real data trace from our platform have proved the efficiency of our solution.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported in part by the National NSF of China (under Grant nos. 61602238, 61672278, 61672283, and 61373128), the Key Project of Jiangsu Research Program Grant (BK20160805), and the China Postdoctoral Science Foundation (no. 2016M590451).

References

[1] R. Ahlswede, N. Cai, S. R. Li, and R. W. Yeung, "Network information flow," *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, vol. 46, no. 4, pp. 1204–1216, 2000.

[2] S. Katti, H. Rahul, W. Hu, D. Katabi, M. Médard, and J. Crowcroft, "XORs in the air: Practical wireless network coding," pp. 243–254.

[3] I. F. Akyildiz, W. Lee, M. C. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: a survey," *Computer Networks*, vol. 50, no. 13, pp. 2127–2159, 2006.

[4] N. Nie and C. Comaniciu, "Adaptive channel allocation spectrum etiquette for cognitive radio networks," in *Proceedings of the 1st IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN '05)*, pp. 269–278, Baltimore, Md, USA, November 2005.

[5] Y. Zhao, J. Wu, F. Li, and S. Lu, "On maximizing the lifetime of wireless sensor networks using virtual backbone scheduling," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 8, pp. 1528–1535, 2012.

[6] Z. Yang, C. Wu, and Y. Liu, "Locating in fingerprint space: wireless indoor localization with little human intervention," in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking (Mobicom '12)*, pp. 269–280, August 2012.

[7] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of WiFi signal based human activity recognition," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking, MobiCom 2015*, pp. 65–76, Paris, France, September 2015.

[8] O. Goussevskaia, Y. A. Oswald, and R. Wattenhofer, "Complexity in geometric SINR," in *Proceedings of the 8th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc '07)*, pp. 100–109, September 2007.

[9] R. Bhatia and M. Kodialam, "On power efficient communication over multi-hop wireless networks: joint routing, scheduling and power control," in *Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '04)*, vol. 2, pp. 1457–1466, Hong Kong, March 2004.

[10] Y. Shi, Y. T. Hou, S. Kompella, and H. D. Sherali, "Maximizing capacity in multihop cognitive radio networks under the SINR model," *IEEE Transactions on Mobile Computing*, vol. 10, no. 7, pp. 954–967, 2011.

[11] E. J. Candes and T. Tao, "Decoding by linear programming," *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[12] D. L. Donoho, "Compressed sensing," *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[13] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: universal encoding strategies?" *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[14] Y. Zhao, J. Wu, and S. Lu, "Efficient SINR estimating with accuracy control in large scale cognitive radio networks," in *Proceedings of the 2011 17th IEEE International Conference on Parallel and Distributed Systems, ICPADS 2011*, pp. 549–556, Tainan, Taiwan, December 2011.

[15] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling: a sensing/sampling paradigm that goes against the common knowledge in data acquisition," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[16] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation. An International Journal for Approximations and Expansions*, vol. 28, no. 3, pp. 253–263, 2008.

[17] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathématique. Académie des Sciences. Paris*, vol. 346, no. 9-10, pp. 589–592, 2008.

[18] W. Lu, W. Li, K. Kpalma, and J. Ronsin, "Near-optimal binary compressed sensing matrix," https://arxiv.org/abs/1304.4071.

[19] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[20] S. A. Khayam, *The Discrete Cosine Transform (Dct): Theory and Application*, Michigan State University, 2003.

[21] Q. Li and D. Rus, "Global clock synchronization in sensor networks," *IEEE Transactions on Computers*, vol. 55, no. 2, pp. 214–226, 2006.

[22] http://cs.nju.edu.cn/lwz/swim/swim.html.

[23] Y. Zhao, J. Wu, and S. Lu, "Throughput maximization in cognitive radio based wireless mesh networks," in *Proceedings of the 2011 IEEE Military Communications Conference, MILCOM 2011*, pp. 260–265, Baltimore, MD, USA, November 2011.

[24] X. Zhao, L. Razoumov, and L. J. Greenstein, "Path loss estimation algorithms and results for rf sensor networks," in *Proceedings of IEEE Vehicular Technology Conference (VTC)*, 2004.

[25] N. Benvenuto and F. Santucci, "A least squares path-loss estimation approach to handover algorithms," *IEEE Transactions on Vehicular Technology*, vol. 48, no. 2, pp. 437–447, 1999.

[26] G. Mao, B. D. O. Anderson, and B. Fidan, "Path loss exponent estimation for wireless sensor network localization," *Computer Networks*, vol. 51, no. 10, pp. 2467–2483, 2007.

[27] J. Huang, S. Liu, G. Xing, H. Zhang, J. Wang, and L. Huang, "Accuracy-aware interference modeling and measurement in wireless sensor networks," in *Proceedings of the 31st International Conference on Distributed Computing Systems (ICDCS '11)*, pp. 172–181, Minneapolis, Minn, USA, July 2011.

[28] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak, "Compressive wireless sensing," in *Proceedings of the Fifth International Conference on Information Processing in Sensor Networks, IPSN '06*, pp. 134–142, Nashville, Tenn, USA, April 2006.

[29] C. Luo, F. Wu, J. Sun, and C. W. Chen, "Efficient measurement generation and pervasive sparsity for compressive data gathering," *IEEE Transactions on Wireless Communications*, vol. 9, no. 12, pp. 3728–3738, 2010.

[30] M. Roughan, Y. Zhang, W. Willinger, and L. Qiu, "Spatio-temporal compressive sensing and internet traffic matrices (Extended Version)," *IEEE/ACM Transactions on Networking*, vol. 20, no. 3, pp. 662–676, 2012.

[31] Z. Liu, Z. Li, M. Li, W. Xing, and D. Lu, "Path reconstruction in dynamic wireless sensor networks using compressive sensing," in *Proceedings of the 15th ACM International Symposium on Mobile Ad Hoc Networking and Computing, MobiHoc 2014*, pp. 297–306, Philadelphia, Pa, USA, August 2014.

[32] J. Guo, B. Song, and X. Du, "Significance Evaluation of Video Data over Media Cloud Based on Compressed Sensing," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1297–1304, 2016.

[33] J. Guo, B. Song, Y. He, F. R. Yu, and M. Sookhak, "A survey on compressed sensing in vehicular infotainment systems," *IEEE Communications Surveys & Tutorials*, vol. 99, 2017.

[34] Y. Zhao, W. Li, J. Wu, and S. Lu, "Efficient RSS measurement in wireless networks based on compressive sensing," in *Proceedings of the 34th IEEE International Performance Computing and Communications Conference, IPCCC 2015*, Nanjing, China, December 2015.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 8949430, 8 pages
https://doi.org/10.1155/2017/8949430

Research Article

Collaborative Covert Communication Design Based on Lattice Reduction Aided Multiple User Detection Method

Baoguo Yu,[1,2] Yachuan Bao,[1,2] Haitao Wei,[1,2] Xin Huang,[3] and Yuquan Shu[1,2]

[1]*State Key Laboratory of Satellite Navigation System and Equipment Technology, Shijiazhuang, China*
[2]*The 54th Research Institute of CETC, Shijiazhuang, China*
[3]*Northwestern Polytechnical University, Xi'an, China*

Correspondence should be addressed to Yachuan Bao; baoyachuan@126.com

Received 13 April 2017; Accepted 28 June 2017; Published 6 August 2017

Academic Editor: Xiaoqiang Ma

Spread spectrum communication is a typical scheme for covert communication because of its low detectability and antijam characteristic. However, the associated design concerns multiple factors, such as cochannel multiple access interference (MAI) and spread spectrum gain. In this paper, the lattice reduction theory is applied to MAI cancellation of spread spectrum communication and a novel lattice reduction aided multiple user detection method is proposed. The near maximum likelihood (ML) performance of MAI resistance is verified by simulation and theoretical analysis. The superiority of detection performance in strong MAI scenarios is especially addressed. Based on the algorithm, a collaborative covert communication system design is proposed. Low-power covert signals can be transmitted at a higher bit rate with the same coverage as more high-power cochannel signals. The covert transmission performance can be improved significantly compared to traditional designs.

1. Introduction

For the purposes of information transmission security, covert communication is applied in many special scenarios, including military and national security applications. The development of covert communication systems shows a trend for diversity, where different designs are tailored to different applications. Covert communication systems can be divided into two main groups: covert information transmission and covert signal transmission. The covert transmission of information can be satisfied by securing both the information source and the communication protocol encryption. A typical schema includes information transmission with image and video compression [1] and covert P2P channels over the Internet [2], among others. The covert transmission of signals is realized by signal or transmitter-receiver mode design with low detectability. Typical schemes include designs based on direct antenna modulation (DAM) [3] and spread spectrum (SS) communication.

The most common method for covert communication design based on SS involves reducing the signal power to hide it in noise or transmit the covert signal covered by several high-power signals. An improved method involving frequency and code hopping is proposed in this paper, with an improvement in anticapture performance [4]. SS signals are also designed to transmit coupled with satellite television signals and broadcast signals to achieve covert communication [5].

Transmitting low-power covert signals coupled with high-power signals is an effective way to reduce the detectability of a signal. However, the design will be restricted by multiple access interference (MAI) and wireless link resources. To achieve high imperceptibility, signal power and information rates will be limited.

Research on covert communication based on SS is discussed in this paper. Lattice reduction (LR) is applied to multiple user detection (MUD) of SS communication, and an LR-MUD algorithm is also proposed. Near maximum likelihood (ML) performance of lower power signals is achieved with low complexity in serious MAI scenarios. Based on the LR-MUD algorithm, a novel design of collaborative SS covert communication is proposed. Covert signals

can be transmitted with the coverage of a greater number of high-power signals. Thus, the transmission concealment and robustness of covert communication will be improved significantly.

The remainder of the paper is organized as follows: Section 2 gives the covert communication model based on SS communication, while the traditional MUD method is analyzed in Section 3. The lattice theory and lattice aided MUD algorithm is given in Section 4, followed by simulation experiments of the algorithm in Section 5. The design of a system based on LR-MUD is given in Section 6. Finally, the conclusions are provided in Section 7.

2. Covert Communication Model Based on Spread Spectrum

Consider a covert SS communication system consisting of one low-power covert signal and $N-1$ high-power signals. The received signal at the receiver is

$$r(t) = \sum_{k=1}^{N} A_k(t)\, b_k(t)\, \mathrm{PN}_k(t) + n(t), \tag{1}$$

where A is the signal amplitude, b is the information bit which takes values in the interval $\{-1, +1\}$, and PN is the pseudo noise code used for spread spectrum modulation.

Suppose signal k is the desired signal. Then, the output of the matched filter for signal k is as follows:

$$\begin{aligned} y_k &= \int_0^{T_b} r(t)\, \mathrm{PN}_k(t)\, dt \\ &= \int_0^{T_b} \left[\sum_{k=1}^{N} A_k b_k \mathrm{PN}_k(t) + n(t) \right] \mathrm{PN}_k(t)\, dt \\ &= A_k b_k + \sum_{j=1, j\neq k}^{N} A_j b_j \rho_{jk} + n_k, \end{aligned} \tag{2}$$

where $\rho_{jk} = \int_0^{T_b} \mathrm{PN}_k(t)\mathrm{PN}_j(t)dt$, $n_k = \int_0^{T_b} n(t)\mathrm{PN}_k(t)dt$, and T_b is the time length of the information bit.

The first item is the expected signal, whereas the second item is the correlation sum of the spread spectrum code and other signals, which is the MAI. The third item is the channel noise.

The output of a matched filter of N signals is given below:

$$\begin{aligned} \mathbf{Y} &= [y_1, y_2, \ldots, y_N]^T, \\ \mathbf{Y} &= \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1N} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{N1} & \rho_{N2} & \cdots & \rho_{NN} \end{bmatrix} \begin{bmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_N \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix} \\ &\quad + \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_N \end{bmatrix} = \mathbf{RAb} + \mathbf{n}. \end{aligned} \tag{3}$$

Definite $\mathbf{H} = \mathbf{RA}$ and formula (3) can be written as follows:

$$\mathbf{Y} = \mathbf{Hb} + \mathbf{n}. \tag{4}$$

In traditional detection, the output of a matched filter will be sampled at every bit interval, and the bit will be decided on the basis of the decision threshold:

$$\widehat{\mathbf{b}} = \mathrm{dec}(\mathbf{Y}) = \mathrm{dec}(\mathbf{Hb} + \mathbf{n}), \tag{5}$$

where dec( ) represents the decision value of the bit. Getting the bit directly with the decision of the matched filer output, critical errors could possibly be caused by MAI, especially for the covert low-power signal.

3. Multiple User Detection Method for Spread Spectrum Communication

As an effective method to improve the capacity of a spread spectrum communication system, a MAI suppression method has been developed in depth [6, 7]. The typical method includes two types: multiple user detection (MUD) and an interference cancellation method. The foundation of the MUD method is the calculation of correlation between signals, and the interference is suppressed by the decorrelation process. The traditional MUD method includes the zero forcing (ZF) method and the minimum mean square error (MMSE) method, among many other new methods proposed over the years.

Here, a MUD method aided by a Hopfield neural network is proposed [8]. With the method, detection performance can be improved, but the improvement will decrease with an increase in signal number. A weighted orthogonal matched filter method based on quantum signal processing is proposed [9]. With this method, the estimation of noise power is not necessary, but the detection performance is reduced compared to the MMSE method. A blind MUD method based on Schmidt-orthogonalization and subspace-tracking Kalman filtering is also presented, and with this method, the complexity of the blind MUD method is reduced [10]. There are two types of interference cancellation methods: serial interference cancellation (SIC) and parallel interference cancellation (PIC). Based on the correlation computation of signals, the signal interference influence of other signals can be cancelled by the serial or parallel mode. The performance of this kind of method is suboptimal compared with the MUD method. Its performance is limited by the initial detection accuracy of multiple signals. When multiple interference is serious, the platform effect will emerge early and the detection performance will not be ideal [11, 12]. In practical applications, the detection performance can be improved with usage combined with high-gain error correction coding. The benefit from the low computation complexity and flexible processing architecture leads to this interference cancellation method being applied in some satellite communication system designs.

The article is focused on the MUD method. The basic idea of the MUD method is to cancel MAI between signals with

the calculation and transformation of a correlation matrix. Maximum likelihood (ML), ZF, and MMSE methods are typical algorithms applied.

The ML algorithm is the theoretical optimum MUD algorithm. Its principle is shown as follows. Generate a transmit signal traversal set and complete the following operation:

$$\widehat{\mathbf{b}} = \operatorname{argmin}(\mathbf{Y} - \mathbf{Hb}) \quad \mathbf{b} \in \{\mathbf{b}\}. \tag{6}$$

The main process of ZF algorithm involves calculating the inverse matrix G of the correlation matrix $\mathbf{H}$ and then calculating the dot product of $\mathbf{G}$ and the output of the matched filter group. After that, estimation of transmitter signals can be obtained with the following decision [13]:

$$\begin{aligned} \mathbf{G} &= \left(\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{H}^H, \\ \widehat{\mathbf{b}} &= \operatorname{dec}(\mathbf{G}(\mathbf{Hb} + \mathbf{n})) = \mathbf{b} + \mathbf{n}'. \end{aligned} \tag{7}$$

$\mathbf{n}'$ is the noise component after ZF conversion and bit decision.

The calculation complexity of the ZF method is relatively low and scales with the number of signals.

The estimation error covariance matrix is as follows:

$$\boldsymbol{\varphi}_{\mathrm{MMSE}} = E\left\{\left(\widehat{\mathbf{b}} - \mathbf{b}\right)\left(\widehat{\mathbf{b}} - \mathbf{b}\right)^H\right\} = \sigma^2\left(\mathbf{H}^H\mathbf{H}\right)^{-1}. \tag{8}$$

σ^2 is the variance of noise. The influence of noise is always increased with the ZF method, which is why ZF performs poorly with a low signal noise ratio (SNR) scenario. $(\mathbf{H}^H\mathbf{H})^{-1}$ can be defined as the demodulation error coefficient for ZF, which measures how the influence of noise is enlarged by the ZF method.

The basic idea of MMSE is to minimize the mean square error between transmit signal and estimation result. Unlike the ZF method, MMSE takes noise suppression into account, and thus the estimation performance is improved to some extent. The process of MMSE is as follows:

$$\begin{aligned} \mathbf{G}' &= \left(\mathbf{H}^H\mathbf{H} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{H}^H, \\ \widehat{\mathbf{b}} &= \operatorname{dec}\left(\mathbf{G}'(\mathbf{Hb} + \mathbf{n})\right) = \mathbf{b} + \mathbf{n}''. \end{aligned} \tag{9}$$

$\mathbf{n}''$ is the noise component after MMSE conversion and bit decision.

The estimation error covariance matrix of MMSE is

$$\begin{aligned} \boldsymbol{\varphi}_{\mathrm{MMSE}} &= E\left\{\left(\widehat{\mathbf{b}} - \mathbf{b}\right)\left(\widehat{\mathbf{b}} - \mathbf{b}\right)^H\right\} \\ &= \sigma^2\left(\mathbf{H}^H\mathbf{H} + \sigma^2\mathbf{I}\right)^{-1}. \end{aligned} \tag{10}$$

Based on formal analysis, it can be shown that the performance of MUD is influenced by noise and the orthogonality of the correlation matrix. If the correlation matrix is orthogonal, the detection performance of MUD will equal that of the performance of a single signal without MAI. In addition, for special users, it is possible to suppress noise with the transformation of the correlation matrix. Searching for a method, which can improve the orthogonality of the matrix, is an important pathway to fulfill enhanced MAI cancellation.

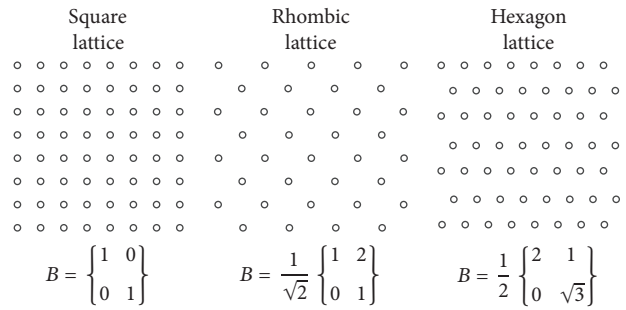


Figure 1: Several typical lattices.

4. Multiple User Detection Methods Based on Lattice Reduction

4.1. Lattice Theory. A lattice is a congregation of scatter points arranged with scheduled rule [14–16]. Any lattice can be generated by a group of linear unrelated vectors. Figure 1 shows several typical types of lattice.

Suppose a group of n dimensional vectors is defined by $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m) \in R_n$. The number of vectors is m and the vectors are linearly independent. The integrated linear combination of the vectors can form an m-dimensional lattice. It can be written as follows:

$$L(\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m) = \sum_{i=1}^{m} t_i b_i, \quad t_i \in Z. \tag{11}$$

Vector $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m)$ is one foundation of the lattice, and $\mathbf{t} = [t_1, t_2, \ldots, t_n]^T$ is the coefficient vector composed of integers. It can be written in matrix mode:

$$L = \{\mathbf{l} = \mathbf{Bt}\}. \tag{12}$$

The three different lattices are generated from the basis below the figures. The dimensionality is decided by the number of base vectors. The dimensionality of the base vector is called the rank of the generated lattice.

The basis of a lattice is diversified. A lattice can be denoted by different bases. An m-dimensional space can be generated by any group of m linearly independent vectors, but not any group of m linearly independent vectors can form a lattice.

There is a fixed relation between the different base vectors of one lattice. For the lattice $L(\mathbf{B})$ generated by vector group $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m)$, all of the vectors transformed from $\mathbf{B}$ by an elementary column transfer from the base vectors of the lattice. The product of several elementary column transfer matrices equals one unimodular transfer matrix $\mathbf{T}$. The elements of $\mathbf{T}$ are complex integers and the determinant of $\mathbf{T}$ equals ± 1. As long as $\mathbf{T}$ is a unimodular matrix, the same lattice can be generated by $\mathbf{B}$ and $\mathbf{BT}$.

$$L(\mathbf{B}) = L(\mathbf{BT}). \tag{13}$$

The shortest base vector, that is, the base vector of shortest length, is a common research object for a lattice. If the shortest base vector cannot be reached, the nearest short base vector is always desired. The basis of a lattice has the

Input: $\mathbf{Q}$, $\mathbf{R}$
Reduction process:
Initialization: $\mathbf{Q}' = \mathbf{Q}$, $\mathbf{R}' = \mathbf{R}$, $k = 2$;
While $k \le n$
for $l = k-1, \ldots, 1$
$b = \mathbf{R}'(l,k)/\mathbf{R}'(l,l)$, b rounded to complex integer μ
if $(\mu \neq 0)$
$\mathbf{R}'(1:l,k) = \mathbf{R}'(1:l,k) - \mu\mathbf{R}'(1:l,l)$
$\mathbf{P}(:,k) = \mathbf{P}(:,k) - \mu\mathbf{P}(:,l)$
end
end
if $(\delta\|\mathbf{R}'(k-1,k-1)\|^2 > \|\mathbf{R}'(k,k)\|^2 + \|\mathbf{R}'(k-1,k)\|^2)$
Interchange column $k-1$ and column k of matrix $\mathbf{R}'$ and $\mathbf{P}$.
Calculate the Givens rotation matrix $\boldsymbol{\theta}$, to make $\mathbf{R}'(k,k-1) = 0$:
$\mathbf{R}'\left(k-1:k,k-1:N_t\right) = \boldsymbol{\theta}\mathbf{R}'\left(k-1:k,k-1:N_t\right)$
$\mathbf{Q}'(:,k-1:k) = \mathbf{Q}'(:,k-1:k)\,\boldsymbol{\theta}^H$
$\left(\boldsymbol{\theta} = \begin{bmatrix} \mathbf{c} & \mathbf{s} \\ -\mathbf{s} & \mathbf{c} \end{bmatrix},\ \mathbf{c} = \dfrac{\mathbf{R}'(k-1,k-1)}{\|\mathbf{R}'(k-1:k,k-1)\|},\ \mathbf{s} = \dfrac{\mathbf{R}'(k,k-1)}{\|\mathbf{R}'(k-1:k,k-1)\|}\right)$
$k = \max\{k-1, 2\}$
else
$k = k+1$
end
If $k > n$, export $\mathbf{R}'$ $\mathbf{Q}'$ and transfer matrix $\mathbf{P}$
Reduction completed.

Algorithm 1: Algorithm LLL.

feature that a shorter basis will be more orthometric. Lattice reduction is the process by which the nearest short base vector is found while the orthogonality of the base vector is improved.

4.2. Lattice Reduction Algorithms. Lattice reduction is a process to obtain the nearest short base vector while the orthogonality of the base vector is improved. Common lattice reduction algorithms include the Lenstra–Lenstra–Lovász (LLL) reduction [17] and the Seysen reduction [18, 19].

To measure the orthogonality of a lattice basis, the degree of orthogonality defect is defined. Suppose $(\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m)$ is one group of base vectors of lattice L, and the base vector matrix is $\mathbf{H}$. The degree of orthogonality defect of $\mathbf{H}$ is defined as

$$\Delta(\mathbf{H}) = \frac{\prod_{m=1}^{M} \|\mathbf{h}_m\|^2}{\det\left(\mathbf{H}^H\mathbf{H}\right)}. \tag{14}$$

In the formula, M is the column number of $\mathbf{H}$, $\mathbf{h}_m$ is the column m of $\mathbf{H}$, and det( ) is the determinant operator. $\Delta(\mathbf{H}) \ge 1$, and only when $\mathbf{H}$ is an orthogonal matrix, $\Delta(\mathbf{H}) = 1$.

Suppose $\mathbf{H}'$ is the matrix generated by an LLL reduction:

$$\mathbf{H}' = \mathbf{HP}, \quad \mathbf{P} \text{ is a unimodular matrix.} \tag{15}$$

If $\mathbf{R}'$ is found via a QR decomposition of $\mathbf{H}'$, it satisfies the following two conditions:

$$\begin{aligned} \left\|r'_{l,k}\right\| &\le \frac{1}{2}\left\|r'_{l,l}\right\|, \quad 1 \le l \le k \le n, \\ \delta\left\|r'_{k-1,k-1}\right\| &\le \left\|r'_{k,k}\right\| + \left\|r'_{k-1,k}\right\|, \quad k = 2, \ldots, n. \end{aligned} \tag{16}$$

Then, $\mathbf{H}'$ is the matrix processed by LLL reduction.

The basic procedure of LLL [20–22] is shown in Algorithm 1

Compared with the LLL reduction, the main difference in the Seysen reduction is a different definition of the degree of orthogonality defect:

$$S(\mathbf{H}) = \sum_{m=1}^{M} \|\mathbf{h}_m\|^2 \left\|\mathbf{h}_m^{\#}\right\|^2. \tag{17}$$

In the formula, $\mathbf{H}$ has M columns. $\mathbf{h}_m$ is the m column of $\mathbf{H}$, and $\mathbf{h}_m^{\#}$ is the m column of the dual matrix $\mathbf{H}^{\#} = ((\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H)^H$ [12, 13].

When $\mathbf{H}$ is an orthogonal matrix,

$$S(\mathbf{H})_{\min} = M. \tag{18}$$

The Seysen algorithm takes both the raw matrix and dual matrix into account; as a result, a better reduction performance can be achieved.

1: Estimation of signals Y is calculated by matched filters.
Signal power, relative delay, and carrier phase difference of signals are estimated.
2: A correlation matrix between signals $\mathbf{H}$ is calculated.
3: Process $\mathbf{H}$ with lattice reduction algorithm (LLL or Seysen).
Conversion matrix $\mathbf{P}$ and optimized matrix $\mathbf{H}' = \mathbf{HP}$ is obtained.
4: The ZF or MMSE method is implemented

$$\widetilde{\mathbf{Z}}_{\mathrm{LLL\_ZF}} = \left(\mathbf{H}'^{H}\mathbf{H}'\right)^{-1}\mathbf{H}'^{H}\mathbf{Y}$$

$$\widetilde{\mathbf{Z}}_{\mathrm{LLL\_MMSE}} = \left(\mathbf{H}'^{H}\mathbf{H}' + \sigma^2\mathbf{P}^{H}\mathbf{P}\right)^{-1}\mathbf{H}'^{H}\mathbf{Y}$$

5: The modified quantization in improved lattice space is taken:

$$\widehat{\mathbf{Z}} = a\left(\left\lceil \frac{1}{a}\widetilde{\mathbf{Z}} - \frac{1}{2}\mathbf{P}^{-1}\mathbf{I}\right\rfloor + \frac{1}{2}\mathbf{P}^{-1}\mathbf{I}\right).$$

a is the power normalization parameter, and $\lceil\ \rfloor$ is the roundness of real value.
6: Final detection result $\widehat{\mathbf{X}}$ is calculated.

$$\widehat{\mathbf{X}} = \mathbf{P}\widehat{\mathbf{Z}}$$

Algorithm 2: Algorithm LR-MUD.

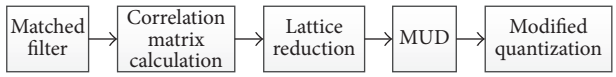


Figure 2: Process of the LR-MUD algorithm.

4.3. Lattice Reduction Aided Multiple User Detection (LR-MUD). With lattice reduction, the correlation matrix $\mathbf{H}$ can be transformed to matrix $\mathbf{H}'$, which has better orthogonality. The analysis of the MUD method in Section 2 shows that with a correlation matrix with better orthogonality the influence of noise can be better suppressed.

The process of lattice reduction aided multiple user detection algorithm is shown in Algorithm 2

The basic process of the algorithm is shown in Figure 2. The signal power and signal delay estimation accuracy is important to the performance of the algorithm. To estimate the signal power in the MAI scene, a synchronization-head can be used to improve the estimation accuracy. The signal delay and carrier phase information can be given by the signal tracking loop. The measurement accuracy can be improved by extending integration time and by using a large correlator design. Related research shows that week spread spectrum signals like GPS can be tracked stably when SNR is lower than −4 dB.

The complexity of ML is $O({N_s}^M)$, where N_s is the modulation order and M is the signal number. The complexity of ML increases with N_s exponentially. The complexity of ZF and MMSE is relatively low, given as $O(M^3)$. The main computation of the MUD method is in the inversion of the correlation matrix. The lattice reduction process is added based on the MUD in the LR-MUD method. The increased computation includes QR decomposition and inversion of matrix $\mathbf{P}$. The increase in complexity is linear, and thus the complexity is still $O(M^3)$. Compared with ML, LR-MUD has superior performance in regard to complexity.

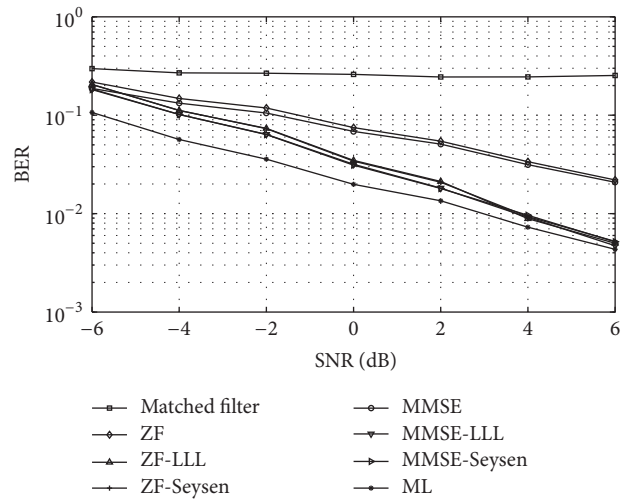


Figure 3: BER of LR-MUD with different SNR scenarios.

5. Simulation and Analysis

In this section, simulations of the LR-MUD method of SS communication are presented. The signal modulation is QPSK, and the spread code sequence is a group of gold codes.

Figure 3 shows a demodulation BER in different SNR scenarios. The signal number is 6, and the spread ratio is 16. Signal power is distributed randomly in the range of 0–6 dB, and the BER of all signals is counted. MAI can be cancelled effectively by the ZF or MMSE method. The BER of the two algorithms is lower than the BER of the basic matched filter method, but the performance is much weaker than that of the ML method. In the figure, ZF-LLL and ZF-Seysen are the tested LR-MUD methods based on a combination of the ZF method using a lattice reduction by LLL and Seysen. Likewise, MMSE-LLL and MMSE-Seysen are the tested LR-MUD methods based on the MMSE method using a lattice reduction by LLL and Seysen. The BER of LR-MUD is

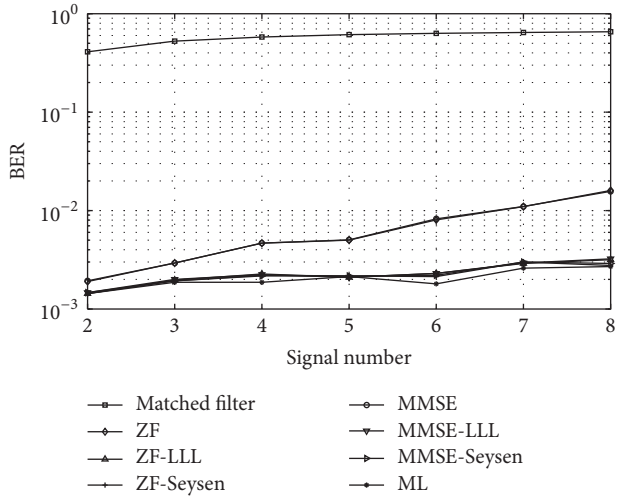


Figure 4: BER of desired signal with different signal numbers.

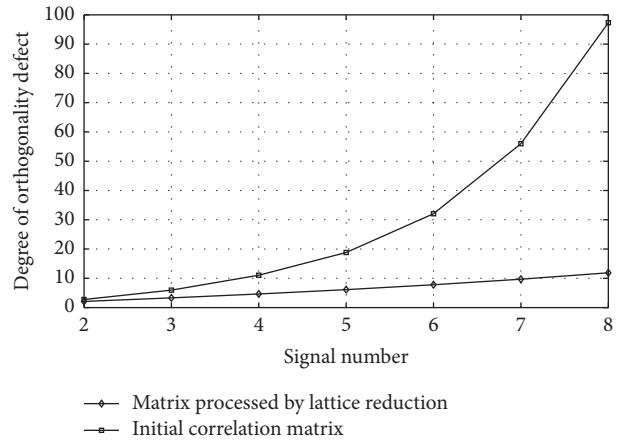


Figure 5: Degree of orthogonality defect of a correlation matrix with different signal numbers.

reduced significantly compared with the MUD methods, and its performance approaches ML with an increase of SNR. The algorithm gain compared to MUD is more than 4 dB.

Figure 4 shows the demodulation BER of a desired signal with different numbers of interference signals. Here, the spread ratio is 16 and the SNR of the covert signal is 0 dB. The power of the other signals is randomly distributed in the 0–20 dB range compared to the desired signal. The BER of the traditional MUD method increases with an increase in signal number. Performance near ML is achieved by LR-MUD and remains steady even with an increase in signal number. LR-MUD shows superiority when subjected to a greater number of interference signals. With the same BER requirements, the number of interfering high-power signals is 7 for LR-MUD, compared to 2 for the traditional MUD method.

Figure 5 shows the variation in the orthogonality of the correlation matrix. Along with the increase of signal number, the degree of orthogonality defect for the initial correlation matrix increases, which results in the demodulation deterioration in the traditional MUD method. In contrast, the degree of orthogonality defect of the matrix processed by LR increases only slightly. Furthermore, this is the explanation for why LR-MUD performs better than the traditional MUD method.

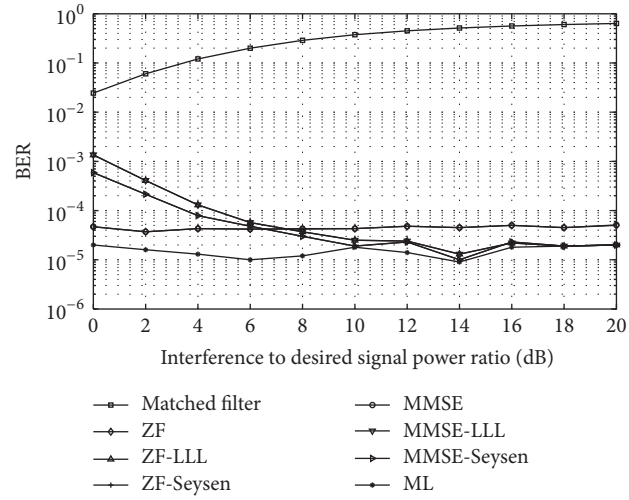


Figure 6: BER of a desired signal with different interference to signal power ratios.

Figure 6 shows the simulation result with different signal power ratios. The spread ratio is 32 and the signal number is 6. SNR of covert signal is 0 dB. The power of the covert signal and noise remain stable, while the power of 5 interference signals increases gradually. The BER of the desired signal is counted. The BER of the MUD method does not change with the increase in power; a fixed gap remains between ML and MUD method. The BER of LR-MUD decreases with the power increase of interference signals, and finally near ML performance is achieved when the interference to signal power ratio is larger than 10 dB. This shows the effect of lattice reduction. When the correlation between signals is good, the algorithm gain of the LR is relatively limited. When the correlation between signals is poor, the algorithm gain of LR will increase. Therefore, LR-MUD is suitable to use in intense near-far scenarios.

Figure 7 shows the variation of the degree of orthogonality defect for a correlation matrix with an increase of interference to signal power ratio. The orthogonality defect of the correlation matrix does not change. With a lattice reduction, the orthogonality of the correlation matrix is improved, and the effect improves further with an increase in interference power.

Figure 8 shows the variation of demodulation error coefficient for a desired signal with an increase of interference to signal power ratio. The change of the demodulation error coefficient is consistent with the BER change shown in Figure 6. With additional increase of interference to signal power ratio, the amplification of noise decreases with the LR-MUD method. This is the main reason why LR-MUD performs better than traditional the MUD method.

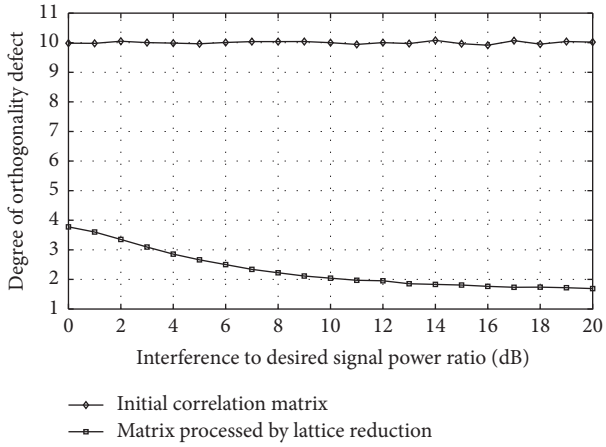


Figure 7: The degree of orthogonality defect for a correlation matrix with different interference to signal power ratios.

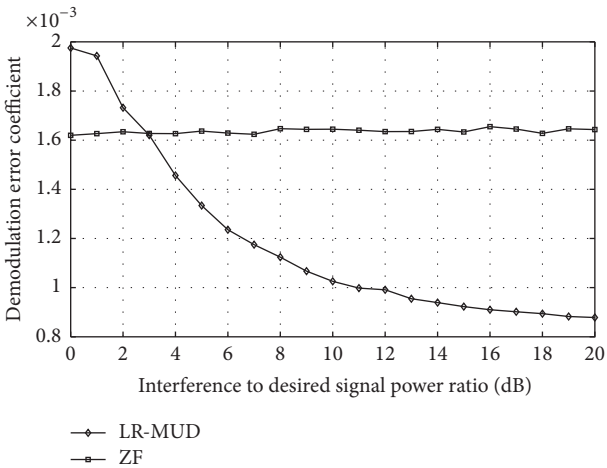


Figure 8: Demodulation error coefficient with different interference to signal power ratios.

6. Design of a Covert Communications System Based on LR-MUD

Because of the excellent performance gain in intense near-far scenarios, LR-MUD is appropriate for implementation in covert communication systems based on SS principles. A block diagram for a covert communications system design is given in Figure 9.

The system includes a covert signal transmitter, covert signal receiver, and common signal transmitter. High-power signals can be transmitted by the covert transmitter as well as a number of common signal transmitters. The covert signal and at least one high-power signal should be transmitted by a covert transmitter synchronously. At the receiver, the capture and tracking of covert signals should be performed using the synchronous high-power signal. All signals will be demodulated with the LR-MUD method.

Figure 10 is the simulation result of the covert transmission performance with different spread ratios. SNR of covert signal is 0 dB and coverage signal number is 5. With

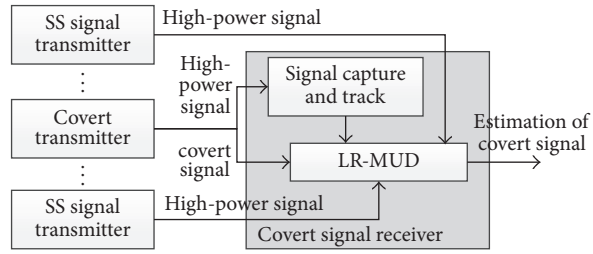


Figure 9: System diagram of a covert communications system based on LR-MUD.

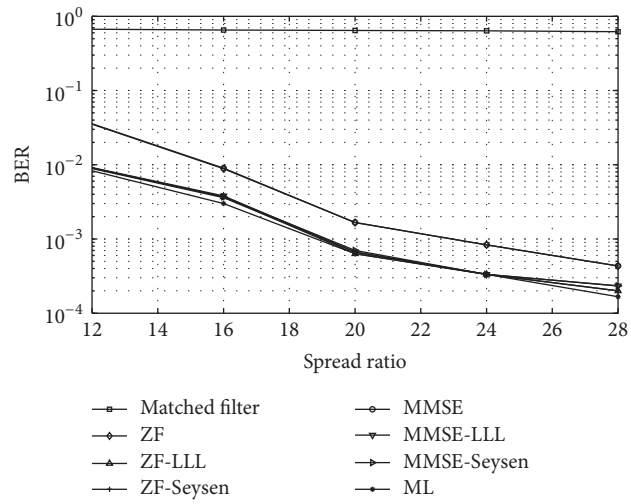


Figure 10: BER of a covert signal with different spread ratios.

the increase of spread ratio, BER decreases with the cost of information rate. The same BER can be achieved by LR-MUD with a lower spread ratio compared to the common MUD method. Higher information rates can be realized by LR-MUD with the same wireless link resource and concealment requirements.

7. Summary and Conclusions

In this paper, lattice reduction theory and related algorithms are applied to the MUD of a spread spectrum communications system, and an algorithm called LR-MUD is presented. Theoretical analyses and simulation results on the LR-MUD method are carried out, showing excellent near-far effect suppression performance. Based on the LR-MUD method, a design of a covert communications system can be feasibly realized. The covert signal can be transmitted at a higher information rate with the coverage of more high-power cochannel signals. Thus, higher transmission rates and concealment performance are achieved using the same link resources.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors acknowledge the National Natural Science Foundation of China (Grant 91638203) and the National Key Research and Development Program of China (Grant 2016YFB0502102).

References

[1] Y. Cao, H. Zhang, X. Zhao, and H. Yu, "Covert communication by compressed videos exploiting the uncertainty of motion estimation," *IEEE Communications Letters*, vol. 19, no. 2, pp. 203–206, 2015.

[2] M. Cunche, M.-A. Kaafar, and R. Boreli, "Asynchronous covert communication using bittorrent trackers," in *Proceedings of the 16th IEEE International Conference on High Performance Computing and Communications, HPCC 2014, 11th IEEE International Conference on Embedded Software and Systems, ICESS 2014 and 6th International Symposium on Cyberspace Safety and Security, CSS 2014*, pp. 827–830, fra, August 2014.

[3] H. Shi and A. Tennant, "Covert communication using a directly modulated array transmitter," in *Proceedings of the 8th European Conference on Antennas and Propagation, EuCAP 2014*, pp. 352–354, nld, April 2014.

[4] L. Li, J. Zheng, and L. Zheng, "The Chaotic Code-hopping Spread Spectrum Communication System[J]," *Science Technology and Engineering*, vol. 13, no. 30, pp. 176–180, 2014.

[5] X. Zhao, X. Ma, and W. Qu, "Analysis of the parasitic small signal coupling with RDSS RF signal [J]," *Telecommunication Engineering*, vol. 49, no. 8, pp. 40–44, 2009.

[6] Y. Hou, M. Li, X. Yuan, Y. T. Hou, and W. Lou, "Cooperative Interference Mitigation for Heterogeneous Multi-Hop Wireless Networks Coexistence," *IEEE Transactions on Wireless Communications*, vol. 15, no. 8, pp. 5328–5340, 2016.

[7] Q. Zhou and X. Ma, "Receiver designs for differential UWB systems with multiple access interference," *IEEE Transactions on Communications*, vol. 62, no. 1, pp. 126–134, 2014.

[8] G. I. Kechriotis and E. S. Manolakos, "Hopfield neural network implementation of the optimal CDMA multiuser detector," *IEEE Transactions on Neural Networks*, vol. 7, no. 1, pp. 131–141, 1996.

[9] S.-N. Shi, Y. Shang, and Q.-L. Liang, "A novel linear multi-user detector," *Acta Electronica Sinica*, vol. 35, no. 3, pp. 426–429, 2007.

[10] Y. Yu, J. Li, Z. Wang, S. Wang, and H. Zhang, "Blind multiuser detection in MC-CDMA: Schmidt-orthogonalization and subspace tracking Kalman filtering," in *Proceedings of the 2011 3rd International Conference on Communications and Mobile Computing, CMC 2011*, pp. 375–380, chn, April 2011.

[11] Z. Xie, R. T. Short, and C. K. Rushforth, "A Family of Suboptimum Detectors for Coherent Multiuser Communications," *IEEE Journal on Selected Areas in Communications*, vol. 8, no. 4, pp. 683–690, 1990.

[12] W. Zheng, J. Li, Y. Luo, J. Chen, and J. Wu, "Multi-user interference pre-cancellation for downlink signals of multi-beam satellite system," in *Proceedings of the 2013 3rd International Conference on Consumer Electronics, Communications and Networks, CECNet 2013*, pp. 415–418, chn, November 2013.

[13] Y. Bao and B. Yu, "A MAI cancellation algorithm with near ML performance," in *Proceedings of the IEEE International Conference on Communication Software and Networks, ICCSN 2015*, pp. 196–200, chn, June 2015.

[14] M. R. Bremner, *Lattice Basis Reduction: An introduction to the LLL algorithm and its applications*, vol. 300, CRC Press, Inc., Boca Raton, FL, USA, 2012.

[15] S. M. Dias and N. J. Vieira, "Concept lattices reduction: Definition, analysis and classification," *Expert Systems with Applications*, vol. 42, no. 20, pp. 7084–7097, 2015.

[16] B. A. Lamacchia, *Basis reduction algorithms and subset sum problems [D]. [Master's dissertation]*, Massachusetts Inst Technol, 1991.

[17] A. K. Lenstra, J. Lenstra, and L. Lovász, "Factoring polynomials with rational coefficients," *Mathematische Annalen*, vol. 261, no. 4, pp. 515–534, 1982.

[18] P. Q. Nguyen and J. Stern, "Lattice reduction in cryptology: an update," in *Algorithmic number theory (Leiden, 2000)*, vol. 1838 of *Lecture Notes in Comput. Sci.*, pp. 85–112, Springer, Berlin, 2000.

[19] C.-P. Schnorr and M. Euchner, "Lattice basis reduction: improved practical algorithms and solving subset sum problems," *Mathematical Programming*, vol. 66, no. 2, Ser. A, pp. 181–199, 1994.

[20] P. Q. Nguyen and J. Stern, "Lattice Reduction in Cryptology: An Update [C]," *Lecture Notes in Computer Science*, no. 4, pp. 85–112, 1838.

[21] X. Ma and W. Zhang, "Performance analysis for MIMO systems with lattice-reduction aided linear equalization," *IEEE Transactions on Communications*, vol. 56, no. 2, pp. 309–318, 2008.

[22] L. G. Barbero, T. Ratnarajah, and C. Cowan, "A comparison of complex lattice reduction algorithms for MIMO detection," in *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 2705–2708, usa, April 2008.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 1858532, 15 pages
https://doi.org/10.1155/2017/1858532

Research Article

Energy Harvesting for Internet of Things with Heterogeneous Users

Desheng Wang,[1] Haizhen Liu,[1] Xiaoqiang Ma,[1] Jun Wang,[1] Yanrong Peng,[1] and Yanyan Wu[2]

[1]*School of EIC, Huazhong University of Science and Technology, Wuhan, China*
[2]*School of Public Management, South-Central University for Nationalities, Wuhan, China*

Correspondence should be addressed to Xiaoqiang Ma; maxiaoqiang@hust.edu.cn

Received 17 March 2017; Accepted 11 June 2017; Published 30 July 2017

Academic Editor: Pierre-Martin Tardif

We study the energy harvesting problem in the Internet of Things with heterogeneous users, where there are three types of single-antenna users: *ID users* that only receive information, *EH users* that can only receive energy, and *ID/EH users* that receive information and energy simultaneously from a multiantenna base station via power splitting. We aim to maximize the minimum signal-to-interference-plus-noise ratio (SINR) of the ID users and ID/EH users by jointly designing the power allocation at the transmitter and the power splitting strategy at the ID/EH receivers under the maximum transmit power and the minimum energy harvesting constraints. Specifically, we first apply the semidefinite relaxation (SDR), zero-forcing (ZF), and maximum ratio transmission (MRT) techniques to solve the nonconvex problems. We then apply the zero-forcing dirty paper coding (ZF-DPC) technique to eliminate the multiuser interference and derive the closed-form optimal solution. Numerical results show that ZF-DPC provides higher achievable minimum SINR than SDR and ZF in most cases.

1. Introduction

Recently, the Internet of Things (IoT) [1–3] has been invading various industries. Given the fact that IoT normally consists of resource-constrained devices and relies on wireless communication for data transmission, energy efficiency is an important issue. For example, in wireless sensor networks (WSNs), a key component and enabler of IoT, the sensor nodes are normally powered by batteries that have very limited lifetime. The network lifetime can be extended by replacing or recharging the batteries, which, however, are inconvenient, costly, and environmentally unfriendly [4, 5]. As an alternative solution to prolonging the network lifetime, energy harvesting [6] has become an appealing solution that potentially provides unlimited power supply to wireless networks by scavenging energy from the environment. The radio frequency (RF) signals emitted by ambient transmitters are a viable source for wireless energy harvesting (EH). As an emerging energy harvesting technique, simultaneous wireless information and power transfer (SWIPT) [7] has drawn an upsurge of interests, where RF signals are used to charge users' devices wirelessly. Through power splitting (PS) [8, 9], users are provided with information decoding (ID) and energy harvesting simultaneously, which brings great convenience.

How to get the trade-off between the achievable information and energy harvesting is still a challenging problem, and various techniques have been developed, for example, zero-forcing (ZF), maximum ratio transmission (MRT), and semidefinite relaxation (SDR) [10–14]. It is worth noting that most existing studies only consider *ID/EH users* that can receive information and energy simultaneously. Yet, in practice, there are still many *ID users* that can only transmit/receive information and do not have the ability to harvest energy as well as *EH users* that only harvest energy from RF signals, which should be also considered in a unified framework. Hence, in this paper, we, for the first time, consider the coexistence of all the three types of users under multiuser MISO broadcast channels. Our aim is to maximize

the minimum SINR of all users except for the EH users by jointly designing the power allocations at the transmitter and the PS ratios for ID/EH users who need to harvest energy under both the maximum transmit power and the minimum EH constraints.

In the remainder of this paper, we first present the system model and problem formulation and prove the feasibility of the problem. We then apply three traditional techniques, namely, semidefinite relaxation (SDR), zero-forcing (ZF), and maximum ratio transmission (MRT), to solve the problem. Moreover, it is well known that dirty paper coding (DPC), a nonlinear precoding scheme, can precancel noncausal interference without loss of information and also achieve the capacity region for MIMO broadcast channels (BCs) [15], which, however, has a relatively high computational complexity. To this end, we develop a suboptimal yet efficient solution based on zero-forcing dirty paper coding (ZF-DPC) [16]. We compare the three schemes through extensive simulations, and the results show that ZF-DPC provides better performance than SDR and ZF in most cases.

2. Related Work

Initial research works in the field of IoT mainly focus on the building management [17, 18] and the related security as well as energy issues [19]. Building management includes many aspects, for example, traffic control, surveillance, energy management, and indoor environmental and air quality (IEAQ) control. The authors in [3] propose an energy-efficient large-scale and diffusive object monitoring mechanism to reduce communication costs and thus to reduce the energy consumption. In this paper, we propose using SWIPT for energy harvesting in IoT, which is convenient, costless, and environmentally friendly.

SWIPT has drawn an upsurge of interests. The authors in [20] study a multiuser multiple-input single-output (MISO) broadcast SWIPT system, where there exist two kinds of users, namely, ID users and EH users; the aim is to maximize the weighted sum power transferred to all EH receivers subject to given minimum SINR constraints at different ID users. The authors in [21] consider a point-to-point wireless link over the narrow band flat-fading channel subject to time-varying cochannel interference and then propose two different schemes, namely, time switching (TS) and power splitting (PS), for distributed information and energy receivers to evaluate the performance of the system. The recent research mainly focuses on the users who can receive information and harvest energy simultaneously. The authors in [22] consider SWIPT in the multiuser single-input single-output (SISO) interference channels and maximize the minimum SINR of users under the maximum transmit power and the minimum EH constraints; then the authors propose two algorithms: a centralized algorithm to optimally solve the nonconvex problem and a distributed algorithm designed to update its transmit power and PS ratio through an iterative process. The authors in [10] utilize the ZF and MRT schemes to minimize the total transmit power at the BS under the given SINR and the minimum EH constraints. The authors in [11–13] apply the SDR technique [14] to solve the nonconvex problem.

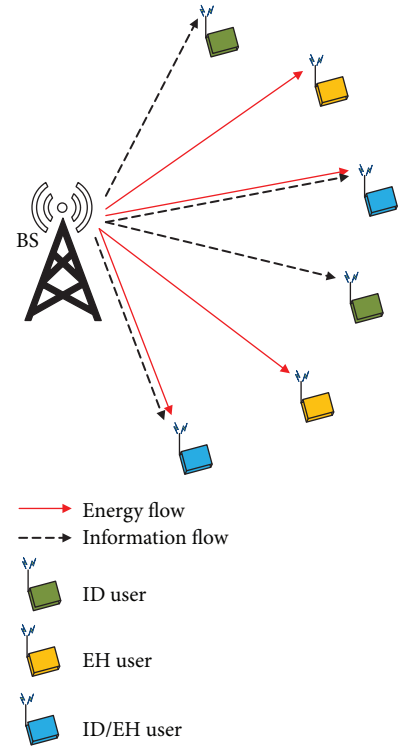


Figure 1: A general model for the coexistence of heterogeneous users.

Different from most of the existing studies that only consider ID/EH users, the novelty of our work is to investigate the coexistence of ID users, ID/EH users, and EH users under the multiuser MISO broadcast channels, as shown in Figure 1. Our objective is to maximize the minimum SINR of the ID and ID/EH users by jointly designing the power allocations at the transmitter and the PS ratios. Specifically, we discuss how to adapt the traditional schemes in the literature such as SDR, ZF, and MRT to the new scenario and examine their effectiveness. We further propose a new nonlinear ZF-DPC scheme for the new scenario and compare it with the traditional ones.

3. System Model and Problem Formulation

We consider a multiuser MISO broadcast system consisting of one BS and K users. The system model is depicted in Figure 2, where the BS is equipped with $N_t > 1$ antennas, and each user is equipped with one antenna. We summarized the notations in Notations. The first M users can only receive information (namely, ID users), while the next N users can receive information and harvest energy through a power splitter simultaneously (namely, ID/EH users), and the other $(K-M-N)$ users can only harvest energy (namely, EH users). We assume linear transmit precoding at the BS, where each

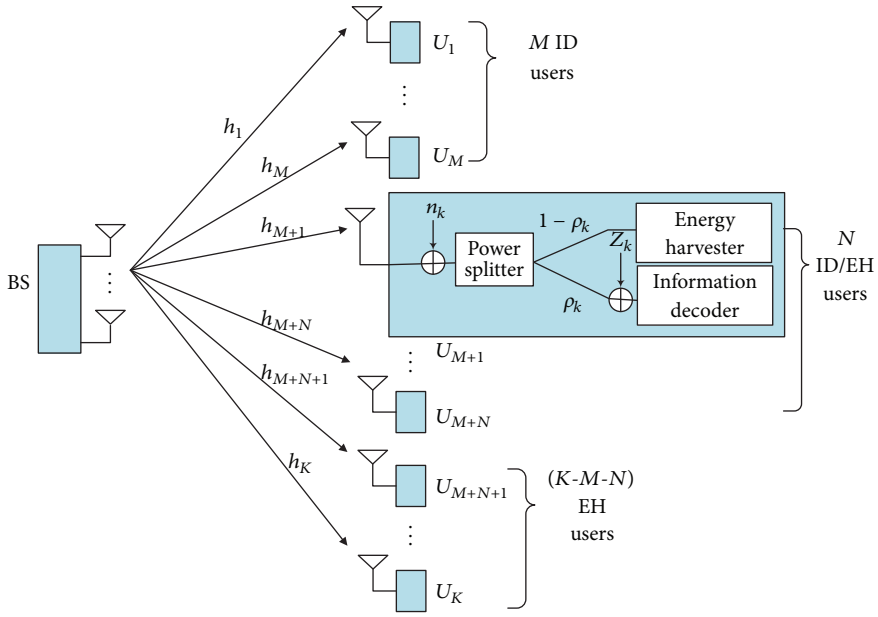


Figure 2: A multiuser MISO broadcast system.

user is assigned one dedicated information beam [14]. The complex baseband transmitted signal at the BS is

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{v}_k s_k, \tag{1}$$

where s_k denotes the transmitted data symbol for user k and $\mathbf{v}_k$ is the corresponding transmit beamforming vector. It is assumed that $s_k, k = 1, \ldots, K$, are independent and identically distributed (i.i.d.) cyclic symmetric complex Gaussian (CSCG) random variables with zero mean and unit variance, denoted by $s_k \sim \mathcal{CN}(0, 1)$.

We assume quasi-static flat-fading channel for all users and for convenience denote $\mathbf{h}_k$ as the conjugated complex channel vector from BS to user k. Then the received signal at user k is given by

$$y_k = \mathbf{h}_k^H \sum_{j=1}^{K} \mathbf{v}_j s_j + n_k, \quad k = 1, \ldots, K, \tag{2}$$

where $n_k \sim \mathcal{CN}(0, \sigma_k^2)$ denotes the antenna noise at user k. Therefore, the SINR of the first M ID users is

$$\mathrm{SINR}_k = \frac{\left|\mathbf{h}_k^H \mathbf{v}_k\right|^2}{\sum_{j \neq k} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \sigma_k^2}, \quad k = 1, \ldots, M. \tag{3}$$

As for the next N ID/EH users, we assume that each applies the PS scheme [8] to coordinate the process of information decoding and energy harvesting from the received signal, which allocates a ρ_k $(0 \leq \rho_k \leq 1)$ portion of the signal power to ID and the remaining $(1 - \rho_k)$ portion to EH. As a result, the signal split to user k's ID part is

$$y_k^{\mathrm{ID}} = \sqrt{\rho_k} \left(\mathbf{h}_k^H \sum_{j=1}^{K} \mathbf{v}_j s_j + n_k \right) + z_k, \quad k = M+1, \ldots, M+N, \tag{4}$$

where $z_k \sim \mathcal{CN}(0, \sigma_z^2)$ is the additional noise introduced by user k's ID part.

Accordingly, the SINR at user k's ID part is given by

$$\mathrm{SINR}_k = \frac{\rho_k \left|\mathbf{h}_k^H \mathbf{v}_k\right|^2}{\rho_k \sum_{j \neq k} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \rho_k \sigma_k^2 + \sigma_z^2}, \quad k = M+1, \ldots, M+N. \tag{5}$$

On the other hand, the signal split to user k's EH part is

$$y_k^{\mathrm{EH}} = \sqrt{1 - \rho_k} \left(\mathbf{h}_k^H \sum_{j=1}^{K} \mathbf{v}_j s_j + n_k \right), \quad k = M+1, \ldots, M+N. \tag{6}$$

Then, the harvested power of user k's EH part is given by

$$E_k = \zeta_k (1 - \rho_k) \left(\sum_{j=1}^{K} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \sigma_k^2 \right), \quad k = M+1, \ldots, M+N, \tag{7}$$

where $\zeta_k \in (0, 1]$ denotes the energy conversion efficiency of user k.

For the remaining $(K - M - N)$ EH users, we have

$$E_k = \zeta_k \left(\sum_{j=1}^{K} \left| \mathbf{h}_k^H \mathbf{v}_j \right|^2 + \sigma_k^2 \right), \quad k = M + N + 1, \ldots, K. \quad (8)$$

Assume that each EH user k $(k = M + 1, \ldots, K)$ has a minimum harvest energy demand e_k and that the BS has a maximum transmit power constraint $P^{\max}$. Our aim is to maximize the minimum SINR by jointly designing the transmit beamforming vectors $\{\mathbf{v}_k\}, k = 1, \ldots, K$, and the PS ratios $\{\rho_k\}, k = M + 1, \ldots, M + N$, under these constraints, which is formulated as follows:

$$\begin{aligned} \max_{\mathbf{v},\boldsymbol{\rho}} \quad & \min_{k=1,\ldots,M+N} \mathrm{SINR}_k \\ \text{s.t.} \quad & \sum_{k=1}^{K} \left\| \mathbf{v}_k \right\|^2 \le P^{\max}, \\ & E_k \ge e_k, \quad k = M + 1, \ldots, K, \\ & 0 \le \rho_k \le 1, \quad k = M + 1, \ldots, M + N, \end{aligned} \quad (9)$$

where

$$\begin{aligned} \mathrm{SINR}_k &= \frac{\left| \mathbf{h}_k^H \mathbf{v}_k \right|^2}{\sum_{j \ne k} \left| \mathbf{h}_k^H \mathbf{v}_j \right|^2 + \sigma_k^2}, \quad k = 1, \ldots, M, \\ \mathrm{SINR}_k &= \frac{\rho_k \left| \mathbf{h}_k^H \mathbf{v}_k \right|^2}{\rho_k \sum_{j \ne k} \left| \mathbf{h}_k^H \mathbf{v}_j \right|^2 + \rho_k \sigma_k^2 + \sigma_z^2}, \\ & \qquad k = M + 1, \ldots, M + N, \\ E_k &= \zeta_k \left(1 - \rho_k\right) \left(\sum_{j=1}^{K} \left| \mathbf{h}_k^H \mathbf{v}_j \right|^2 + \sigma_k^2 \right), \\ & \qquad k = M + 1, \ldots, M + N, \\ E_k &= \zeta_k \left(\sum_{j=1}^{K} \left| \mathbf{h}_k^H \mathbf{v}_j \right|^2 + \sigma_k^2 \right), \\ & \qquad k = M + N + 1, \ldots, K. \end{aligned} \quad (10)$$

By introducing an auxiliary variable $\tau = \min_{k=1,\ldots,M+N} \mathrm{SINR}_k$, problem (9) can be reformulated as

$$\begin{aligned} \max_{\mathbf{v},\boldsymbol{\rho},\tau} \quad & \tau \\ \text{s.t.} \quad & \mathrm{SINR}_k \ge \tau, \quad k = 1, \ldots, M + N, \\ & \sum_{k=1}^{K} \left\| \mathbf{v}_k \right\|^2 \le P^{\max}, \\ & E_k \ge e_k, \quad k = M + 1, \ldots, K, \\ & 0 \le \rho_k \le 1, \quad k = M + 1, \ldots, M + N. \end{aligned} \quad (11)$$

It can be shown that problem (11) becomes a feasibility problem for any given τ (please refer to Appendix A for details). Therefore, by applying the bisection search method on τ over a specific interval and solving the feasibility problem at each step with the associated τ, the optimal solution to problem (11) is also that to problem (9).

4. Three Traditional Schemes: SDR, ZF, and MRT

In this section, we introduce three traditional schemes, namely, SDR, ZF, and MRT, to solve the nonconvex problem (11) in a multiuser MISO broadcast system.

4.1. SDR Scheme. Define $\mathbf{X}_k = \mathbf{v}_k \mathbf{v}_k^H, \forall k$; then $\mathrm{rank}(\mathbf{X}_k) = 1$, $\forall k$. By ignoring the rank-one constraint for all $\mathbf{X}_k$, the SDR of problem (11) is given by

$$\begin{aligned} \max_{\mathbf{X},\boldsymbol{\rho},\tau} \quad & \tau \\ \text{s.t.} \quad & \sum_{k=1}^{K} \mathrm{trace}\left(\mathbf{X}_k\right) \le P^{\max}, \\ & \frac{\mathbf{h}_k^H \mathbf{X}_k \mathbf{h}_k}{\sum_{j \ne k} \mathbf{h}_k^H \mathbf{X}_j \mathbf{h}_k + \sigma_k^2} \ge \tau, \quad k = 1, \ldots, M, \\ & \frac{\rho_k \mathbf{h}_k^H \mathbf{X}_k \mathbf{h}_k}{\rho_k \sum_{j \ne k} \mathbf{h}_k^H \mathbf{X}_j \mathbf{h}_k + \rho_k \sigma_k^2 + \sigma_z^2} \ge \tau, \\ & \qquad k = M + 1, \ldots, M + N, \\ & \zeta_k \left(1 - \rho_k\right) \left(\sum_{j=1}^{K} \mathbf{h}_k^H \mathbf{X}_j \mathbf{h}_k + \sigma_k^2 \right) \ge e_k, \\ & \qquad k = M + 1, \ldots, M + N, \\ & 0 \le \rho_k \le 1, \quad k = M + 1, \ldots, M + N, \\ & \zeta_k \left(\sum_{j=1}^{K} \mathbf{h}_k^H \mathbf{X}_j \mathbf{h}_k + \sigma_k^2 \right) \ge e_k, \\ & \qquad k = M + N + 1, \ldots, K, \\ & \mathbf{X}_k \succeq 0, \quad \forall k. \end{aligned} \quad (12)$$

Problem (12) is still nonconvex, since both the SINR and harvested power constraints involve coupled $\mathbf{X}_k$ and ρ_k. This problem can be reformulated as the following problem:

$$\begin{aligned} \max_{\mathbf{X},\boldsymbol{\rho},\tau} \quad & \tau \\ \text{s.t.} \quad & \sum_{k=1}^{K} \mathrm{trace}\left(\mathbf{X}_k\right) \le P^{\max}, \\ & \frac{1}{\tau} \mathbf{h}_k^H \mathbf{X}_k \mathbf{h}_k - \sum_{j \ne k} \mathbf{h}_k^H \mathbf{X}_j \mathbf{h}_k \ge \sigma_k^2, \quad k = 1, \ldots, M, \end{aligned}$$

$$\begin{aligned}
&\frac{1}{\tau}\mathbf{h}_k^H\mathbf{X}_k\mathbf{h}_k - \sum_{j\neq k}\mathbf{h}_k^H\mathbf{X}_j\mathbf{h}_k \geq \sigma_k^2 + \frac{\sigma_z^2}{\rho_k}, \\
&\qquad\qquad k = M+1,\ldots,M+N, \\
&\sum_{j=1}^{K}\mathbf{h}_k^H\mathbf{X}_j\mathbf{h}_k \geq \frac{e_k}{\zeta_k\left(1-\rho_k\right)} - \sigma_k^2, \\
&\qquad\qquad k = M+1,\ldots,M+N, \\
&0 \leq \rho_k \leq 1, \quad k = M+1,\ldots,M+N, \\
&\sum_{j=1}^{K}\mathbf{h}_k^H\mathbf{X}_j\mathbf{h}_k \geq \frac{e_k}{\zeta_k} - \sigma_k^2, \\
&\qquad\qquad k = M+1,\ldots,M+N, \\
&\mathbf{X}_k \geq 0, \quad \forall k.
\end{aligned} \tag{13}$$

Note that problem (13) is convex; therefore, the optimal solution can be obtained by applying the bisection search method on τ. Let $\{\mathbf{X}_k^*\}$ and $\{\rho_k^*\}$ denote the optimal solution to problem (13). If $\{\mathbf{X}_k^*\}$ satisfies $\text{rank}(\mathbf{X}_k^*) = 1$, $\forall k$, then the optimal beamforming solution to problem (11) can be obtained by doing eigenvalue decomposition (EVD) of $\mathbf{X}_k^* = \overline{\mathbf{v}}_k\overline{\mathbf{v}}_k^H$, $\forall k$; otherwise we need to do Gaussian randomization [23] to choose the best solution.

In Appendix B, we have proven that $\mathbf{X}_k^*$ satisfies $\text{rank}(\mathbf{X}_k^*) = 1$ when there are no EH users, where $K = M+N$, and from the numerical analysis in Section 6, we could verify that no matter whether there exist EH users, $\mathbf{X}_k^*$ for the ID users and the ID/EH users always satisfy the rank-one conditions, but it is not suitable for the EH users, and thus we need to do Gaussian randomization to choose the best beamforming vectors $\{\mathbf{v}_k\}$ for them. However, we just need to know the achievable minimum SINR which is obtained directly from $\mathbf{X}_k^*$ rather than $\{\mathbf{v}_k\}$ in this paper, and thus we can ignore the process of Gaussian randomization.

The optimal solution to problem (11) can be obtained via solving problem (13) by the interior-point algorithm [24] using standard solvers, for example, CVX [25], and the complexity of the interior-point algorithm is $\mathcal{O}((K + N_t^2)^{3.5}\log(1/\varepsilon))$ [26], where ε is the desired numerical accuracy.

4.2. ZF Scheme. In ZF, the condition $N_t \geq K$ must be satisfied [10], and $\mathbf{h}_k$ are not linearly dependent. The ZF beamforming scheme can then be used to eliminate the multiuser interference by restricting $\mathbf{v}_k$ to satisfy $\mathbf{h}_i^H\mathbf{v}_k = 0$, $\forall i \neq k$. More specifically, the ZF weight $\mathbf{v}_k$ is provided by the solution to the following problem:

$$\begin{aligned}
\min_{\mathbf{v}_k} \quad & \left|\mathbf{h}_k^H\mathbf{v}_k\right|^2 \\
\text{s.t.} \quad & \mathbf{H}_k^H\mathbf{v}_k = \mathbf{0}_{(K-1)\times 1},
\end{aligned} \tag{14}$$

where $\mathbf{H}_k \triangleq [\mathbf{h}_1,\ldots,\mathbf{h}_{k-1},\mathbf{h}_{k+1},\ldots,\mathbf{h}_K] \in \mathbb{C}^{N_t\times(K-1)}$. Assume that $\mathbf{v}_k = \sqrt{p_k}\mathbf{w}_k$, where $\|\mathbf{w}_k\|_2^2 = 1$ and the solution is given by [27]

$$\mathbf{w}_k^{(\text{ZF})} = \frac{\left(\mathbf{I}_{N_t} - \mathbf{F}\right)\mathbf{h}_k}{\left\|\left(\mathbf{I}_{N_t} - \mathbf{F}\right)\mathbf{h}_k\right\|}, \tag{15}$$

where $\mathbf{F} = \mathbf{H}_k^\dagger\mathbf{H}_k^H$ and $\mathbf{H}_k^\dagger = \mathbf{H}_k(\mathbf{H}_k^H\mathbf{H}_k)^{-1}$ is the Moore-Penrose inverse of $\mathbf{H}_k$.

Let $G_{i,j} = |\mathbf{h}_i^H\mathbf{w}_j^{(\text{ZF})}|^2$ denote the link gain, and we have $G_{i,j} = 0$, $i \neq j$. Hence, problem (11) is equivalent to the following problem:

$$\begin{aligned}
\max_{p_k,\rho,\tau} \quad & \tau \\
\text{s.t.} \quad & \sum_{k=1}^{K} p_k \leq P^{\max}, \\
& G_{k,k}p_k \geq \tau\sigma_k^2, \quad k = 1,\ldots,M, \\
& G_{k,k}p_k \geq \tau\sigma_k^2 + \frac{\tau\sigma_z^2}{\rho_k}, \\
& \qquad k = M+1,\ldots,M+N, \\
& \zeta_k\left(1-\rho_k\right)\left(G_{k,k}p_k + \sigma_k^2\right) \geq e_k, \\
& \qquad k = M+1,\ldots,M+N, \\
& 0 \leq \rho_k \leq 1, \quad k = M+1,\ldots,M+N, \\
& \zeta_k\left(G_{k,k}p_k + \sigma_k^2\right) \geq e_k, \\
& \qquad k = M+N+1,\ldots,K, \\
& p_k \geq 0, \quad \forall k.
\end{aligned} \tag{16}$$

Let $x_k = G_{k,k}p_k + \sigma_k^2$, $\alpha = (1+\tau)\sigma_k^2$, $\beta = \tau\sigma_z^2$, and $\gamma = e_k/\zeta_k$, and the closed-form solution is given by

$$\begin{aligned}
&p_k^* = \frac{\tau\sigma_k^2}{G_{k,k}}, \quad k = 1,\ldots,M, \\
&p_k^* = \frac{1}{G_{k,k}}\left(x^* - \sigma_k^2\right), \\
&\qquad \rho_k^* = 1 - \frac{\gamma}{x^*}, \ k = M+1,\ldots,M+N, \\
&p_k^* = \frac{\gamma - \sigma_k^2}{G_{k,k}}, \quad k = M+N+1,\ldots,K,
\end{aligned} \tag{17}$$

where $x^* = (1/2)(\alpha+\beta+\gamma+\sqrt{(\alpha+\beta+\gamma)^2 - 4\alpha\gamma})$. Please refer to Appendix C for details.

Clearly, the complexity of the ZF scheme is dominated by the K times of SVD operations. Since the complexity of each SVD operation is $\mathcal{O}(KN_t^2 + K^2N_t + K^3)$ [28], the overall complexity of the ZF scheme is $\mathcal{O}(K^2N_t^2 + K^3N_t + K^4)$.

4.3. MRT Scheme. The MRT beamforming maximizes the SNR at each receiver ($|\mathbf{h}_k^H \mathbf{v}_k|^2/\sigma_k^2, \forall k$), and it only requires the knowledge of the direct links $\mathbf{h}_k$; thus, it is of relatively low complexity to obtain the beamforming vectors. Assuming that $\mathbf{v}_k = \sqrt{p_k}\mathbf{w}_k$, where $\|\mathbf{w}_k\|_2^2 = 1$ and p_k is the power allocated to user k, the MRT beamforming [29] can be expressed as

$$\mathbf{w}_k^{\mathrm{MRT}} = \frac{\mathbf{h}_k}{\|\mathbf{h}_k\|}. \tag{18}$$

It is worth noting that MRT does not take into account the signals transmitted to other users and therefore it results in a strong cross-interference. Although this cross-interference is a bottleneck for conventional MISO systems, it could be beneficial for scenarios with EH constraints. Then, let $G_{i,j} = |\mathbf{h}_i^H \mathbf{w}_j^{(\mathrm{MRT})}|^2$ denote the link gain; problem (11) can be cast into a second-order cone programming (SOCP) [30] formulation as follows:

$$\begin{aligned}
\max_{p_k,\rho,\tau} \quad & \tau \\
\text{s.t.} \quad & \sum_{k=1}^{K} p_k \le P^{\max}, \\
& \frac{1}{\tau} G_{k,k} p_k - \sum_{j\neq k} G_{k,j} p_j \ge \sigma_k^2, \quad k = 1, \ldots, M, \\
& \frac{1}{\tau} G_{k,k} p_k - \sum_{j\neq k} G_{k,j} p_j \ge \sigma_k^2 + \frac{\sigma_z^2}{\rho_k}, \\
& \qquad\qquad k = M+1, \ldots, M+N, \\
& \zeta_k (1-\rho_k) \left(\sum_{j=1}^{K} G_{k,j} p_j + \sigma_k^2 \right) \ge e_k, \\
& \qquad\qquad k = M+1, \ldots, M+N, \\
& 0 \le \rho_k \le 1, \quad k = M+1, \ldots, M+N, \\
& \zeta_k \left(\sum_{j=1}^{K} G_{k,j} p_j + \sigma_k^2 \right) \ge e_k, \\
& \qquad\qquad k = M+N+1, \ldots, K, \\
& p_k \ge 0, \quad \forall k.
\end{aligned} \tag{19}$$

The optimal solution to problem (19) can be obtained by using standard solvers, for example, CVX. The complexity of the MRT scheme is dominated by the K times of computing the beamforming vector $\mathbf{w}_k^{\mathrm{MRT}}$ with $\mathcal{O}(1)$ complexity each time. The SOCP algorithm for solving problem (19) is $\mathcal{O}((K+N)^{3.5}\log(1/\varepsilon))$ [26], where ε is the preset search precision. Thus, the complexity of MRT scheme is $\mathcal{O}((K+N)^{3.5}\log(1/\varepsilon))$.

5. Nonlinear Precoding Scheme: ZF-DPC

In this section, we present the ZF-DPC scheme that applies ZF with QR decomposition to eliminate the causal interference and then uses DPC to eliminate the noncausal interference.

Denote the MISO channels by $\mathbf{H} \triangleq [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_K]^H \in \mathbb{C}^{K\times N_t}$. Then the QR decomposition [31] obtained by applying Gram-Schmidt orthogonalization to the rows of $\mathbf{H}^H$ is $\mathbf{H}^H = \mathbf{R}\mathbf{Q}^H$; then $\mathbf{H}$ can be expressed as

$$\mathbf{H} = \mathbf{Q}\mathbf{R}^H. \tag{20}$$

Let $m = \mathrm{rank}(\mathbf{H}) \le \min\{K, N_t\}$. Then $\mathbf{Q}$ is a $K \times m$ lower triangular matrix (i.e., it has zeros above its main diagonal), and denote $q_{i,j}$ as the (i,j)th element of $\mathbf{Q}$. $\mathbf{R}$ is an $N_t \times m$ subunitary matrix with m orthonormal columns. By letting the beamforming matrix $\mathbf{B}_{\mathrm{ZF}} = \mathbf{R}$, the transmit signal of ZF-DPC is given by

$$\mathbf{x}_{\mathrm{ZF}} = \mathbf{B}_{\mathrm{ZF}}\mathbf{u}_{\mathrm{ZF}}, \tag{21}$$

where $\mathbf{u}_{\mathrm{ZF}} = [u_1, u_2, \ldots, u_m]^T$ and it satisfies the power constraints $\|\mathbf{u}_{\mathrm{ZF}}\|^2 \le P^{\max}$. Then the received signal of user k is given by the set of interference channels:

$$y_k = q_{k,k}u_k + \sum_{j<k} q_{k,j}u_j + n_k, \quad k = 1, \ldots, m, \tag{22}$$

while no information is sent to users $m+1, \ldots, K$. In this case, there is no point in maximizing the minimum SINR, and we assume that $\mathbf{h_k}$ are not linearly dependent and the condition $N_t \ge K$ must be satisfied, and we have $m = K$.

Based on (20)–(22), the received signals $\mathbf{y} = [y_1, y_2, \ldots, y_K]^T$ can be written as

$$\mathbf{y} = \mathbf{H}\mathbf{B}_{\mathrm{ZF}}\mathbf{u}_{\mathrm{ZF}} + \mathbf{n} = \mathbf{Q}\mathbf{u}_{\mathrm{ZF}} + \mathbf{n}, \tag{23}$$

where $\mathbf{n} = [n_1, n_2, \ldots, n_K]^T$. According to the QR decomposition of the channel matrix, different ordering among users will lead to different channel gain. For the sake of convenience, we assume that the coding order is $[1, 2, \ldots, K]$. Note that the interuser interference from the off-diagonal entries of $\mathbf{Q}$ can be cancelled by successive DPC by the precoding matrix $\mathbf{W} = \mathbf{Q}^{-1}\mathrm{diag}(q_{1,1}, q_{2,2}, \ldots, q_{K,K})$ and $\mathbf{u}_{\mathrm{ZF}} = \mathbf{W}\bar{\mathbf{s}}_{\mathrm{ZF}}$, where $\bar{\mathbf{s}}_{\mathrm{ZF}}$ is the transmitted signal. Then, the received signals can be written as $\mathbf{y} = \mathrm{diag}(q_{1,1}, q_{2,2}, \ldots, q_{K,K})\bar{\mathbf{s}}_{\mathrm{ZF}} + \mathbf{n}$. Thus the BC is transformed into K independent channels with $q_{k,k}^2$, $k = 1, \ldots, K$, being the channel gain. Denote p_k as the power allocated to user k, and the SINR of user k is given by

$$\begin{aligned}
\mathrm{SINR}_k &= \frac{q_{k,k}^2 p_k}{\sigma_k^2}, \quad k = 1, \ldots, M, \\
\mathrm{SINR}_k &= \frac{\rho_k q_{k,k}^2 p_k}{\rho_k \sigma_k^2 + \sigma_z^2}, \quad k = M+1, \ldots, M+N.
\end{aligned} \tag{24}$$

The EH of user k is

$$E_k = \zeta_k (1-\rho_k)\left(q_{k,k}^2 p_k + \sigma_k^2\right), \quad k = M+1,\ldots,M+N, \quad (25)$$
$$E_k = \zeta_k \left(q_{k,k}^2 p_k + \sigma_k^2\right), \quad k = M+N+1,\ldots,K.$$

Now problem (11) can be reformulated as follows:

$$\begin{aligned}
\max_{p_k,\rho,\tau} \quad & \tau \\
\text{s.t.} \quad & \sum_{k=1}^{K} p_k \le P^{\max}, \\
& q_{k,k}^2 p_k \ge \tau\sigma_k^2, \quad k = 1,\ldots,M, \\
& q_{k,k}^2 p_k \ge \tau\sigma_k^2 + \frac{\tau\sigma_z^2}{\rho_k}, \quad k = M+1,\ldots,M+N, \\
& \zeta_k (1-\rho_k)\left(q_{k,k}^2 p_k + \sigma_k^2\right) \ge e_k, \\
& \qquad k = M+1,\ldots,M+N, \\
& 0 \le \rho_k \le 1, \quad k = M+1,\ldots,M+N, \\
& \zeta_k \left(q_{k,k}^2 p_k + \sigma_k^2\right) \ge e_k, \\
& \qquad k = M+1,\ldots,M+N, \\
& p_k \ge 0, \quad \forall k.
\end{aligned} \quad (26)$$

Similar to the ZF scheme, let $x_k = q_{k,k}^2 p_k + \sigma_k^2$, $\alpha = (1+\tau)\sigma_k^2$, $\beta = \tau\sigma_z^2$, and $\gamma = e_k/\zeta_k$, and then the solution is

$$\begin{aligned}
p_k^* &= \frac{\tau\sigma_k^2}{q_{k,k}^2}, \quad k = 1,\ldots,M, \\
p_k^* &= \frac{1}{q_{k,k}^2}\left(x^* - \sigma_k^2\right), \\
\rho_k^* &= 1 - \frac{\gamma}{x^*}, \quad k = M+1,\ldots,M+N, \\
p_k^* &= \frac{\gamma - \sigma_k^2}{q_{k,k}^2}, \quad k = M+N+1,\ldots,K,
\end{aligned} \quad (27)$$

where $x^* = (1/2)(\alpha+\beta+\gamma+\sqrt{(\alpha+\beta+\gamma)^2 - 4\alpha\gamma})$.

The complexity of the ZF-DPC scheme is divided into two parts. The first part is to perform QR decomposition, the complexity of which is $\mathcal{O}(K^2 N_t)$; the other part is to perform pseudoinverse and matrix multiplication (multiply a diagonal matrix) for both $K \times K$ matrices, and the complexity is $\mathcal{O}(K^3)$ [32] and $\mathcal{O}(K)$, respectively. Thus the overall complexity of the ZF-DPC scheme is $\mathcal{O}(K^2 N_t + K^3)$.

6. Numerical Analysis

In this section, we numerically evaluate the performance of the proposed schemes. For simplicity, we assume that $\zeta_k = \zeta$, $e_k = e$, $k = M+1,\ldots,K$, and $\sigma_c^2 = \sigma_z^2 = 10^{-4}$, $\forall k$; the BS is equipped with $N_t = 8$ antennas, and the transmit power is no more than $P^{\max}$.

We have the following initial observations. First, we can see from Figure 3 that the system's performance in terms of the achievable minimum SINR under different ζ is almost the same when ζ exceeds a certain value, say 0.6, in our simulation. We use $\zeta = 1$ in the next simulations without loss of generality.

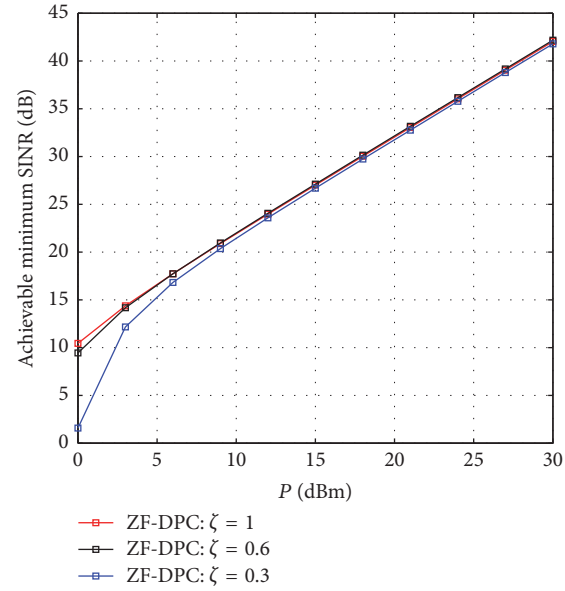


Figure 3: Achievable minimum SINR versus transmit power under different ζ.

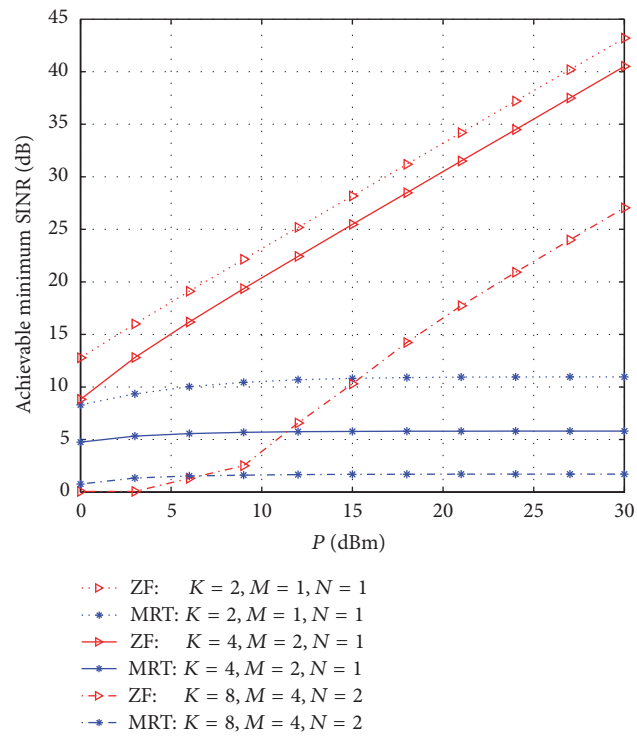


Figure 4: Comparison of ZF and MRT schemes.

Second, we can see from Figure 4 that the performance of the MRT scheme remains almost unchanged when the transmit power increases, due to the strong cross-interference

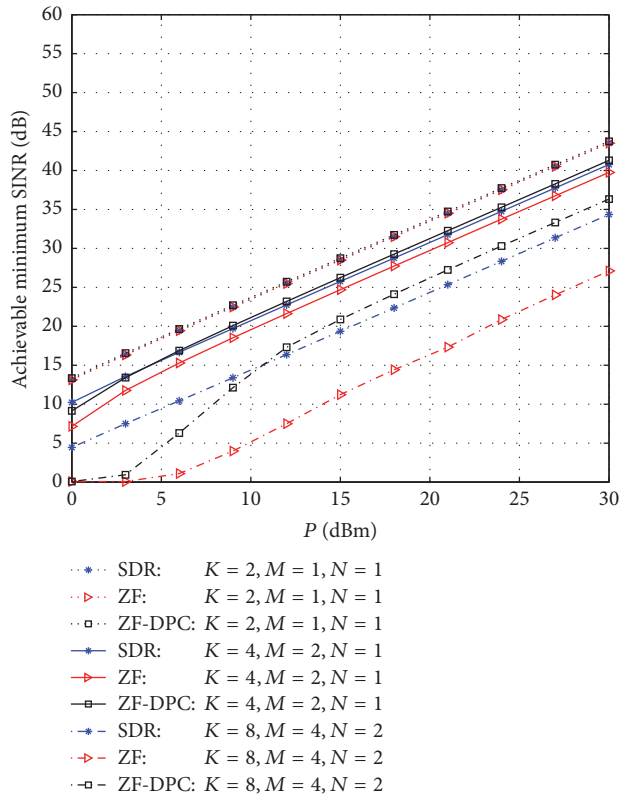


Figure 5: Achievable minimum SINR versus transmit power with different numbers of users.

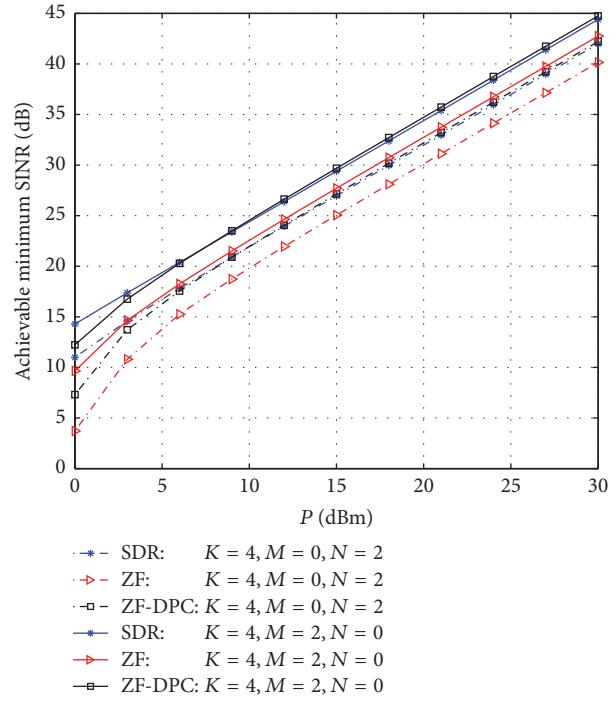


Figure 6: Achievable minimum SINR versus transmit power with fixed number of EH users.

from other users, and the ZF scheme preforms better than the MRT scheme except when $K = 8$, $M = 4$, and $N = 2$ and the transmit power is relatively small, which also indicates that interference is beneficial to EH. A simple method to ensure that the direct links are much stronger than the interference links is to strengthen the direct channel links by a multiplicative constant δ [33]; that is, $\delta \mathbf{h}_{i,i} \to \mathbf{h}_{i,i}$, which is, however, infeasible in our scenario with only one BS. We do not consider the MRT scheme in the remaining part, since it is inferior to ZF in most cases.

6.1. SINR versus Transmit Power. In this part, we investigate the achievable minimum SINR of the proposed schemes under different transmit power of the BS when $e = 0$ dBm.

From Figures 5 and 6, we can see that as the transmit power $P^{\max}$ increases from 0 dBm to 30 dBm, the system's performance of all three schemes increases substantially, and the performance of ZF is always the worst. When the number of users is relatively small, the performances of the three schemes are almost the same, especially when $K = 2$, $M = 1$, and $N = 1$ due to less interference from other users. Yet, as shown in Figure 5, with the increasing number of users (e.g., $K = 4$ and $K = 8$), the performances of different schemes differ significantly.

Specifically, when the transmit power is relatively small, SDR has better performance than ZF-DPC, while with higher transmit power, the performance of ZF can approach SDR, and ZF-DPC has a gain of about 2 dB and about 10 dB over SDR and ZF, respectively, when $K = 8$, $M = 4$, and $N = 2$. We can also see from Figure 6 that when the number of total users K and the number of EH users are fixed, with the increasing number of ID users M (the number of ID/EH users decreasing correspondingly), the performances of the three schemes all improve about 3 dB. The reason is that, with fewer energy harvesters, the BS can allocate more power to the information receivers.

Figure 7 shows that our proposed ZF-DPC technique provides higher achievable minimum SINR than SDR and ZF even for the traditional scenario with ID/EH users only.

6.2. SINR versus EH. In this part, we investigate the relationship between the achievable minimum SINR and the minimum energy harvesting of the users, with fixed transmit power of the BS ($P^{\max} = 30$ dBm).

We can see from Figure 8 that the performances of the three schemes are similar when the number of users is small and the performances of different schemes differ significantly with the increasing number of the users.

From Figure 9, we can see that ZF-DPC performs worse than SDR only in the very high minimum harvest energy e regions. With the increase of the ID users M (the number of ID/EH users decreasing correspondingly), the performances of the three schemes all improve about 4 dB.

Figure 10 compares the proposed schemes when there are only ID/EH users, which indicates that our proposed ZF-DPC scheme is superior to the other schemes in most cases.

All the three figures again indicate the trade-off between energy harvesting and information transmission, which is a fundamental issue in the power splitting design. When the users need to harvest more energy, the achievable SINR

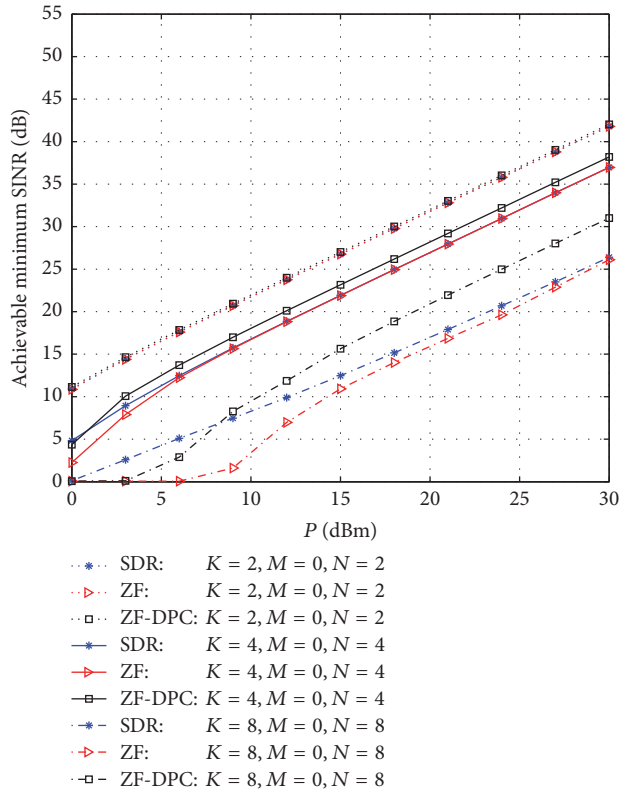


Figure 7: Achievable minimum SINR versus transmit power with ID/EH users only.

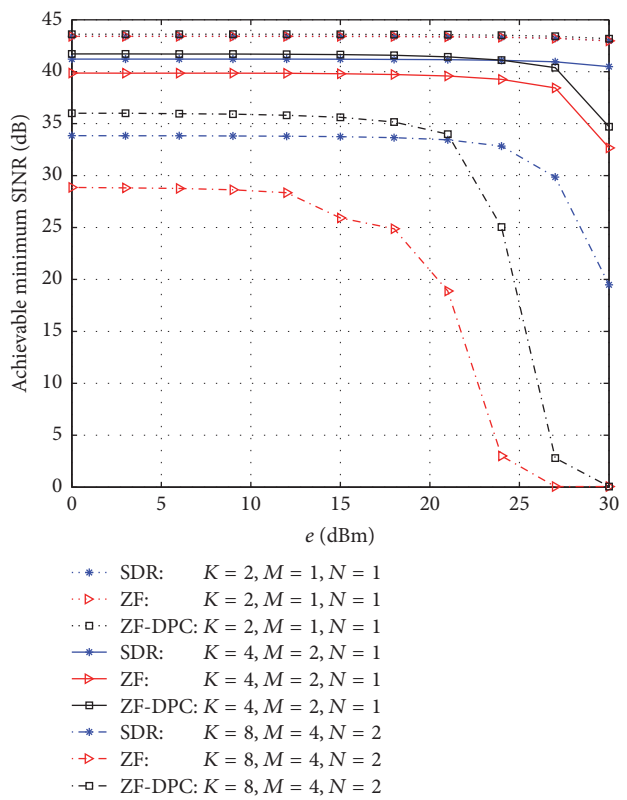


Figure 8: Achievable minimum SINR versus minimum EH constraint with different number of users.

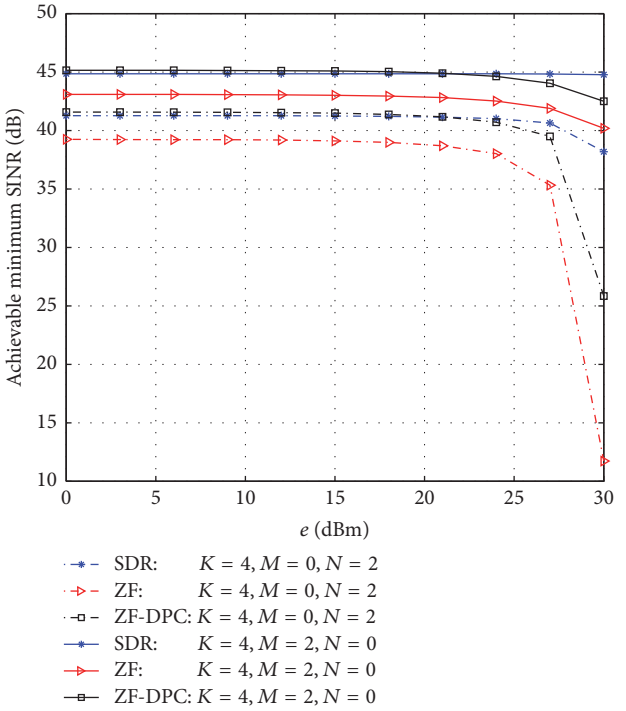


Figure 9: Achievable minimum SINR versus minimum EH constraint with fixed number of EH users.

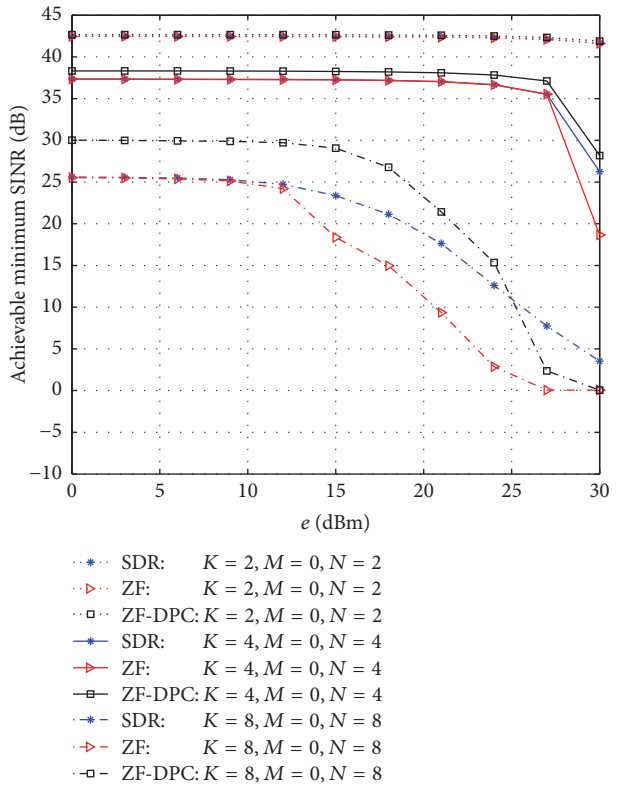


Figure 10: Achievable minimum SINR versus minimum EH constraint with ID/EH users only.

that corresponds to the capacity of information transmission becomes lower.

7. Conclusion

In this paper, we, for the first time, studied the energy harvesting in the IoT for MISO SWIPT systems with heterogeneous users. We developed four schemes, namely, SDR, ZF, MRT, and ZF-DPC, to solve the nonconvex problem. We then compared them through simulations and provided suggestions on how to select the proper scheme in different scenarios. In the future work, we plan to solve the problems when there are multiple transmitters and the perfect CSI is not available.

Appendix

A. The Proof of the Feasibility of Problem (11)

For any given τ, problem (11) can be reformulated as

$$\begin{aligned}
\min_{\mathbf{v},\boldsymbol{\rho}} \quad & \sum_{k=1}^{K} \left\|\mathbf{v}_k\right\|^2 \\
\text{s.t.} \quad & \frac{\left|\mathbf{h}_k^H \mathbf{v}_k\right|^2}{\sum_{j\neq k} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \sigma_k^2} \geq \tau, \quad k = 1, \ldots, M \\
& \frac{\rho_k \left|\mathbf{h}_k^H \mathbf{v}_k\right|^2}{\rho_k \sum_{j\neq k} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \rho_k \sigma_k^2 + \sigma_z^2} \geq \tau, \\
& \qquad\qquad k = M+1, \ldots, M+N, \\
& \zeta_k (1-\rho_k) \left(\sum_{j=1}^{K} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \sigma_k^2 \right) \geq e_k, \\
& \qquad\qquad k = M+1, \ldots, M+N, \\
& 0 \leq \rho_k \leq 1, \quad k = M+1, \ldots, M+N, \\
& \zeta_k \left(\sum_{j=1}^{K} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \sigma_k^2 \right) \geq e_k, \\
& \qquad\qquad k = M+N+1, \ldots, K.
\end{aligned} \tag{A.1}$$

Then examine whether the minimum power is less than the power constraint $P^{\max}$. Thus, we need to verify that problem (A.1) is feasible. And problem (A.1) is feasible if and only if the following problem is feasible:

$$\begin{aligned}
\text{find} \quad & \{\mathbf{v}_k, \rho_k\} \\
\text{s.t.} \quad & \frac{\left|\mathbf{h}_k^H \mathbf{v}_k\right|^2}{\sum_{j\neq k} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \sigma_k^2} \geq \tau, \quad k = 1, \ldots, M \\
& \frac{\rho_k \left|\mathbf{h}_k^H \mathbf{v}_k\right|^2}{\rho_k \sum_{j\neq k} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \rho_k \sigma_k^2 + \sigma_z^2} \geq \tau, \\
& \qquad\qquad k = M+1, \ldots, M+N, \\
& 0 \leq \rho_k \leq 1, \quad k = M+1, \ldots, M+N.
\end{aligned} \tag{A.2}$$

First, it can be easily verified that if problem (A.2) is not feasible, then problem (A.1) cannot be feasible, since problem (A.1) has additional constraints on harvested power. Second, suppose that problem (A.2) is feasible, and let $\{\mathbf{v}_k\}$ and $\{\rho_k\}$ be a feasible solution. It can be shown that there always exists the new solution $\{\alpha \mathbf{v}_k\}$, $\alpha > 1$, $\forall k$, with $\{\rho_k\}$, $k = M+1, \ldots, M+N$, being feasible to problem (A.1). Thus to prove the feasibility of problem (A.1), we need to prove the feasibility of problem (A.2), which is feasible if and only if the following problem is feasible:

$$\begin{aligned}
\text{find} \quad & \{\mathbf{v}_k\} \\
\text{s.t.} \quad & \frac{\left|\mathbf{h}_k^H \mathbf{v}_k\right|^2}{\sum_{j\neq k} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \sigma_k^2} \geq \tau, \quad k = 1, \ldots, M, \\
& \frac{\left|\mathbf{h}_k^H \mathbf{v}_k\right|^2}{\sum_{j\neq k} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \sigma_k^2 + \sigma_z^2} \geq \tau, \\
& \qquad\qquad k = M+1, \ldots, M+N.
\end{aligned} \tag{A.3}$$

To prove the feasibility of problem (A.2), first, suppose that problem (A.3) is feasible, and let $\{\mathbf{v}_k\}$ denote a feasible solution to problem (A.3). Then, given any $0 < \rho < 1$, consider the following solution to problem (A.2): $\overline{\mathbf{v}}_k = \mathbf{v}_k / \sqrt{\rho}$, $\forall k$, and $\overline{\rho}_k = \rho$, $k = M+1, \ldots, M+N$. We have

$$\begin{aligned}
\frac{\left|\mathbf{h}_k^H \overline{\mathbf{v}}_k\right|^2}{\sum_{j\neq k} \left|\mathbf{h}_k^H \overline{\mathbf{v}}_j\right|^2 + \sigma_k^2} &= \frac{\left|\mathbf{h}_k^H \mathbf{v}_k\right|^2}{\sum_{j\neq k} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \rho \sigma_k^2} \\
&> \frac{\left|\mathbf{h}_k^H \mathbf{v}_k\right|^2}{\sum_{j\neq k} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \sigma_k^2} \geq \tau, \quad k = 1, \ldots, M,
\end{aligned} \tag{A.4}$$

$$\begin{aligned}
&\frac{\overline{\rho}_k \left|\mathbf{h}_k^H \overline{\mathbf{v}}_k\right|^2}{\overline{\rho}_k \sum_{j\neq k} \left|\mathbf{h}_k^H \overline{\mathbf{v}}_j\right|^2 + \overline{\rho}_k \sigma_k^2 + \sigma_z^2} \\
&\quad = \frac{\left|\mathbf{h}_k^H \mathbf{v}_k\right|^2}{\sum_{j\neq k} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \rho \sigma_k^2 + \sigma_z^2} \\
&\quad > \frac{\left|\mathbf{h}_k^H \mathbf{v}_k\right|^2}{\sum_{j\neq k} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \sigma_k^2 + \sigma_z^2} \geq \tau, \\
&\qquad\qquad k = M+1, \ldots, M+N.
\end{aligned} \tag{A.5}$$

Combining (A.4) with (A.5), we can see that $\{\overline{\mathbf{v}}_k\}$ and $\{\overline{\rho}_k\}$ are a feasible solution to problem (A.2). Therefore, if problem (A.3) is feasible, problem (A.2) must be feasible too. Second, suppose that problem (A.2) is feasible, and let $\{\mathbf{v}_k\}$ and $\{\rho_k\}$ denote a feasible solution to problem (A.2). Since $\rho_k < 1$, $k = M+1,\ldots,M+N$, we have

$$\begin{aligned}\tau &\le \frac{\rho_k \left|\mathbf{h}_k^H \mathbf{v}_k\right|^2}{\rho_k \sum_{j\neq k} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \rho_k \sigma_k^2 + \sigma_z^2} \\ &= \frac{\left|\mathbf{h}_k^H \mathbf{v}_k\right|^2}{\sum_{j\neq k} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \sigma_k^2 + \sigma_z^2/\rho_k} \\ &< \frac{\left|\mathbf{h}_k^H \mathbf{v}_k\right|^2}{\sum_{j\neq k} \left|\mathbf{h}_k^H \mathbf{v}_j\right|^2 + \sigma_k^2 + \sigma_z^2}, \\ &\qquad k = M+1,\ldots,M+N.\end{aligned} \tag{A.6}$$

As a result, $\{\mathbf{v}_k\}$ is also a feasible solution to problem (A.3). As problem (A.3) is a well-known SINR feasible problem, we can conclude that problem (11) is feasible. Then, given sufficiently small τ, it always admits a feasible solution.

B. The Proof of $\mathbf{X}_k^*$ Satisfying $\mathrm{rank}(\mathbf{X}_k^*) = 1$

The problem of maximizing the minimum SINR is equal to minimizing the total transmission power at BS which is also NP-hard. But, contrary to the QoS approach, it always admits a feasible solution, apart from the trivial case of zero channel vectors. By applying the bisection search method over a specific interval $[L,U]$, we have $\tau = (L+U)/2$, and the problem is transformed into a QoS problem depicted in (B.1). If the minimum power is less than the power constraint $P^{\max}$, then update the lower bound $L = \tau$; otherwise update the upper bound $U = \tau$. Repeat the above steps until the algorithm stops when it satisfies $U - L \le \varepsilon_1$, where ε_1 is the desired accuracy of the search.

$$\begin{aligned}\min_{\mathbf{X},\boldsymbol{\rho}} \quad & \sum_{k=1}^{K} \mathrm{Tr}\left(\mathbf{X}_k\right) \\ \text{s.t.} \quad & \frac{\mathbf{h}_k^H \mathbf{X}_k \mathbf{h}_k}{\sum_{j\neq k} \mathbf{h}_k^H \mathbf{X}_j \mathbf{h}_k + \sigma_k^2} \ge \tau, \quad k = 1,\ldots,M, \\ & \frac{\rho_k \mathbf{h}_k^H \mathbf{X}_k \mathbf{h}_k}{\rho_k \sum_{j\neq k} \mathbf{h}_k^H \mathbf{X}_j \mathbf{h}_k + \rho_k \sigma_k^2 + \sigma_z^2} \ge \tau, \\ & \qquad k = M+1,\ldots,K, \\ & \zeta_k (1-\rho_k) \left(\sum_{j=1}^{K} \mathbf{h}_k^H \mathbf{X}_j \mathbf{h}_k + \sigma_k^2\right) \ge e_k, \\ & \qquad k = M+1,\ldots,K, \\ & 0 \le \rho_k \le 1, \quad k = M+1,\ldots,K, \\ & \mathbf{X}_k \succeq 0, \quad \forall k.\end{aligned} \tag{B.1}$$

After decoupling $\mathbf{X}_k$ and ρ_k, problem (B.1) can be reformulated as the following problem:

$$\begin{aligned}\min_{\mathbf{X},\boldsymbol{\rho}} \quad & \sum_{k=1}^{K} \mathrm{Tr}\left(\mathbf{X}_k\right) \\ \text{s.t.} \quad & \frac{1}{\tau} \mathbf{h}_k^H \mathbf{X}_k \mathbf{h}_k - \sum_{j\neq k} \mathbf{h}_k^H \mathbf{X}_j \mathbf{h}_k \ge \sigma_k^2, \\ & \qquad k = 1,\ldots,M, \\ & \frac{1}{\tau} \mathbf{h}_k^H \mathbf{X}_k \mathbf{h}_k - \sum_{j\neq k} \mathbf{h}_k^H \mathbf{X}_j \mathbf{h}_k \ge \sigma_k^2 + \frac{\sigma_z^2}{\rho_k}, \\ & \qquad k = M+1,\ldots,K, \\ & \sum_{j=1}^{K} \mathbf{h}_k^H \mathbf{X}_j \mathbf{h}_k \ge \frac{e_k}{\zeta_k (1-\rho_k)} - \sigma_k^2, \\ & \qquad k = M+1,\ldots,K, \\ & 0 \le \rho_k \le 1, \quad k = M+1,\ldots,K, \\ & \mathbf{X}_k \succeq 0, \quad \forall k.\end{aligned} \tag{B.2}$$

Let $\{\lambda_k\}$ and $\{\mu_k\}$ denote the dual variables associated with the SINR constraints and harvested power constraints of problem (B.2), respectively. Thus, we have $\lambda_k \ge 0$, $\forall k$, and $\mu_k \ge 0$, $k = M+1,\ldots,K$. Then the Lagrangian [10] of problem (B.2) is defined as

$$\begin{aligned}L\left(\{\mathbf{X}_k,\rho_k,\lambda_k,\mu_k\}\right) \triangleq & \sum_{k=1}^{K} \left(\mathrm{Tr}\left(\mathbf{X}_k\right) - \mathrm{Tr}\left(\mathbf{X}_k \mathbf{S}_k\right)\right) \\ & - \sum_{k=1}^{M} \lambda_k \left(\frac{1}{\tau} \mathbf{h}_k^H \mathbf{X}_k \mathbf{h}_k - \sum_{j\neq k} \mathbf{h}_k^H \mathbf{X}_j \mathbf{h}_k - \sigma_k^2\right) \\ & - \sum_{k=M+1}^{K} \lambda_k \left(\frac{1}{\tau} \mathbf{h}_k^H \mathbf{X}_k \mathbf{h}_k - \sum_{j\neq k} \mathbf{h}_k^H \mathbf{X}_j \mathbf{h}_k - \sigma_k^2 - \frac{\sigma_z^2}{\rho_k}\right) \\ & - \sum_{k=M+1}^{K} \mu_k \left(\sum_{j=1}^{K} \mathbf{h}_k^H \mathbf{X}_j \mathbf{h}_k - \frac{e_k}{\zeta_k (1-\rho_k)} + \sigma_k^2\right).\end{aligned} \tag{B.3}$$

Then the dual function of problem (B.2) is given by

$$\min_{0\le\rho_k\le 1}\left\{\sum_{k=1}^{K}\left(\mathrm{Tr}\left(\mathbf{A}_k\mathbf{X}_k\right)-\mathrm{Tr}\left(\mathbf{S}_k\mathbf{X}_k\right)\right)+\sum_{k=1}^{K}\lambda_k\sigma_k^2+\sum_{k=M+1}^{K}\left(\frac{\lambda_k\sigma_z^2}{\rho_k}+\frac{\mu_k e_k}{\zeta_k\left(1-\rho_k\right)}\right)-\sum_{k=M+1}^{K}\mu_k\sigma_k^2\right\}, \tag{B.4}$$

where

$$\begin{aligned}\mathbf{A}_k=\mathbf{I}_{N_t}+\sum_{j=1}^{K}\lambda_j\mathbf{h}_j\mathbf{h}_j^H-\sum_{j=M+1}^{K}\mu_j\mathbf{h}_j\mathbf{h}_j^H\\-\left(\frac{1}{\tau}+1\right)\lambda_k\mathbf{h}_k\mathbf{h}_k^H,\quad\forall k.\end{aligned} \tag{B.5}$$

Because the problem is feasible and the primal problem and the dual problem are strongly coupled, the duality gap is zero. Let $\{\lambda_k^*\},\{\mu_k^*\}$ denote the optimal dual solution to problem (B.2); accordingly, we define

$$\begin{aligned}\mathbf{A}_k^*=\mathbf{I}_{N_t}+\sum_{j=1}^{K}\lambda_j^*\mathbf{h}_j\mathbf{h}_j^H-\sum_{j=M+1}^{K}\mu_j^*\mathbf{h}_j\mathbf{h}_j^H\\-\left(\frac{1}{\tau}+1\right)\lambda_k^*\mathbf{h}_k\mathbf{h}_k^H,\quad\forall k.\end{aligned} \tag{B.6}$$

According to the KKT conditions, we get

$$\begin{aligned}\mathbf{A}_k^*-\mathbf{S}_k=0,\\\mathbf{S}_k\mathbf{X}_k^*=0,\\\forall k,\end{aligned} \tag{B.7}$$

where the first equation is obtained by equating the gradient of the Lagrangian of problem (B.3) (with respect to $\mathbf{X}_k$) to zero and the second equation is the complementary condition for $\mathbf{X}_k\succeq 0,\forall k$. From (B.7), we have

$$\mathbf{A}_k^*\mathbf{X}_k^*=0,\quad\forall k. \tag{B.8}$$

Moreover, it is observed from (B.4) that the optimal PS solution ρ_k, $k=M+1,\ldots,K$, must be the solution of the following problem:

$$\begin{aligned}\min_{\rho_k}\quad&\frac{\lambda_k^*\sigma_z^2}{\rho_k}+\frac{\mu_k^*e_k}{\zeta_k\left(1-\rho_k\right)}\\\text{s.t.}\quad&0\le\rho_k\le 1,\quad k=M+1,\ldots,K.\end{aligned} \tag{B.9}$$

From (B.9), we can see that if $\lambda_k^*=0$ and $\mu_k^*>0$, the optimal solution will be $\rho_k^*\to 0$. Similarly, if $\mu_k^*=0$ and $\lambda_k^*>0$, then the optimal solution is $\rho_k^*\to 1$. Since $e_k>0,\ \tau>0,\ k=M+1,\ldots,K$, the above two cases cannot happen. Next, we show that $\lambda_k^*=0$ and $\mu_k^*=0$ cannot be true for $k,\ k=M+1,\ldots,K$, by contradiction. Suppose that there exist some $k's$ such that $\lambda_k^*=\mu_k^*=0$. Define a set

$$\Psi\triangleq\left\{k\mid\lambda_k^*=0,\ \mu_k^*=0,\ M+1\le k\le K\right\}, \tag{B.10}$$

where $\Psi\neq\emptyset$. Define

$$\mathbf{B}_k^*=\mathbf{I}_{N_t}+\sum_{j=1}^{K}\lambda_j^*\mathbf{h}_j\mathbf{h}_j^H-\sum_{j=M+1}^{K}\mu_j^*\mathbf{h}_j\mathbf{h}_j^H. \tag{B.11}$$

Then $\mathbf{A}_k^*$ can be expressed as

$$\mathbf{A}_k^*=\begin{cases}\mathbf{B}^*, & \text{if } k\in\Psi,\\ \mathbf{B}^*-\left(\frac{1}{\tau}+1\right)\lambda_k^*\mathbf{h}_k\mathbf{h}_k^H, & \text{otherwise.}\end{cases} \tag{B.12}$$

Given $\mathbf{A}_k^*\succeq 0,\ -(1/\tau+1)\lambda_k^*\mathbf{h}_k\mathbf{h}_k^H\preceq 0$, we have $\mathbf{B}^*\succeq 0$. In the following, we show that $\mathbf{B}^*\succ 0$ by contradiction. Suppose that the minimum eigenvalue of $\mathbf{B}^*\succeq 0$ is zero. Then, there exists at least $\mathbf{x}\neq 0$ such that $\mathbf{x}^H\mathbf{B}^*\mathbf{x}=0$. According to (B.12), it follows that

$$\mathbf{x}^H\mathbf{A}_k^*\mathbf{x}=-\left(\frac{1}{\tau}+1\right)\lambda_k^*\mathbf{x}^H\mathbf{h}_k\mathbf{h}_k^H\mathbf{x}\succeq 0,\quad k\notin\Phi. \tag{B.13}$$

Note that when $k\notin\Psi$, we have $\lambda_k^*>0$; hence, according to (B.13), we get $|\mathbf{h}_k^H\mathbf{x}|^2\le 0,\ k\notin\Psi$. It thus follows that

$$\left|\mathbf{h}_k^H\mathbf{x}\right|^2=0,\quad k\notin\Psi. \tag{B.14}$$

Thus, we have

$$\begin{aligned}\mathbf{x}^H\mathbf{B}^*\mathbf{x}&=\mathbf{x}^H\left(\mathbf{I}_{N_t}+\sum_{j=1}^{K}\lambda_j^*\mathbf{h}_j\mathbf{h}_j^H-\sum_{j=M+1}^{K}\mu_j^*\mathbf{h}_j\mathbf{h}_j^H\right)\mathbf{x}\\&=\mathbf{x}^H\left(\mathbf{I}_{N_t}+\sum_{j=1}^{M}\lambda_j^*\mathbf{h}_j\mathbf{h}_j^H\right)\mathbf{x}\succeq\mathbf{x}^H\mathbf{x}>0\end{aligned} \tag{B.15}$$

which contradicts $\mathbf{x}^H\mathbf{B}^*\mathbf{x}=0$. Thus we have $\mathbf{B}^*\succ 0$; that is, $\mathrm{rank}(\mathbf{B}^*)=N_t$; then, from (B.12), we have $\mathrm{rank}(\mathbf{A}_k^*)=N_t,\ k\in\Psi$. However, according to (B.8), we have $\mathbf{X}_k^*=0$; it is easily verified that $\mathbf{X}_k^*=0$ cannot be optimal for problem (B.1). Thus, we have

$$\begin{aligned}\mathbf{B}_k^*=\mathbf{I}_{N_t}+\sum_{j=1}^{K}\lambda_j^*\mathbf{h}_j\mathbf{h}_j^H-\sum_{j=M+1}^{K}\mu_j^*\mathbf{h}_j\mathbf{h}_j^H,\\\mathbf{A}_k^*=\mathbf{B}^*-\left(\frac{1}{\tau}+1\right)\lambda_k^*\mathbf{h}_k\mathbf{h}_k^H,\quad k=M+1,\ldots,K.\end{aligned} \tag{B.16}$$

Because of $\mathrm{rank}(\mathbf{B}^*)=N_t$, it follows that $\mathrm{rank}(\mathbf{A}_k^*)\ge N_t-1,\ k=M+1,\ldots,K$. We have seen that $\mathbf{X}_k^*=0$ is not optimal for problem (B.1); then we have $\mathrm{rank}(\mathbf{A}_k^*)=N_t-1,\ k=M+1,\ldots,K$. As for the ID users $k,\ k=1,\ldots,M$, it is equivalent to $\mu_k^*=0,\ k=1,\ldots,M$. Then the proof is similar to the situation when $k=M+1,\ldots,K$. Thus, we can conclude that $\mathrm{rank}(\mathbf{X}_k^*)=1,\ \forall k$; then the proof is completed.

C. The Closed-Form Solution of ZF Scheme

We can see from the bisection method that, for given τ, the inner optimization problem is transformed into the problem of minimizing the total transmit power:

$$\begin{aligned}
\min_{p_k,\rho} \quad & \sum_{k=1}^{K} p_k \\
\text{s.t.} \quad & G_{k,k}p_k \geq \tau\sigma_k^2, \quad k = 1,\ldots,M, \\
& G_{k,k}p_k \geq \tau\sigma_k^2 + \frac{\tau\sigma_z^2}{\rho_k}, \\
& \qquad\qquad k = M+1,\ldots,M+N, \\
& \zeta_k(1-\rho_k)\left(G_{k,k}p_k + \sigma_k^2\right) \geq e_k, \\
& \qquad\qquad k = M+1,\ldots,M+N, \\
& 0 \leq \rho_k \leq 1, \quad k = M+1,\ldots,M+N, \\
& \zeta_k\left(G_{k,k}p_k + \sigma_k^2\right) \geq e_k, \\
& \qquad\qquad k = M+N+1,\ldots,K, \\
& p_k \geq 0, \quad \forall k.
\end{aligned} \tag{C.1}$$

Now the variables p_k and ρ_k of user k are decoupled and do not appear in other users' constraints. Hence, problem (C.1) can be decomposed into K subproblems as follows:

$$\begin{aligned}
\min_{p_k} \quad & p_k \\
\text{s.t.} \quad & G_{k,k}p_k \geq \tau\sigma_k^2, \quad k = 1,\ldots,M, \\
& p_k \geq 0, \quad k = 1,\ldots,M,
\end{aligned} \tag{C.2}$$

$$\begin{aligned}
\min_{p_k,\rho_k} \quad & p_k \\
\text{s.t.} \quad & \frac{1}{\tau}G_{k,k}p_k \geq \sigma_k^2 + \frac{\sigma_z^2}{\rho_k}, \\
& \qquad\qquad k = M+1,\ldots,M+N, \\
& \zeta_k(1-\rho_k)\left(G_{k,k}p_k + \sigma_k^2\right) \geq e_k, \\
& \qquad\qquad k = M+1,\ldots,M+N, \\
& 0 \leq \rho_k \leq 1, \quad k = M+1,\ldots,M+N, \\
& p_k \geq 0, \quad k = M+1,\ldots,M+N,
\end{aligned} \tag{C.3}$$

$$\begin{aligned}
\min_{p_k} \quad & p_k \\
\text{s.t.} \quad & \zeta_k\left(G_{k,k}p_k + \sigma_k^2\right) \geq e_k, \\
& \qquad\qquad k = M+N+1,\ldots,K, \\
& p_k \geq 0, \quad k = M+N+1,\ldots,K.
\end{aligned} \tag{C.4}$$

For problem (C.2), we have $p_k^* = \tau\sigma_k^2/G_{k,k}$, $k = 1,\ldots,M$. And, for problem (C.4), we have $p_k^* = (\gamma - \sigma_k^2)/G_{k,k}$, $k = M+N+1,\ldots,K$. Then, we show that, for problem (C.3), with the optimal solution p_k^* and ρ_k^*, the SINR constraint and harvested power constraint should both hold with equality by contradiction. First, suppose that both the two SINR and harvested power constraints are not tight, given the solution $\overline{p}_k^*, \overline{\rho}_k^*$. Thus, there must exist an α_k, $0 < \alpha_k < 1$, such that, with the new solution $p_k = \alpha_k\overline{p}_k^*$, $\rho_k = \overline{\rho}_k^*$, either the SINR or harvested power constraint is tight. Moreover, with this new solution, the transmission power is reduced, which contradicts the fact that $\overline{p}_k^*$ and $\overline{\rho}_k^*$ are optimal for problem (C.3). Thus, the case where both the SINR and harvested power constraints are not tight is not valid. Next, consider the case when the SINR constraint is tight but the harvested power constraint is not tight. In this case, we can increase the value of $\overline{\rho}_k^*$ by a sufficiently small amount such that both the SINR and harvested power constraints become nontight. Then, we can conclude that this case cannot be true either. Similarly, we can show that the case where the harvested power constraint is tight but the SINR constraint is not tight cannot be true. To summarize, the SINR and harvested power constraints must both hold with equality. Hence, we have the following equation:

$$\rho_k = \frac{\tau\sigma_z^2}{G_{k,k}p_k - \tau\sigma_k^2} = 1 - \frac{e_k}{\zeta_k\left(G_{k,k}p_k + \sigma_k^2\right)}. \tag{C.5}$$

Substituting x_i, α, β, and γ into (C.5), we can get

$$\rho_i = 1 - \frac{\gamma}{x_i} = \frac{\beta}{x_i - \alpha}. \tag{C.6}$$

Then, there are two distinct solutions:

$$\begin{aligned}
x_1 &= \frac{1}{2}\left(\alpha + \beta + \gamma - \sqrt{(\alpha+\beta+\gamma)^2 - 4\alpha\gamma}\right), \\
x_2 &= \frac{1}{2}\left(\alpha + \beta + \gamma + \sqrt{(\alpha+\beta+\gamma)^2 - 4\alpha\gamma}\right).
\end{aligned} \tag{C.7}$$

Given the existence of ID/EH users, $0 < \rho_k < 1$ should be satisfied. Thus, according to (C.6), we have $x_i > \gamma$ and $x_i > \alpha + \beta$; that is, $x_i > \max(\gamma, \alpha+\beta)$. We can easily show that $x_1 < \max(\gamma, \alpha+\beta)$ and $x_2 > \max(\gamma, \alpha+\beta)$, which imply that the optimal solution is $x^* = x_2$. Thus, we can derive the optimal solution $p_k^* = (1/G_{k,k})(x^* - \sigma_k^2)$ and $\rho_k^* = 1 - \gamma/x^*$. Then, check whether the constraint $\sum_{k=1}^{K} p_k^* \leq P^{\max}$ is tight; if it is not, update until the constraint is tight.

Notations

Boldface uppercase letter **X**:	Matrix
Boldface lowercase letter **x**:	Column vector
Lowercase letter x:	Scalar
S:	Square matrix
$\mathbf{S}^{1/2}$:	Square root of **S**
trace(**S**):	The trace of **S**
$\mathbf{S} \succeq 0$:	**S** is a positive semidefinite matrix
A:	Arbitrary-size matrix **A**
rank(**A**):	The rank of **A**
$\mathbf{A}^*$:	Complex conjugate of **A**
$\mathbf{A}^T$:	Transpose of **A**
$\mathbf{A}^H$:	Hermitian (conjugate) transpose of **A**
$\text{diag}\{x_1, x_2, \ldots, x_M\}$:	$M \times M$ diagonal matrix with $x_1, x_2, \ldots, x_M$ being the diagonal elements
$\mathbf{I}_n$:	$n \times n$ identity matrix
$\|\cdot\|$:	The Euclidean norm of a complex vector
$\|\cdot\|$:	The absolute value of a complex scalar
$\mathbf{x} \sim \mathcal{CN}(\mu, \sigma^2)$:	The distribution of a circularly symmetric complex Gaussian (CSCG) random vector **x** with mean μ and covariance matrix σ^2
"$\sim$":	Stands for "distributed as"
$\mathbb{C}^{m \times n}$:	The space of $m \times n$ complex matrices.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant no. 61471408, in part by the Fundamental Research Funds for the Central Universities under Grant no. 2016YXMS293, and in part by CCF-Tencent Grant no. IAGR20160106.

References

[1] D. Minoli, K. Sohraby, and B. Occhiogrosso, "IoT considerations, requirements, and architectures for smart buildings – energy optimization and next generation building management systems," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 269–283, Feb 2017.

[2] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): a vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[3] S. Oh, J. Lee, and S. Park, "Energy efficient and accurate monitoring of large-scale diffusive objects in internet of things," *IEEE Communications Letters*, vol. 21, no. 3, pp. 612–615, 2017.

[4] H. Ju and R. Zhang, "Throughput maximization in wireless powered communication networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 1, pp. 418–428, 2014.

[5] L. Liu, R. Zhang, and K. C. Chua, "Multi-antenna wireless powered communication with energy beamforming," *IEEE Transactions on Communications*, vol. 62, no. 12, pp. 4349–4361, 2014.

[6] Y. Che, J. Xu, L. Duan, and R. Zhang, "Multiantenna wireless powered communication with cochannel energy and information transfer," *IEEE Communications Letters*, vol. 19, no. 12, pp. 2266–2269, 2015.

[7] B. Xu, Y. Zhu, and R. Zhang, "Optimal power allocation for a two-link interference channel with SWIPT," in *Proceedings of the 6th International Conference on Wireless Communications and Signal Processing (WCSP '14)*, pp. 1–5, IEEE, Hefei, China, October 2014.

[8] R. Zhang and C. K. Ho, "MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 1989–2001, 2013.

[9] X. Zhou, R. Zhang, and C. K. Ho, "Wireless information and power transfer: architecture design and rate-energy tradeoff," *IEEE Transactions on Communications*, vol. 61, no. 11, pp. 4754–4761, 2013.

[10] S. Timotheou, I. Krikidis, G. Zheng, and B. Ottersten, "Beamforming for MISO interference channels with QoS and RF energy transfer," *IEEE Transactions on Wireless Communications*, vol. 13, no. 5, pp. 2646–2658, 2014.

[11] S. Timotheou, I. Krikidis, and B. Ottersten, "MISO interference channel with QoS and RF energy harvesting constraints," in *Proceedings of the IEEE International Conference on Communications (ICC '13)*, pp. 4191–4196, IEEE, Budapest, Hungary, June 2013.

[12] Q. Shi, L. Liu, W. Xu, and R. Zhang, "Joint transmit beamforming and receive power splitting for MISO SWIPT systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 6, pp. 3269–3280, 2014.

[13] M.-M. Zhao, Y. Cai, Q. Shi, B. Champagne, and M.-J. Zhao, "Robust transceiver design for miso interference channel with energy harvesting," *IEEE Transactions on Signal Processing*, vol. 64, no. 17, pp. 4618–4633, 2016.

[14] Z. Q. Luo, W. Ma, S. A. M. C et al., "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Processing Magazine*, vol. 64, no. 17, pp. 4618–4633, Sept 2010.

[15] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, vol. 52, no. 9, pp. 3936–3964, 2006.

[16] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.

[17] K. Xu, Y. Qu, and K. Yang, "A tutorial on the internet of things: From a heterogeneous network integration perspective," *IEEE Network*, vol. 30, no. 2, pp. 102–108, 2016.

[18] F. Saffre, "Tutorial ii: the green internet of things," in *Proceedings of the 11th International Conference on Innovations in Information Technology (IIT '15)*, pp. XXXVII–XXXVII, Dubai, United Arab Emirates, November 2015.

[19] C. Park, "A Secure and efficient ECQV implicit certificate issuance protocol for the internet of things applications," *IEEE Sensors Journal*, vol. 17, no. 7, pp. 2215–2223, 2017.

[20] J. Xu, L. Liu, and R. Zhang, "Multiuser miso beamforming for simultaneous wireless information and power transfer," *IEEE Transactions on Signal Processing*, vol. 62, no. 18, pp. 4778–4810, 2014.

[21] L. Liu, R. Zhang, and K.-C. Chua, "Wireless information transfer with opportunistic energy harvesting," *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 288–300, 2013.

[22] B. Xu, Y. Zhu, and R. Zhang, "Optimized power allocation for interference channel with SWIPT," *IEEE Wireless Communications Letters*, vol. 5, no. 2, pp. 220–223, 2016.

[23] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Transactions on Signal Processing*, vol. 54, no. 6 I, pp. 2239–2251, 2006.

[24] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[25] M. Grant and S. Boyd, "Cvx: Matlab software for disciplined convex programming, version 2.0 beta," Sept 2012.

[26] B. Gopalakrishnan and N. D. Sidiropoulos, "High performance adaptive algorithms for single-group multicast beamforming," *IEEE Transactions on Signal Processing*, vol. 63, no. 16, pp. 4373–4384, 2015.

[27] J. Lei, Z. Han, M. A. Vazquez-Castro, and A. Hjorungnes, "Secure satellite communication systems design with individual secrecy rate constraints," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 661–671, 2011.

[28] Z. Shen, R. Chen, J. G. Andrews, R. W. Heath Jr., and B. L. Evans, "Low complexity user selection algorithms for multiuser MIMO systems with block diagonalization," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3658–3663, 2006.

[29] E. A. Jorswieck, E. G. Larsson, and D. Danev, "Complete characterization of the Pareto boundary for the miso interference channel," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, part 2, pp. 5292–5296, 2008.

[30] F. Alizadeh and D. Goldfarb, "Second-order cone programming," *Mathematical Programming—Series B*, vol. 95, pp. 3–51, 2003.

[31] Z. Luo, M. Zhao, S. Liu, and Y. Liu, "Greville-to-Inverse-Freville algorithm for V-BLAST systems," in *Proceedings of the IEEE International Conference on Communications (ICC '06)*, pp. 4214–4218, Istanbul, Turkey, July 2006.

[32] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge university press, 2012.

[33] F. Rashid-Farrokhi, "Joint optimal power control and beamforming in wireless networks using antenna arrays," *IEEE Transactions on Communications*, vol. 46, no. 10, pp. 1313–1324, 1998.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 8959356, 13 pages
https://doi.org/10.1155/2017/8959356

Research Article

SmartFix: Indoor Locating Optimization Algorithm for Energy-Constrained Wearable Devices

Xiaoliang Wang,[1,2] Ke Xu,[1,2] and Ziwei Li[1,2]

[1]*Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*
[2]*Tsinghua National Laboratory for Information Science and Technology, Beijing 100084, China*

Correspondence should be addressed to Ke Xu; xuke@mail.tsinghua.edu.cn

Received 21 March 2017; Accepted 11 June 2017; Published 26 July 2017

Academic Editor: Zhe Yang

Indoor localization technology based on Wi-Fi has long been a hot research topic in the past decade. Despite numerous solutions, new challenges have arisen along with the trend of smart home and wearable computing. For example, power efficiency needs to be significantly improved for resource-constrained wearable devices, such as smart watch and wristband. For a Wi-Fi-based locating system, most of the energy consumption can be attributed to real-time radio scan; however, simply reducing radio data collection will cause a serious loss of locating accuracy because of unstable Wi-Fi signals. In this paper, we present SmartFix, an optimization algorithm for indoor locating based on Wi-Fi RSS. SmartFix utilizes user motion features, extracts characteristic value from history trajectory, and corrects deviation caused by unstable Wi-Fi signals. We implemented a prototype of SmartFix both on Moto 360 2nd-generation Smartwatch and on HTC One Smartphone. We conducted experiments both in a large open area and in an office hall. Experiment results demonstrate that average locating error is less than 2 meters for more than 80% cases, and energy consumption is only 30% of Wi-Fi fingerprinting method under the same experiment circumstances.

1. Introduction

With massive application demands, recent decade has witnessed remarkable achievements on indoor localization based on different schemes [1–10], such as IOT [11, 12] and smart home system [13]; indoor localization plays a significant role. Thanks to the evident convenience where Wi-Fi technology has been widely used, indoor localization strategy by using Wi-Fi RSS has caught extensive attention since 2000. This strategy deploys the localization algorithm on smart terminals such as smartphones or tablets and assumes that the devices locate at the place exactly the same as the tester does. Hence, we can locate through locating the smart devices. Successful as it is, concerning relative special applications (such as smart home or office settings), it may encounter some restrictions due to false binding relationship between users and traditional smart devices. Luckily, wearable devices (such as smart watches, wristbands, rings, and glasses) have been developed rapidly and became increasingly popular in recent years. Their closer binding relationship and longer usage time improve the applicability of indoor localization algorithm.

On the other hand, with new opportunities due to the popularity of wearable devices and indoor localization technology, challenges have yet been proposed. For example, wearable devices have lower capability in computing than traditional devices. Also, their less storage capacity and simpler hardware functionality block the direct usage of indoor localization technology on wearable devices. More importantly, considering current battery capacity of wearable devices is only one-tenth of that of traditional smart devices. The improvements on high energy consumption of localization algorithms must be the top issue to make localization technology on wearable devices possible.

In one single real-time localization phase, energy consumption includes two main parts: computation and collection of Wi-Fi fingerprints (the same as Wi-Fi RSS). Because of various kinds of influences including random fluctuations and multipath loss due to walls, furniture, and people moving, it demands more RSS collection to ensure indoor localization accuracy. According to our experiment, energy consumption caused by real-time collection occupies 99% of all in that of localization algorithm (See Table 1).

Table 1: Energy consumption in localization phase (mAh).

Device	Method	Signal collection	Calculation	Proportion
HTC One	SmartFix	0.1917	0.0005	99.74%
	MoLoc	0.3105	0.0003	99.90%
	Wi-Fi	0.1917	0.0003	99.84%
Moto 360	SmartFix	0.00806	0.00036	95.68%
	MoLoc	0.01310	0.00008	99.38%
	Wi-Fi	0.00806	0.00008	98.98%

The question is how to minimize the frequency of collection and at the same time guarantee acceptable accuracy? Nowadays, auxiliary sensors, for example, gyroscope, accelerate sensor, infrared, and camera, are introduced to dynamically adjust the collection frequency, reduce the energy consumption, and ensure the localization accuracy. The assisted sensors collect the mobile information and help improve the localization results, such as MoLoc [14], a strategy which implements inertial sensors to directly obtain the information of the user. However, all the methods mentioned above have to function with the help of auxiliary hardware and sensors, and, according to our survey, those devices will not cause evident energy burden to traditional smart devices, but, in terms of wearable devices which are sensitive to energy consumption, energy consumption of such sensors is unneglectable. Therefore, considering these constraints, current strategies should be further improved.

In this paper, we propose a novel indoor localization strategy, SmartFix. It can cooperate with any indoor localization technology based on Wi-Fi RSS and enhance the accuracy with a very little extra energy cost of calculation but a large save of signal collection energy cost. SmartFix is an indoor localization technology free from auxiliary sensors. Aided with machine-learning algorithm, we obtain the relative features given the trajectories of users in certain areas and modify the localization results. Compared to traditional localization technology, SmartFix can achieve certain level of accuracy provided with only one RSS value. It also performs well in energy saving. In comparison to original Wi-Fi fingerprinting method, SmartFix can save 70% of energy by achieving same localization accuracy.

2. Motivation and Challenges

The main problem for designing wearable-based indoor localization technology is to improve the energy efficiency. To find out the key point of that, we conducted experiments on HTC One and Moto 360 and recorded and figured out the present components of localization algorithms. After that, according to the results, we analyzed the common questions on Wi-Fi-based algorithm and introduced how SmartFix takes advantage of history trajectories to improve the performance in localization.

2.1. Analysis of Energy Consumption. We divide the localization phase into two parts, that is, the signal collection and the calculation. We did experiments on Moto 360 Smartwatch and HTC One Smartphone. For the experiments that we discuss in this paper, we scan Wi-Fi channel numbers 1, 6, and 13 for every location and scan and collect RSS data of all Wi-Fi hotspots nearby in every 3 seconds. For the build-in sensor (accelerometer and gyroscope), we collect the data with 50 Hz and 500 Hz sampling rate on Smartphone and 25 Hz and 200 Hz on Smartwatch.

Firstly, we recorded the capacity of battery during the RSS signal collection stage and calculated the average power consumption for collecting one set of RSS signal data. Secondly, we operated the locating algorithm on Smartwatch and Smartphone for many times, recorded the capacity of battery, and calculated the average energy consumption of running the algorithm one time for one location. Data analysis shows that the proportion of energy consumption of signal collection and that of calculation can reach up to 99% (see Table 1).

In fact, due to the low stability of Wi-Fi signal detection and the influence of environmental factors, the signal strength always fluctuates. One method to tackle this problem is to use probability estimation which relies on a special amount of real-time data. We did experiments to verify that increasing the amount of real-time data and using probability estimation will enhance localization accuracy.

We implemented the basic Wi-Fi fingerprinting method (K-NN) and FreeLoc [6] which relies on the relative RSS order of different APs. Since there are only several APs in the experiment, the localization performance of FreeLoc is not good or even worse compared to that applying K-NN. We use simple probability estimation to simulate the effect of different amounts of real-time data. Sharing the same real-time data, both methods will lose 40% accuracy when the number of real-time data reduces from nine to two per location because of the unstable Wi-Fi signals [15]. Although the amount of real-time data is very important to locating accuracy, it will certainly cause more energy consumption.

We utilize the features of user motion to meet the demands of energy efficiency and the locating accuracy. There are several locating methods using user motion to help optimize the locating performance. The existing methods mainly use built-in sensors to directly obtain motion information which will add the diversity built by RSS fingerprints and help distinguish different locations. MoLoc uses digital compass and accelerometer to make user motion available. Judging from our experiment, it shows that although the energy consumption of built-in sensors does not seem to be a burden to traditional smart devices such as smart phones, it cannot be negligible when deployed in wearable devices which are more sensitive to energy efficiency. Table 2 shows the proportions

Table 2: Energy consumption of different parts with MoLoc.

Device	Type	Wi-Fi	Built-in sensor	Calculation
HTC One	mAh	0.1916	0.1189	0.0003
	proportion	61.60%	38.30%	0.10%
Moto 360	mAh	0.00806	0.00504	0.00008
	proportion	61.17%	38.21%	0.62%

of energy consumption of different parts while carrying out a localization process with MoLoc on Moto 360 Smartwatch and HTC One Smartphone. Despite the differences in energy consumption of two devices, their proportion is nearly the same. The result validates our conclusion.

2.2. Features of User Motion in History Trajectory. In avoidance of extra energy consumption from build-in sensors, SmartFix does not require data directly from sensors but instead advances its localization accuracy by implementing machine-learning algorithms on human motion feature along the trajectories.

Consistent location coordinates indicate information of trajectory; as a whole, we can learn its motion features in certain areas. Firstly, human motion is limited by physical space, and it also depends on his/her destination. What is more, whether a person is familiar with the surroundings should also be considered when learning his/her motion features. We believe that, given certain area, one has specific destination and is quite familiar with that area, and his/her trajectories are bound to follow some fixed rules and display some similar patterns. If we can obtain and learn from this motion mode, it will do much good to help improve the performance and decrease the cost in localization, which is a big deal to Location Based Service. For example, under smart home circumstances, by obtaining the location information of users, smart home management system can automatically operate on temperature, humidity, lights, videos, and security system and manage various smart devices, which can enhance its applicability and become more user-friendly.

On the other hand, concerning an unfamiliar environment, though the destination is specific, when one person first steps into this area, there will be relative more random moves. Typical examples such as shopping mall, train station, and airport, in which random motions occur much more often and hence motion features which we pay attention to cannot be extracted. However, after awareness of the area when giving a large bunch of records of trajectories and eliminating those random variables, motion features can be captured by majority of actual trajectories.

According to conclusion mentioned before, we figure out when we provide Location Based Service in a given area; we can benefit from history information on the trajectories or those similar patterns that the majority follows to improve the localization result, which is key to this paper.

3. SmartFix Architecture

This section elaborates the design for SmartFix localization algorithm. Also, there is an specific example at the end of the part to help explain the operation and details for SmartFix.

3.1. Feature Selection of History Trajectory. As a localization strategy is aided by history information, the first problem we are faced with is how to efficiently collect the location information or coordinates of points. In this paper, we take advantage of pedometer and gyroscope equipped on smart devices to collect the path and turning angles of trajectories. Firstly, we conducted small scale experiments recording the trajectories of users, which is labelled as Test A with 500 records. On this basis, we implement machine-learning algorithms to learn from these data and they automatically cluster into three feature values which are probability of continuous turnings, range of turning angles, and proportions of turning during one trajectory. All these features cannot be presented by Wi-Fi fingerprint matching algorithms due to its random distributed RSS value, which can cause irregular fluctuations on the localization.

For further testing of those three features, we chose two areas as use cases. One is a laboratory with 200 square meters. And another is an open area with a square ring shape. It is about 1898 square meters. Total 2000 records have been collected and saved as Test B. As experiment results, all these three features also follow the rules in a larger scale test base. Therefore, we believe that, through learning from these three features, we can efficiently eliminate error fluctuations as much as possible and hence increase localization accuracy. Below we will separately discuss three modes and the patterns they follow. Wi-Fi RSS value will be influenced by many factors such as multipath effects, temperature, and humidity change or electromagnetic change. All these effects pose a negative impact on localization accuracy. Continuous turnings happen much more often under the circumstances by applying Wi-Fi RSS localization technology. However, according to our experiment, this abnormal mode can hardly be spotted in actual trajectory as shown in Figure 1.

In this case, a trajectory moves from point (1) to point (5) keeping the straight direction. However, based on Wi-Fi detection, the result displayed is located randomly on either side of the actual trajectory. Here comes the question that is it safe enough for us to conclude that the continuous turning pattern given by Wi-Fi detection is only a result from error matching and should not be considered by deciding the correct result? To answer this question, we cluster the motion features from Test A and the result shows that it is positive as is shown in Figure 2.

In Figure 2, abnormal turning refers to continuous turning in one single trajectory. Like the trajectory from point (1) to point (4) through point (2) and point (3) in Figure 1, the directions of paths (1)-(2), (2)-(3), (3)-(4) are changing all the time. This form of path located by basic Wi-Fi RSS method is more likely to contain the error points caused by unstable Wi-Fi signal, and that is why we focus on the abnormal turning

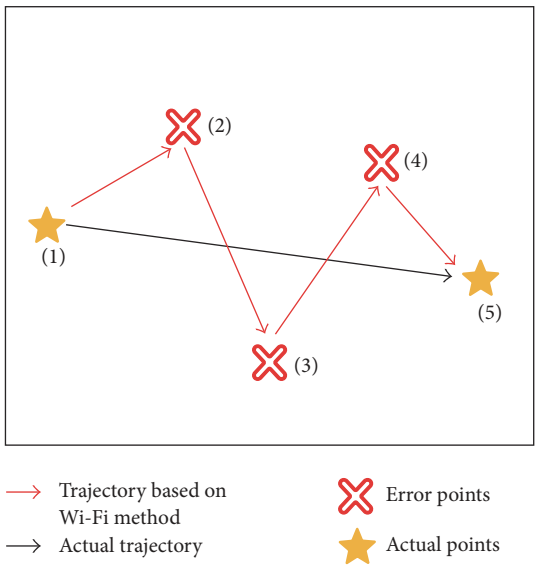


Figure 1: User motion A.

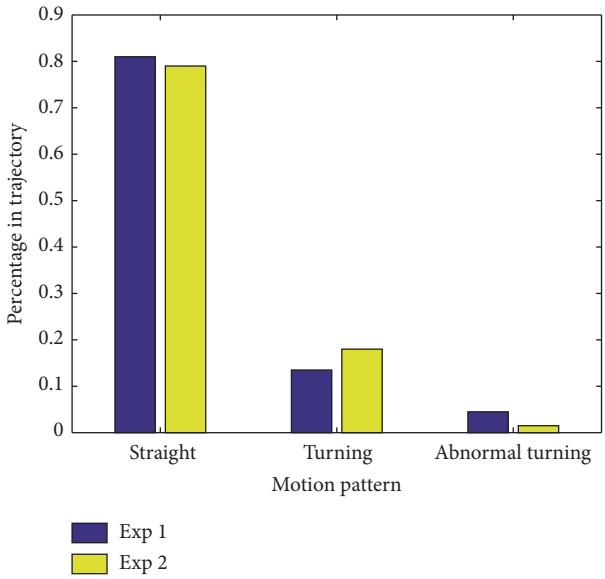


Figure 2: Percentage of motion pattern.

rate in actual trajectory. As a result, 81% of actual trajectory remains straight and about 13% of moves are turnings. The rest, lower than 6% of situations, might present a continuous turning in a single trajectory. In order to be more convincing, we tested with a much larger scale of data of Test B and the result is accorded with our presumption, which indicates that by eliminating such abnormal patterns in the trajectories can help improve the performance in localization.

Throughout the clustering for Test A, distribution of turning angles is another evident feature learned from the trajectories. Under traditional Wi-Fi detection, there is always a problem in locating turning points. On the other hand, range of turning angles depends much on the physical distribution of the place. For example, if the area is relatively spacious such as the building hall, trajectories will display the patterns with

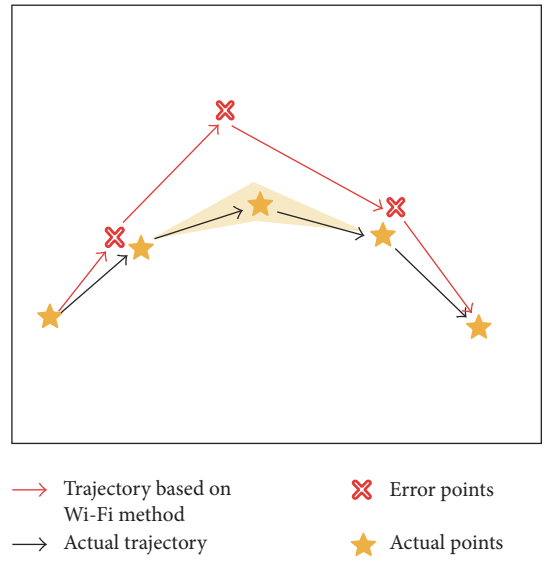


Figure 3: User motion B.

much more straight moves and turnings with small angles. Or, if the place is physically limited in space such as laboratory and office, straight moves will be cut apart and more turnings with larger angles will be presented. Hence, this feature is highly dependent on the certain place. It might be applicable within areas which share similar physical distribution. But in terms of even large scale, it is not fully representative for data from a given area to be widely used as a test base. Its limitation on robustness from different areas is presented in the following experiments; however, there is no denying for its efficiency. First of all, we clustered on Test A and obtained the distribution of turning angles from 0 to 180. Due to the dominance of straight moves, angles less than 10 degree will be considered straight moves in this operation. After that, with a larger scale Test B, including experiment areas 1 and 2, we separately clustered the trajectories and results are shown in Figure 3.

In this experiment, we find out that there is a different range of turning angles for each od the experiment areas. The majority of each of the areas are different because of their specific physical distribution. In the case of experiment area 1, turning angles are mainly clustered around 30 degrees, with maximum value of 40 degrees. 90% of all are distributed within the range of 20 to 40 degrees. Meanwhile, for area 2, the majority stays at 50 degrees, with most of the cases being lower than 60 degrees. And turning angles in the range from 30 to 50 occupy 90% of total records. Based on these rules, as is shown in Figure 4, we can improve the localization by modifying the turning points with more convincing turning angles. The third feature we captured from Test A is the proportion of turning in a whole trajectory. As is shown in Figure 5, we calculate the percentage of turnings in one trajectory given enough length and set it as a threshold. During the decision process, we calculate the percentage of turnings of the trajectory and compare it with our threshold. If that value excesses the threshold, we manually put a penalty value on its probability to prevent the situation for turnings

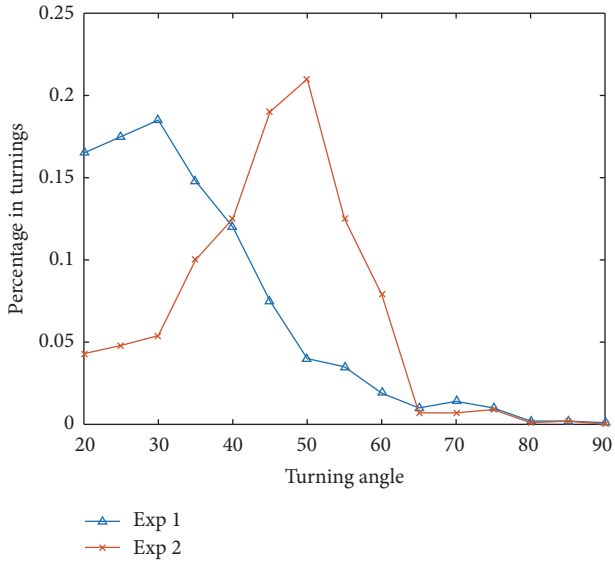


Figure 4: Turning angle distribution.

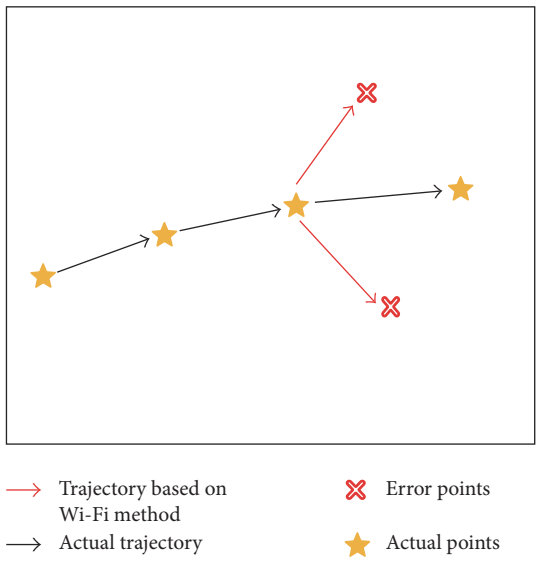


Figure 5: User motion C.

with high frequency. Through this way, it can also help us to adjust the trajectory.

However, like second one, this feature is also limited by physical factors. Directed by this limitation, we conducted another experiment to test its robustness. Different from previous experiment, for those trajectories with less than 10 moves, no valuable information can be extracted and is considered unrepresentative. We select the samples which has much more steps and the result is shown in Figure 6. This can conclude that there will always be a threshold for each experiment area. But these values differ from each other given different physical factors. Meanwhile, the curve displays with a wave shape which indicates that as the length of trajectories increases, the turnings occur regularly in a relatively stable range in both areas 1 and 2. Based on these strategies, we design our localization algorithm, SmartFix.

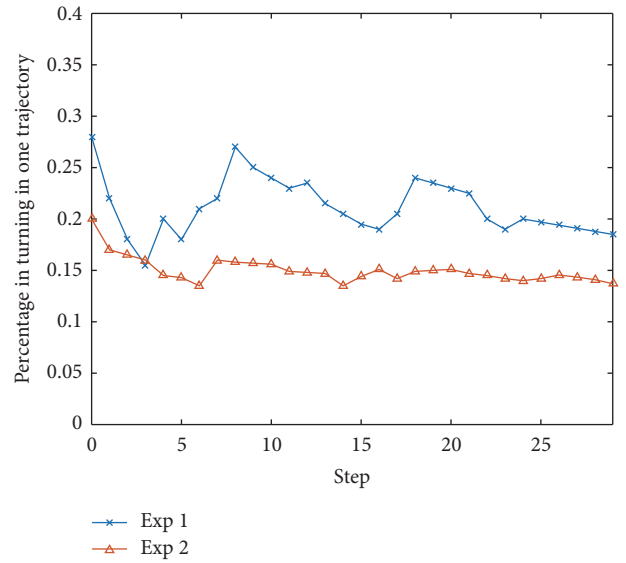


Figure 6: Proportion of turning.

Next part will evaluate the performance of SmartFix in a more detailed degree.

3.2. Background Data and Location Model. SmartFix uses trajectory connectivity and motion tendency to locate and enhance accuracy, so in the collection phase, in addition to the RSS values in each position; SmartFix also uses the relationship between various locations. For example, in a region of N positions, surveyors need to collect the RSS fingerprint information of the APs at each location: $\text{fingerprint}_i = (\text{RSS}_1, \text{RSS}_2, \ldots, \text{RSS}_m)$, where m represents the number of APs. Besides, SmartFix requires relative positions of those N nodes and connectivity relations to develop an indoor location model. The indoor location model can be represented as a weighted adjacency matrix M:

$$M = \begin{pmatrix} \varphi_{1,1} & \varphi_{1,2} & \cdots & \varphi_{1,N} \\ \varphi_{2,1} & \varphi_{2,2} & \cdots & \varphi_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{N,1} & \varphi_{N,2} & \cdots & \varphi_{N,N} \end{pmatrix}, \quad (1)$$

$$\varphi_{i,j} = \begin{cases} k, & \text{loc}_i \text{ and } \text{loc}_j \text{ are adjacent} \\ 0, & \text{loc}_i \text{ and } \text{loc}_j \text{ are not adjacent.} \end{cases}$$

We set k as an integer greater than 0, $\varphi_{i,j} = k$ indicates that location j is on the number of k directions of location i, which will be used for the judgment of motion tendency. The adjacency matrix M indicates the connectivity relations corresponding to indoor locations.

SmartFix assumes that the location model is available from floor plan or direct input from users. There are some studies focusing on the construction of indoor location models which may be used for SmartFix in the future.

On that basis, we set up some auxiliary points located between every two localization points. Their density is much larger than that of localization points and those points do not

Input: *path*
Output: estimated path *path*
(1) $path \leftarrow [\,]$
(2) **for all** $i \in$ every possible points **do**
(3) $PatternRecognization(\,)$
(4) $targetPoint[i] \leftarrow LocalizationPoint$
(5) $replaceP_{err}[i]$ with $targetPoint[i]$
(6) $p_{\text{pre}}[i] = \prod_{i=1}^{n} P_{\text{point}} * P_e$
(7) $P_{\text{cur}}[i] = \prod_{i=1}^{n} P_{\text{point}}$
(8) **for all** $i \in$ every possible points **do**
(9) **if** $(p_{Q[i]} > MAX)$ **then**
(10) $MAX = P_{Q[i]}$
(11) $path \leftarrow P_{Q[i]}max$
(12) **return** *path*

Algorithm 1: SmartFix.

have RSS fingerprints. Due to underlying accuracy limits of Wi-Fi localization, the difference for those auxiliary points only by applying fingerprint matching cannot be told. But judging motion pattern along the trajectory is consistent with all the points not only those localization points. Hence, we add those auxiliary points to help improve the performance in SmartFix.

3.3. Localization Algorithm considering Motion Feature. The core idea of SmartFix is to use the features of indoor human motion, for example, to use the continuity of motion trajectory to modify the moving path. For indoor LBS triggering on smart home scenes, users cannot jump from one to another nonadjacent location. So the current localization result must be related to the last one. For example, the last location along a path should be the reference factor of the possible current location. Judging by the value of the element $\varphi_{i,j}$ in adjacency matrix M, we learn the connectivity of loc_i and loc_j. If the last location loc_i is adjacent to the current location loc_j, which indicates that moving from loc_i to loc_j is possible, the possibility will be multiplied with an enhancement factor. Instead, if loc_i and loc_j are not adjacent, then random error is bound to occur in the current or last measurement; we weaken the probability of the occurrence by multiplying a dull factor to amend the estimated trajectory.

The essence of the SmartFix algorithm is applying pruning and BFS to a tree with fixed height which contains all possible paths and its probability, finding the leaf node with the biggest cumulative probability as the current localization results. The algorithm introduces a queue cur$Q$ containing the results of the possible current paths; the elements of the queue are tuples which hold the current position, the current path, and the cumulative probability. In every localization, it will calculate the cumulative probability Q of each possible path based on the current result and the last queue pre$Q$ and rebuild cur$Q$ for the next locating.

$$Q_n = \begin{cases} Q_{n-1} \times q_i \times P, & \text{loc}_i \text{ is adjacent and matching} \\ Q_{n-1} \times q_i \times R, & \text{loc}_i \text{ is not adjacent} \\ Q_{n-1} \times U, & \text{loc}_i \text{ is adjacent but not matching,} \end{cases} \tag{2}$$

where loc_i is the current position and q_i is the probability of loc_i by the current measurement. First of all, in judging motion patterns, if current motion is a turning, and then we will consider previous pattern, and together we can choose a specific strategy to modify the result. According to Algorithm 1, if there is a left turning followed by a right turning or a right one followed by a left one, considering our strategy in fixing continuous turnings, we assume that this pattern hardly happens under indoor circumstances. We therefore revise this motion into a straight move. Given three points which form these two motions, our method is to find an alternate point which can allow these two motions to be straight moves. Also, each of these points should be linked with its previous one to ensure its connectivity. After that, we generate this new trajectory and replace previous one with it.

Additionally, if either of conditions, left to left, left to right, right to right, and right to left, is satisfied, applying the same rules according to out learning data, we also assume that turning angles should stay at certain range which depends on its place. Therefore, we also search for an alternative to ensure that its turning angle lies in that range.

Besides these two modification strategies, we also introduce another assisted judging rule by setting a threshold for turning proportion. Given enough length of trajectories, if turning proportion excesses that threshold, we manually give its probability a penalty factor to decrease its possibility. Because the threshold is learned by actual trajectories provided in this area, so if there is a higher percentage in turnings, we believe it is wrong matching of Wi-Fi that causes the problem. Therefore, we also add this motion features into our localization strategy.

The core algorithm calculates the probability of each possible trajectory and chooses the one with maximum likelihood to be optimal trajectory for this localization. With RSS matching kept on, we continuously acquire a modified trajectory with updated probability, and, finally, ultimate trajectory with maximum possibility can be obtained as best localization result for this strategy (see Algorithm 1). Our method is to implement BFS to prune a probability tree. The subjects are all the traces given certain length, including all possible trajectory. Current trace with maximum likelihood

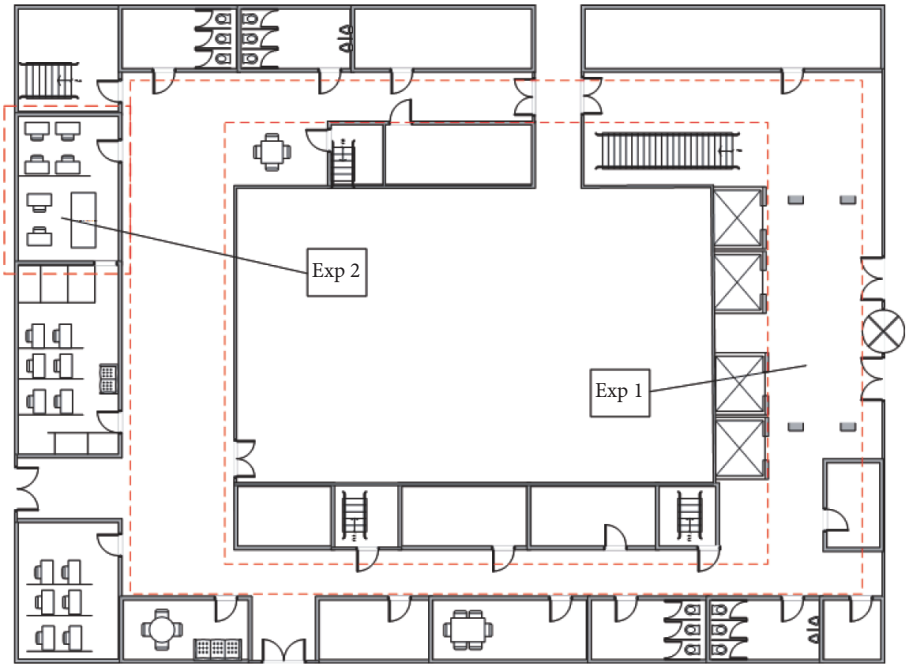


Figure 7: Floor plan of experiments.

will be considered as optimal trace for now. Also, previous likelihood is saved for further judgment. After repeating the process till the end, we will finally get a best trace which fits our strategy best. The strategy of SmartFix has been described by Algorithm 1.

4. Evaluation

This section presents the experimental setups, results, and analyses of our experiment.

4.1. Experiment Settings. We implemented the prototypes of TinyLoc [16], MoLoc [14], and basic Wi-Fi fingerprinting method (k-neighbor algorithm) on Moto 360 2nd-generation Smartwatch (as shown in Figure 10) to compare the localization accuracy and energy efficiency. TinyLoc is one of our previous works which is based on two certain principles to optimize the locating results. TinyLoc is more focused on the energy efficiency than locating accuracy. Certain principles will simplify the modeling and reduce the amount of calculation. In contrast, SmartFix analyzes the history trajectory of people in given area to generate the character value of given physical area and special locations to optimize locating results. Moloc [14] also leverages user motion against unstable Wi-Fi RSS fingerprint. The basic idea of MoLoc is that user motion patterns collected by built-in sensors of mobile phones add to the diversity built by RSS fingerprints and improve the locating accuracy. At the same time, we implemented the prototypes of both original version and optimized version using SmartFix for comparison.

Additionally, considering the variety of indoor environment, we conducted the experiments in an open area of 1898 m^2 and also an office hall of 200 m^2, respectively (as shown in Figures 7, 8, and 9).

Experiment area I is located in an indoor corridor area of 1898 m^2, with 176 locations. By taking 10 RSS samples at each location in the construction phase and the localization phase, respectively, we recorded RSS fingerprints of at most 9 APs, for the purposes of fingerprint database establishment and location estimation. For MoLoc, we recorded the directions and steps of every adjacent location from the digital compass and accelerometer readings for 4 times, three of which are used for building the motion database and one for localization. Experiment area II is located in a laboratory of 13.2 m and 15.6 m with furniture including tables, chairs, server racks, and electronic devices such as computers, servers, and switches, where the electromagnetic environment is complex. We also selected 9 APs of which signals are able to cover the whole area, and 18 reference locations of which spacing is 2–4 m pairwise. The method of data collection and other experimental settings are the same as experimental area I.

4.2. Performance Evaluation. We select MoLoc and TinyLoc for a comparison. These two algorithms both use user motion to enhance accuracy. In this section, we compared the effectiveness of MoLoc, TinyLoc, Smart-MoLoc (using SmartFix to optimize MoLoc), Smart-TinyLoc (using SmartFix to optimize TinyLoc), and basic Wi-Fi fingerprinting method in aspects of localization accuracy and energy efficiency both on classic smart devices and on portable devices. In this section, results and detailed analyses will be represented.

4.2.1. Locating Accuracy. We compared the localization accuracy of SmartFix, MoLoc, Smart-MoLoc (using SmartFix to optimize MoLoc), Smart-TinyLoc (using SmartFix to

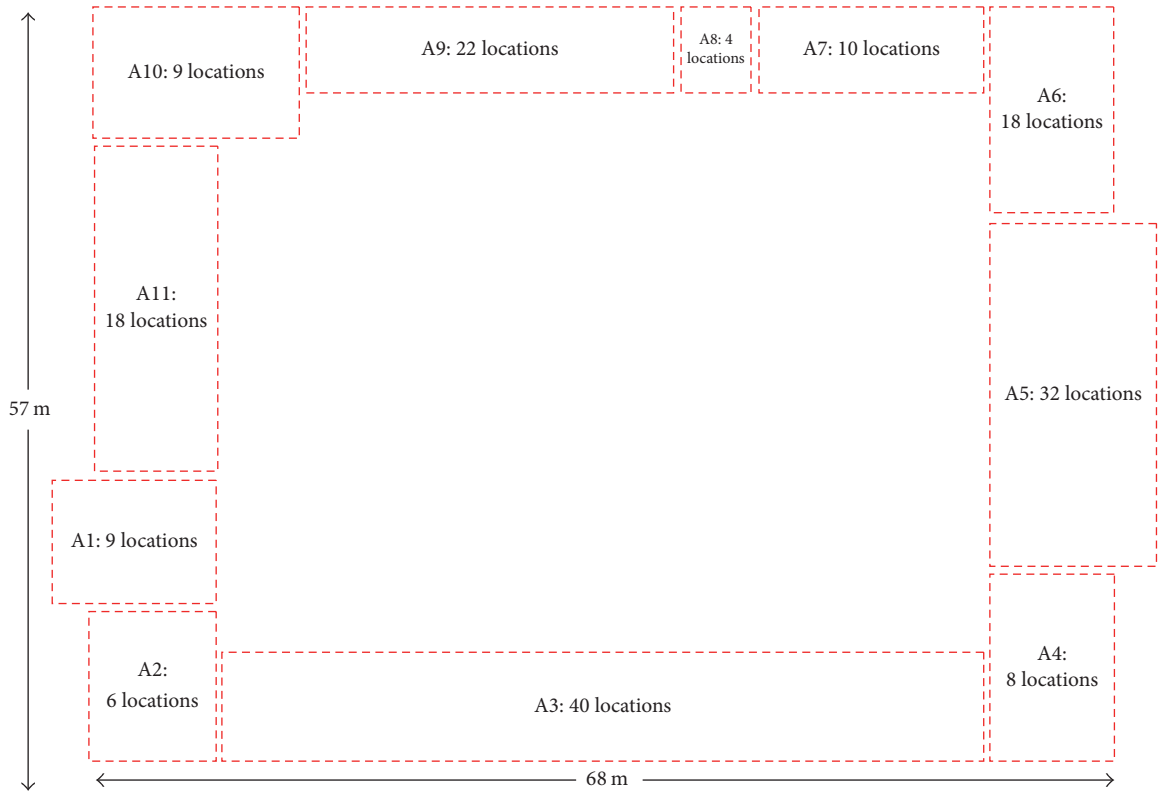


Figure 8: Exp I.

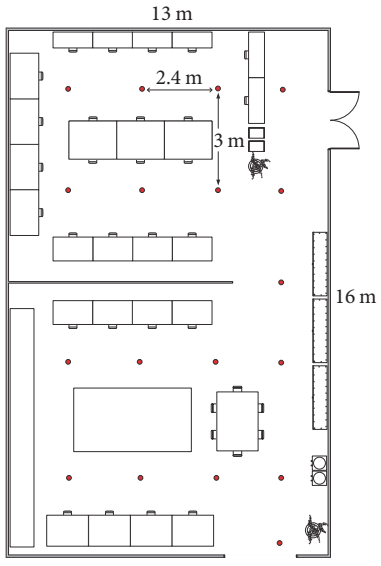


Figure 9: Exp II.

Figure 10: Moto 360.

optimize TinyLoc), and Wi-Fi fingerprinting method from two perspectives: cumulative distribution function (CDF) of average errors and locating accuracy. Then, we analyzed the locating accuracy with different number of APs, trying to explain the difference between SmartFix and MoLoc (or TinyLoc) with respect to their utilization of user motion.

$$D_{\text{loc}} = \frac{\sum_{n=1}^{\text{step}} \text{Distance}\left(\text{loc}_n^{\text{estimated}}, \text{loc}_n^{\text{real}}\right)}{\text{step}},$$
$$p_x = \frac{\text{count}\left(\text{loc} \in \text{path}_{\text{step}=1,2,3,\ldots,N} \wedge D_{\text{loc}} \leq x\right)}{\sum \text{length}\left(\text{path}_{\text{step}} = 1, 2, 3, \ldots, N\right)}. \quad (3)$$

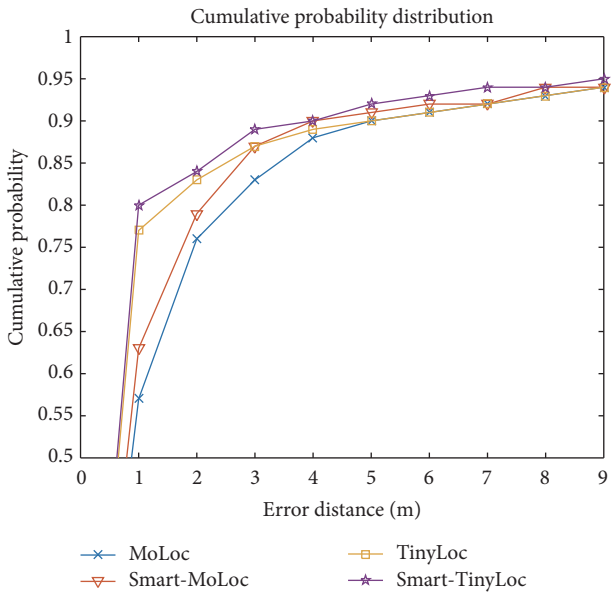


Figure 11: CDF of errors in area I.

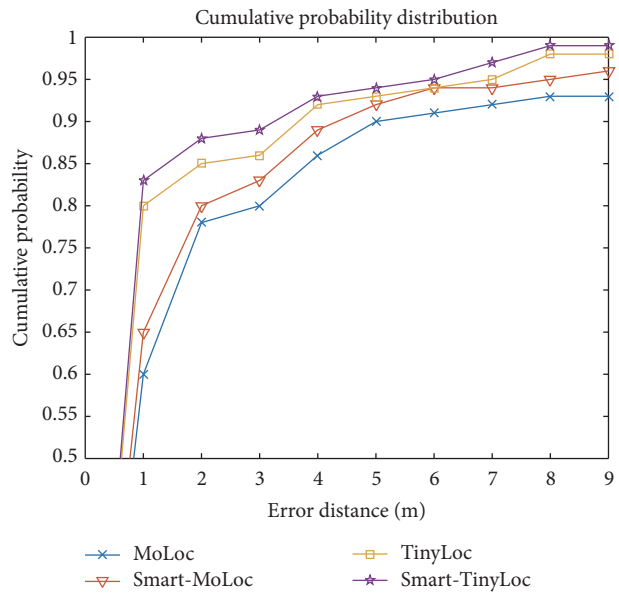


Figure 12: CDF of errors in area II.

We first made the cumulative distribution function of average localization errors both on Moto 360 and on HTC One. The calculation method for average error D_{loc} and the cumulative probability p_x are displayed by formula (3). The CDFs in errors of the two experiments are shown as Figures 11 and 12.

Evidently, Smart-MoLoc and Smart-TinyLoc outperform original indoor locating algorithm, and TinyLoc surpasses MoLoc when compared in the same environment. When using Smart-TinyLoc or Smart-MoLoc, the average probabilities of localization error within 2 m are over 80% in area II, and those probabilities decreased to 75% and 80%, respectively, in area I. Additionally, within 3.5 m, the probability of localization error for Smart-TinyLoc reached up to 90% in area II, while, for Smart-MoLoc, the error range should be released to 5 m to acquire an equal probability. Comparing the results in two different areas, those algorithms achieved better locating accuracy in area II, and the improvement of accuracy caused by SmartFix is little bigger in area II than that in area I. In area II, the improvement is about 2% and 5% for TinyLoc and 3% and 4% for MoLoc. For example, the probability of localization over 80% for Smart-TinyLoc is about 3.5 m and about 4 m for original TinyLoc. In area I, the average improvement caused by SmartFix decreased to 2% and 3% for MoLoc and TinyLoc. That is because area I is a large open space; there are much more changes for each person and less obvious character values in history path. According to this experiment, we find that SmartFix will work better in an indoor area such as office room, home, or restaurant. The experiment results indicate that SmartFix has a 2% and 5% improvement compared to original MoLoc with inner sensors and TinyLoc using certain principles to correct locating results, especially within a small interval.

We then calculated locating accuracy (point-matching probability) in two different areas with different number of APs, of which results are shown in Figures 13 and 14. In this

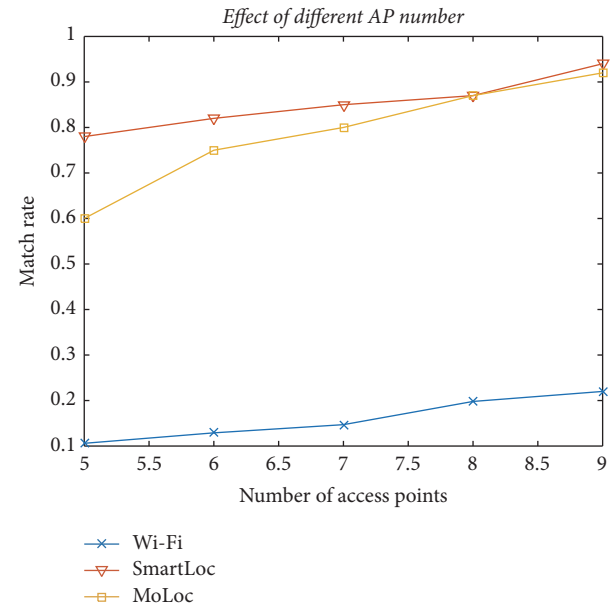


Figure 13: AP Number Change in area I.

experiment, we compared three locating algorithms: Smart-Los (Smart-TinyLoc), MoLoc, and basic Wi-Fi fingerprinting method. With the increase of APs amount, all three methods obtained better locating accuracy, and when the number of APs overpassed a fixed value, locating accuracy nearly remained the same, which is compliant with [17]. SmartFix and MoLoc outperform basic Wi-Fi fingerprinting method in every amount of APs. In particular, in area I, for the sake of being an open space, when the number of APs is set to 5, SmartFix achieved the highest accuracy, about 78%, followed by MoLoc, approximately 60%, and Wi-Fi fingerprinting method had only about 12% locating accuracy. When the number of APs rose to the maximum 9, SmartFix and MoLoc

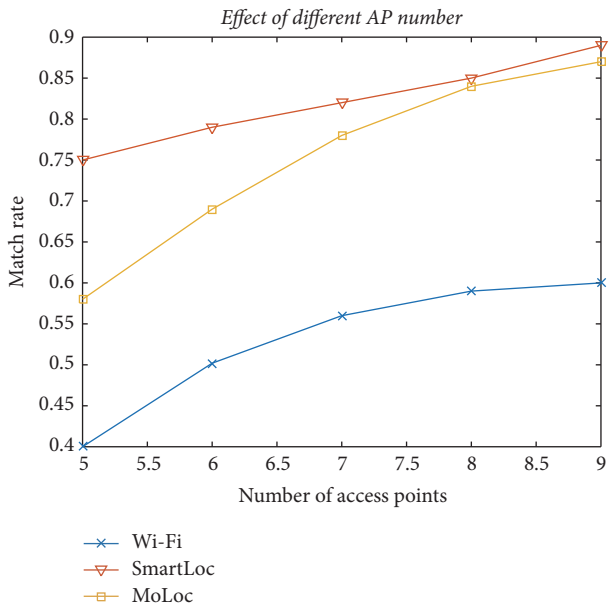


Figure 14: AP number change in area II.

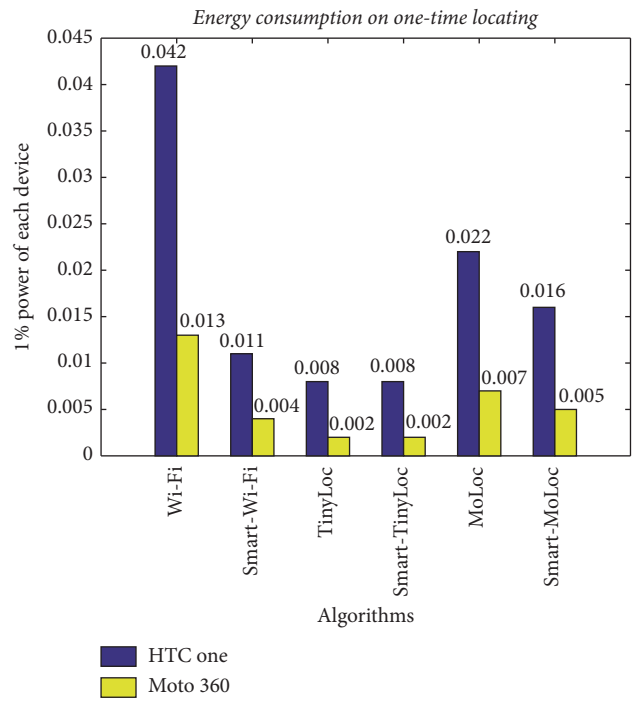


Figure 15: Energy consumption.

obtained 94% and 92% locating accuracy, respectively, and Wi-Fi fingerprinting method also had as approximately twice improvement as before. This results indicate that the number of APs has a significant effect on locating accuracy. The rising tendency of locating accuracy in area II is similar to that in area I. Because of area II's smaller space and lower density of reference positions, much more influences are exerted upon point-matching probability rather than locating errors. Therefore, in area II, despite the same AP number 5, except Wi-Fi fingerprinting method, the other two methods obtained lower locating accuracy than that obtained in area I, and their maximum locating accuracy is also lower than that in area I.

By studying the CDF of average errors and locating accuracy, we learned that SmartFix and MoLoc can significantly improve the locating accuracy of the Wi-Fi fingerprinting method. As for SmartFix which uses trace property, and MoLoc which uses motion property, they reached almost the same level of locating accuracy, while SmartFix is a little better in most cases.

4.2.2. Energy Efficiency. We implemented the localization algorithms both on the Moto 360 2nd-generation Smartwatch and on HTC One Smartphone to compute the average energy consumption on algorithm calculating and repeatedly scanned the Wi-Fi signal to compute the average energy consumption on the data collecting. The results showed that SmartFix can significantly reduce the energy consumption of MoLoc and Wi-Fi fingerprinting method. Endurance time of locating for Smartwatch using Wi-Fi fingerprinting method with SmartFix is about 1.96 times of that using original method.

For MoLoc, we repeatedly carried out the pedometer program to calculate the average energy consumption of built-in motion sensors. Figure 15 shows the total energy consumption of Smart-TinyLoc, Smart-MoLoc, Smart Wi-Fi fingerprinting method, and original methods for performing one-time locating at the same accuracy level. The power consumption for one-time locating is Wi-Fi method: 0.042%; SmartFix Wi-Fi method: 0.011%; TinyLoc: 0.008%; Smart-TinyLoc: 0.008%; MoLoc: 0.022%; and Smart-MoLoc: 0.016%, respectively. And on Moto 360, the power consumption of those methods is Wi-Fi method: 0.013%; SmartFix Wi-Fi method: 0.013%; TinyLoc: 0.002%; Smart-TinyLoc: 0.002%; MoLoc: 0.007%; and Smart-MoLoc: 0.005%. The total energy consumption of SmartFix and the Wi-Fi fingerprinting method is produced from data collection and calculation, while MoLoc has extra energy consumption by built-in motion sensors. Since the energy consumption of calculation is negligible compared with that of data collection, the built-in sensors are the main cause of the extra energy consumption of MoLoc, and SmartFix can reduce work frequency of built-in sensors to achieve lower energy consumption. And for Wi-Fi method, SmartFix can significantly reduce the amount of real-time RSS data and save almost 70% energy. In fact, the locating accuracy and energy efficiency are the two related aspects. It will produce extra energy consumption if the localization algorithm takes measure to improve locating accuracy in most occasions. Locating algorithms suitable for wearable devices must consider the balance between locating accuracy and energy efficiency. Through the estimation, implementing the localization algorithms to the Moto 360 and scanning RSS signal once every 3 seconds, the time that it can perform localization with different localization algorithms is shown in Figure 16. It indicates that the stand-by times to localization are remarkably different by using different localization algorithm and the stand-by times of

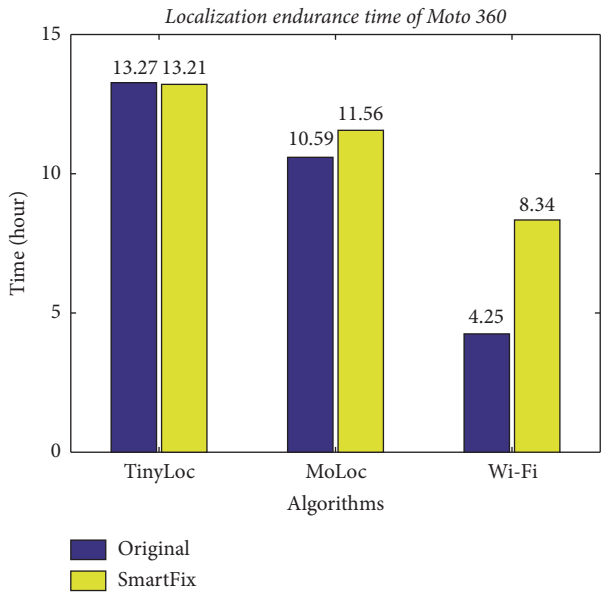


Figure 16: Stand-by time of Moto 360.

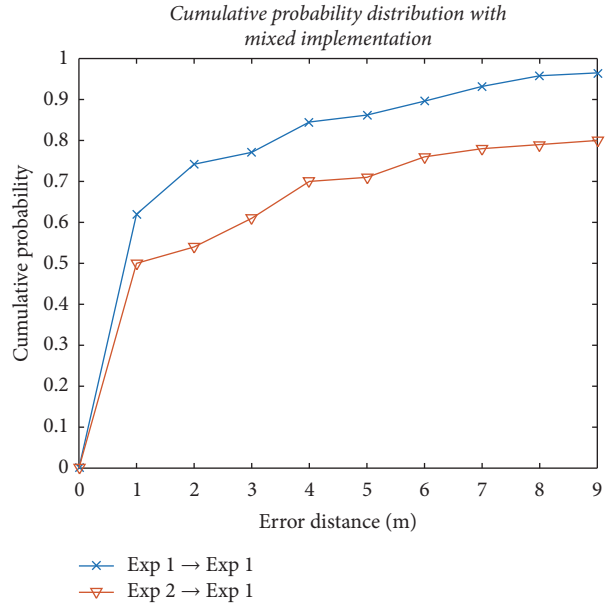


Figure 17: Training set exchange I.

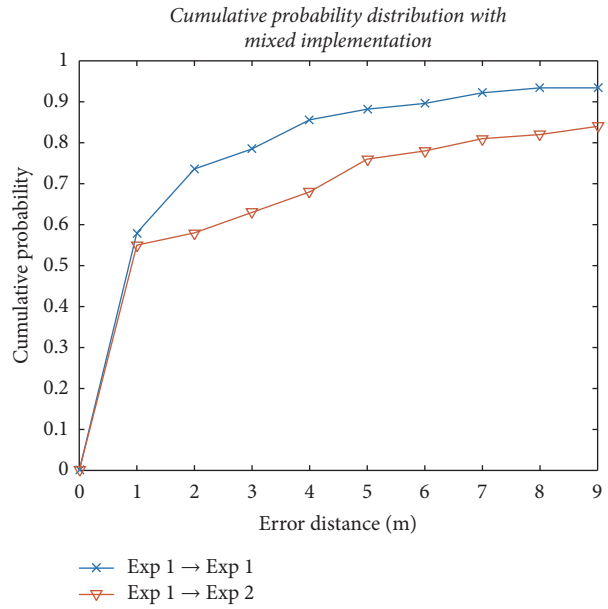


Figure 18: Training set exchange II.

SmartFix methods are almost longer than the original ones. The stand-by time of Moto 360 is almost 24 hours. Working time of locating using Smart-TinyLoc for Moto 360 is the longest and can achieve 13.21 hours which is 13% longer than that using Smart-MoLoc (11.56 hours) and 3.11 times of that using original Wi-Fi fingerprinting method (4.25 hours) when it collects RSS data four times per location and achieves the acceptable accuracy. Working time of locating using SmartFix Wi-Fi fingerprinting method also extends almost 2 times than the original method.

4.2.3. Compatibility. We conducted experiments to attest whether the motion features retrieved from a certain environment can be applied to other environments. To analyze the compatibility of such motion features, we exchanged the feature acquired in areas I and II, of which results are shown in Figures 17 and 18. By studying the experiment results, when applying area II features to area I, we got about 15% fall-off in accuracy. Meanwhile, the results of applying area I features to area II also show about 10% accuracy decrease. Therefore, the calculation of features should be conducted within specific environment. Such features have little compatibility among different areas.

4.2.4. Running Cost. As a history based indoor locating algorithm, the activating speed of SmartFix is shown in Figure 19.

The figure indicates that, with the enlargement of training data, SmartFix acquired better locating accuracy, and when the number of history data reached up to 1000, locating accuracy achieved the maximum. Nevertheless, different environment results in different amount of data required. In area II, SmartFix needs more data to achieve the same locating accuracy with that in area I, which may results in area II's small in space but complication in motion.

5. Related Work

Many indoor localization techniques have been proposed over the past decade and many researches study the application of indoor localization [18]. Wi-Fi-based indoor localization is always one of the most attractive techniques because of its ubiquitous deployment in indoor environment. Our work focuses on designing a low-power locating technology which can be deployed on the wearable devices and deals with the accuracy problems caused by reducing the amount of real-time data.

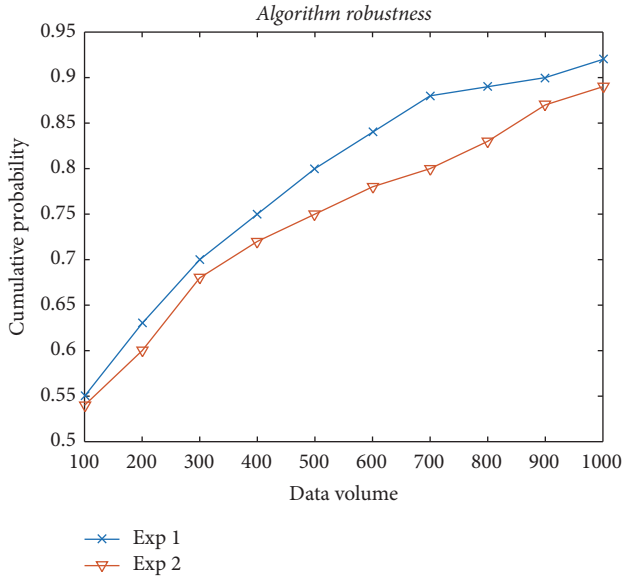


Figure 19: Running cost of SmartFix.

Early indoor locating technology using GSM signal, RFID, Infrared, ultrasound pulses, or UWB required specialized hardware to determine the devices location [1]. RF signal intensity was first used in RADAR for indoor localization in 2000 [2]. In recent years, more indoor localization technologies have paid increasing attention to Wi-Fi RSS data, such as accuracy [3, 5], reducing measurement area [7], training of system [8], or using high-quality audio sensing system [19] as assistant locating method. However, as we know, the signal strength always fluctuates, so it requires a certain amount of real-time data to ensure proper locating accuracy. How to compensate the influence of measurement errors is an important research subject. FreeLoc [6] abandons methods that use RSS value directly but uses AP's RSS relative order to locate a position and solve the problems caused by irregular changes of RSS. However, environment with less APs or adjacent regions RSS without significant change will cause some troubles in locating accuracy. MoLoc [14] is a motion-assisted locating scheme implemented on mobile phones which leverages user motion against fingerprint ambiguity. MoLoc can easily be integrated in existing locating systems by simply adding a motion database. However, the user motion detected by built-in sensors costs extra energy consumption which needs to be improved when implemented on the energy-constrained wearable devices. Another idea is to put aside the RSS and measure other stable physical data in Wi-Fi environment, such as FILA [9] proposed measure CSI (channel state information), in order to achieve higher accuracy. But it is not suitable for the wearable devices because of the low capability of detection.

Energy efficiency is always a popular research topic in indoor localization [20] and also a key issue for wearable devices. Perceiving environment information to dynamically adjust the rate of data collection is the main idea of the current energy-saving mechanism, such as using velocity of the nodes to dynamically adjust the frequency of acquiring the signal strength [10, 21–24], reducing the rate of use channel responses from multiple OFDM subcarriers [25], and using the environment information [26, 27]. There are other methods to reduce the total energy consumption in a system, like GreenLoc [28]. GreenLoc considers that people generally have similar mobility patterns, so it selects a few people from the group as samples instead of detecting every person in order to lower the average energy cost for the whole system. But GreenLoc is not good at locating each of the individuals, which exactly needs to be done in smart home scenes.

6. Conclusion and Future Work

The rapid development of smart home and wearable devices provides a good foundation for the high availability of indoor LBS in home setting scenario. However, energy efficiency is the essential issue that needs to be significantly improved for the existing locating technology before they can be implemented on energy-constrained wearable devices. In this paper, we propose a novel indoor localization technology called SmartFix with its focus on energy efficiency, the first one that can fit in wearable computing in smart home scenes.

SmartFix only needs single real-time RSS signal in the locating phase to guarantee excellent energy-saving performance. In another aspect, by referring to user motion features, SmartFix modifies locating results to achieve satisfying locating accuracy. According to the experiment, the probability of error within 2 meters can reach more than 80%. Meanwhile, energy consumption is 35% lower than that of MoLoc when achieves the same accuracy, and SmartFix obtains best accuracy with minimal energy cost. In the experiment, we applied the core idea of SmartFix to other location technology, achieved good results, and proved that SmartFix have good portability and technical compatibility.

In addition to using the motion features, we will focus more on optimizing strategies for SmartFix, so that SmartFix can be applied to more diverse indoor application scenarios.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Foundation of China (Grant no. 61472212), National Science and Technology Major Project of China (Grant no. 2015ZX03003004), the National High Technology Research and Development Program of China (863 Program) (Grant no. 2013AA013302 and Grant no. 2015AA015601), and EU Marie Curie Actions CROWN (Grant no. FP7-PEOPLE-2013-IRSES-610524).

References

[1] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Transactions on Systems, Man and Cybernetics C*, vol. 37, no. 6, pp. 1067–1080, 2007.

[2] P. Bahl and V. N. Padmanabhan, "RADAR: an in-building RF-based user location and tracking system," in *Proceedings of the 19th Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE INFOCOM '00)*, vol. 2, pp. 775–784, Tel Aviv, Israel, March 2000.

[3] M. Youssef and A. Agrawala, "The Horus WLAN location determination system," in *Proceedings of the 3rd International Conference on Mobile Systems, Applications, and Services (MobiSys '05)*, pp. 205–218, ACM, New York, NY, USA, June 2005.

[4] M. Youssef, M. Abdallah, and A. Agrawala, "Multivariate analysis for probabilistic WLAN location determination systems," in *Proceedings of the MobiQuitous 2005: Second Annual International Conference on Mobile and Ubiquitous Systems - Networking and Services*, pp. 353–362, San Diego, Calif, USA, July 2005.

[5] Widyawan, M. Klepal, and S. Beauregard, "A novel backtracking particle filter for pattern matching indoor localization," in *Proceedings of the 2008 International Conference on Mobile Computing and Networking, MobiCom'08 - 1st ACM International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments, MELT'08*, pp. 79–83, San Francisco, Calif, USA, September 2008.

[6] Y. Sungwon, P. Dessai, M. Verma, and M. Gerla, "FreeLoc: calibration-free crowdsourced indoor localization," in *Proceedings of the 32nd IEEE Conference on Computer Communications (INFOCOM '13)*, pp. 2481–2489, Turin, Italy, April 2013.

[7] A. Eleryan, M. Elsabagh, and M. Youssef, "Synthetic generation of radio maps for device-free passive localization," in *Proceedings of the 54th Annual IEEE Global Telecommunications Conference: "Energizing Global Communications", GLOBECOM 2011*, Kathmandu, Nepal, December 2011.

[8] H. Wang, S. Sen, A. Elgohary, M. Farid, M. Youssef, and R. R. Choudhury, "No need to war-drive: unsupervised indoor localization," in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services (MobiSys '12)*, pp. 197–210, ACM, New York, NY, USA, June 2012.

[9] K. Wu, J. Xiao, Y. Yi, M. Gao, and L. M. Ni, "FILA: fine-grained indoor localization," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM '12)*, pp. 2210–2218, Orlando, Fla, USA, March 2012.

[10] S. Tilak, V. Kolar, N. B. Abu-Ghazaleh, and K.-D. Kang, "Dynamic localization control for mobile sensor networks," in *Proceedings of the 24th IEEE International Performance, Computing, and Communications Conference (IPCCC '05)*, pp. 587–592, Phoenix, Ariz, USA, April 2005.

[11] Y. Qu, K. Xu, J. Liu, and W. Chen, "Toward a Practical Energy Conservation Mechanism with Assistance of Resourceful Mules," *IEEE Internet of Things Journal*, vol. 2, no. 2, pp. 145–158, 2015.

[12] K. Xu, Y. Qu, and K. Yang, "A tutorial on the internet of things: from a heterogeneous network integration perspective," *IEEE Network*, vol. 30, no. 2, pp. 102–108, 2016.

[13] K. Xu, X. Wang, W. Wei, H. Song, and B. Mao, "Toward software defined smart home," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 116–122, 2016.

[14] W. Sun, J. Liu, C. Wu, Z. Yang, X. Zhang, and Y. Liu, "MoLoc: on distinguishing fingerprint twins," in *Proceedings of the IEEE 33rd International Conference on Distributed Computing Systems (ICDCS '13)*, pp. 226–235, Philadelphia, Pa, USA, July 2013.

[15] W. Li, S. Wang, Y. Cui et al., "AP association for proportional fairness in multirate WLANs," *IEEE/ACM Transactions on Networking*, vol. 22, no. 1, pp. 191–202, 2014.

[16] W. Xiaoliang, X. Ke, Y. Zheng, and Z. Ge, "Tinyloc:Indoor localization for energy-constrained wearable devices[j/ol]," *Chinese Journal of Computers*, 2016 (Chinese), http://www.cnki.net/kcms/detail/11.1826.TP.20161106.1649.002.html.

[17] H. Li, L. Sun, H. Zhu, X. Lu, and X. Cheng, "Achieving privacy preservation in WiFi fingerprint-based localization," in *Proceedings of the 33rd IEEE Conference on Computer Communications, IEEE INFOCOM 2014*, pp. 2337–2345, Toronto, Canada, May 2014.

[18] C. Huijie, L. Fan, and W. Yu, "Echotrack: Acousticdevice-free hand tracking on smart phones," in *Proceedings of the in Proceedingsof IEEE 36th Conference on Computer Communications INFOCOM*, 2017.

[19] F. Li, H. Chen, X. Song, Q. Zhang, Y. Li, and Y. Wang, "CondioSense: high-quality context-aware service for audio sensing system via active sonar," *Personal and Ubiquitous Computing*, vol. 21, no. 1, pp. 17–29, 2017.

[20] L. Wang, Y. Cui, I. Stojmenovic, X. Ma, and J. Song, "Energy efficiency on location based applications in mobile cloud computing: A survey," *Computing*, vol. 96, no. 7, pp. 569–585, 2014.

[21] C.-W. You, Y.-C. Chen, J.-R. Chiang, P. Huang, H.-H. Chu, and S.-Y. Lau, "Sensor-enhanced mobility prediction for energy-efficient localization," in *Proceedings of the 2006 3rd Annual IEEE Communications Society on Sensor and Ad hoc Communications and Networks, Secon 2006*, pp. 565–574, Reston, Va, USA, September 2006.

[22] I. Shafer and M. L. Chang, "Movement detection for power-efficient smartphone WLAN localization," in *Proceedings of the 13th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, MSWiM 2010*, pp. 81–90, Bodrum, Turkey, October 2010.

[23] D. Mizell, "Using gravity to estimate accelerometer orientation," in *Proceedings of the Seventh IEEE International Symposium on Wearable Computers, 2003.*, pp. 252-253, White Plains, NY, USA.

[24] I. Constandache, R. R. Choudhury, and I. Rhee, "Towards mobile phone localization without war-driving," in *Proceedings of the IEEE 29th Conference on Computer Communications (INFOCOM '10)*, pp. 1–9, San Diego, Calif, USA, March 2010.

[25] S. Sen, B. Radunovic, R. R. Choudhury, and T. Minka, "You are facing the Mona Lisa: spot localization using PHY layer information," in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services (MobiSys '12)*, pp. 183–196, ACM, Lake District, UK, June 2012.

[26] Y. Liu, S. Lu, and Y. Liu, "COAL: Context Aware Localization for high energy efficiency in wireless networks," in *Proceedings of the 2011 IEEE Wireless Communications and Networking Conference, WCNC 2011*, pp. 2030–2035, Quintana Roo, Mexico, March 2011.

[27] M. Azizyan, I. Constandache, and R. R. Choudhury, "SurroundSense: mobile phone localization via ambience fingerprinting," in *Proceedings of the 15th Annual ACM International Conference on Mobile Computing and Networking (MobiCom '09)*, pp. 261–272, ACM, Beijing, China, September 2009.

[28] M. Abdellatif, A. Mtibaa, K. A. Harras, and M. Youssef, "GreenLoc: An energy efficient architecture for WiFi-based indoor localization on mobile phones," in *Proceedings of the 2013 IEEE International Conference on Communications, ICC 2013*, pp. 4425–4430, Budapest, Hungary, June 2013.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 8754020, 10 pages
https://doi.org/10.1155/2017/8754020

Research Article

QoE-Driven D2D Media Services Distribution Scheme in Cellular Networks

Mingkai Chen,[1,2] Lei Wang,[1,2] Jianxin Chen,[1] and Xin Wei[1]

[1]*Key Lab of Broadband Wireless Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, Ministry of Education, Nanjing, China*
[2]*The State Key Laboratory of Integrated Services Networks, Xidian University, Xian, China*

Correspondence should be addressed to Mingkai Chen; mingkaichen1989@163.com

Received 15 April 2017; Accepted 25 May 2017; Published 19 July 2017

Academic Editor: Feng Wang

Device-to-device (D2D) communication has been widely studied to improve network performance and considered as a potential technological component for the next generation communication. Considering the diverse users' demand, Quality of Experience (QoE) is recognized as a new degree of user's satisfaction for media service transmissions in the wireless communication. Furthermore, we aim at promoting user's Mean of Score (MOS) value to quantify and analyze user's QoE in the dynamic cellular networks. In this paper, we explore the heterogeneous media service distribution in D2D communications underlaying cellular networks to improve the total users' QoE. We propose a novel media service scheme based on different QoE models that jointly solve the massive media content dissemination issue for cellular networks. Moreover, we also investigate the so-called Media Service Adaptive Update Scheme (MSAUS) framework to maximize users' QoE satisfaction and we derive the popularity and priority function of different media service QoE expression. Then, we further design Media Service Resource Allocation (MSRA) algorithm to schedule limited cellular networks resource, which is based on the popularity function to optimize the total users' QoE satisfaction and avoid D2D interference. In addition, numerical simulation results indicate that the proposed scheme is more effective in cellular network content delivery, which makes it suitable for various media service propagation.

1. Introduction

With the propagation of the content delivery cooperative transmission technologies, D2D communication is considered as one of the most promising techniques for the next generation communication or new mobile applications, and it has recently received a substantial amount of interest [1–3]. Therefore, such D2D communication makes it possible to provide large heterogeneous media content services, such as HD (High Definition) video stream service, lossless music service, or website service for moving UE (User Equipment) [4]. Moreover, the heterogeneous media service over D2D communication is a very interesting topic. Meanwhile, it can make a difference on our daily life and provide a high Quality of Experience (QoE) service entertainment [5–7].

1.1. Related Work. At present, most of existing works on the content distribution for D2D communication underlaying cellular networks is focused on developing the transmission rate about the popularity files and D2D transmission radius. In [8], by means of exploiting the optimal collaboration distance, Golrezaei et al. have proposed a novel scheme to increase the throughput of video files, which in the cellular networks proves the possibility of improving spectral efficiency. The new architecture to achieve wireless D2D cache collaboration is proposed in [9] from the viewpoint of asymptotic scaling characteristics and video content popularity. In [10], considering the improvement of the area spectral efficiency of video transmission, Shanmugam et al. present a small cells heterogeneous architectures. In addition, [11] studies and formulates a max-flow optimization problem to maximizes the content downloading flows in D2D communication underlaying cellular networks and obtains the content downloading performance upper bound. Although this data dissemination protocols and schemes can be studied in D2D communication, which just focus

on the maximum capacity, optimal relationship between D2D radius, and service popularity for D2D networks, they ignore the characteristics of different media service, mobile opportunistic environment, and D2D interference. Moreover, they deprive the cochannel interference of frequency reuse in D2D communications and the QoE is out of consideration in a dynamic network scheduling problem [12]. Essentially, D2D communication technology belongs to point-to-point communication technology, and there is a certain similarity with P2P (Peer-to-Peer) technology in the dynamic network topology and data processing. But D2D communication pays more attention to the scope of users local service, and the P2P service is based on IP network location. At the same time, D2D communication is for wireless physical approaching users, and P2P users are near for a wired or wireless network virtual location. Therefore, in the D2D communication we need to consider problems in wireless transmission issue, while the P2P service does not consider.

Based on our previous work [13], the problem of heterogeneous media service over D2D cellular networks includes the following items: (1) content QoE satisfaction: how do we distribute the media content service to adapt dynamic complex networks and achieve optimal user satisfaction? (2) cache update: how to update each UE's cache in the context of mobile opportunistic meeting environment without exceeding UE capacity or bandwidth resource? These two issues interact with each other; thus, the challenging problem concerning on QoE-driven network service has arisen across the dynamic networks and multiple UE [12]. Consequently, these issues may also perplex transmission quality of media service in the D2D networks. Therefore, we attempt to layout a suitable D2D distributed content distribution scheme actually corresponding to different QoE models.

1.2. Contributions. In this paper, we consider a media service content delivery issue through opportunistic D2D communication underlaying cellular networks and establish a diversified media services spontaneous propagation framework among cellular users. Our objective is to maximize the total user' QoE MOS value and achieve a balance between the amount of media services and cellular resource. We first identify an objective function that incorporates the different media characteristics and QoE model. Different from the traditional content distribution issue in the D2D networks where only the popularity of media service is considered, we here take additionally the channel impact of the different media service into account, in addition to the transmission packet loss rate of D2D direct link. Although there are some mature content delivery algorithms known in Peer-to-Peer (P2P) networks, these algorithms are inadaptable to the media service in the mobile opportunistic environment and neglect the transmitting interference in the D2D network. The contributions of this paper are twofold as follows:

(1) We introduce the dynamic distributed heterogeneous Media Service Adaptive Update Scheme (MSAUS) for the D2D communication underlaying the cellular networks according to the popularity function and priority function. Then, we set up a D2D media service distribution scheme based on the popularity function and priority function to weigh the data dissemination characteristic and attempt to achieve the maximum satisfaction for cellular networks in the mobile opportunistic environment.

(2) We adopt resource greedy algorithm Media Service Resource Allocation (MSRA) to optimize the total users' satisfaction by reducing the interference from D2D media service update. This method reduces the impact of D2D interference caused by the band multiplexing on the prevalence media service.

1.3. Outline and Notation. The rest of this paper is organized as follows: Section 2 describes the media service system model for the cellular networks. In Section 3, we show an optimal distributed media service scheme based on the user cache and the popularity of media service. And the cellular resource allocation issue has been solved in Section 4. Then, extensive simulation results and comparisons are provided in Section 5. Section 6 provides conclusion and an outlook for future work. The following notations will be employed throughout the paper. Moreover, In Notations, we summarize the main notations used in this paper.

2. System Model and Problem Formulation

The scenario analyzed in this paper is depicted in Figure 1. There is a heterogeneous media service architecture for cellular and D2D communication networks, which contains three components, namely, D2D content delivery networks, network provider, and media content service provider [13]. In the D2D and cellular network component, a UE can access the BS (Base Station) via cellular link when the UE is under the base station coverage or chooses the vicinity UE to achieve the service through D2D link. In terms of the network provider part, the BS acts as gateway to connect the users to the media service providers. [8–10].

In our model, each user that is mobile and always connected to the BS is indicated as $u \in U \equiv \{u_1, \ldots, u_i\}$. Time is separated into time slots with a discrete index $t \in T$. During each time slot, we assume that the request media service is received with no errors and the media service updating can finish with no duration. Moreover, the BS controls the resource allocation in the D2D link and cellular link. According to the scared bandwidth resource $B \equiv \{B_1, \ldots, B_i\}$ the interferences should be one of the crucial factors to affect communication quality. Note that we assume the max number of interference source is Z. Then, the total bandwidth resource in the cellular network is expressed as

$$B = \sum_{i=0}^{U} ZB_i. \tag{1}$$

2.1. Media Service Request Generation Model. At the beginning of a time slot t, each user requests a random media service from a media service library $m \in M \equiv \{m_1, \ldots, m_i\}$. For each media service, we assume that it follows the Zipf distribution, which determines a ranking order of the media

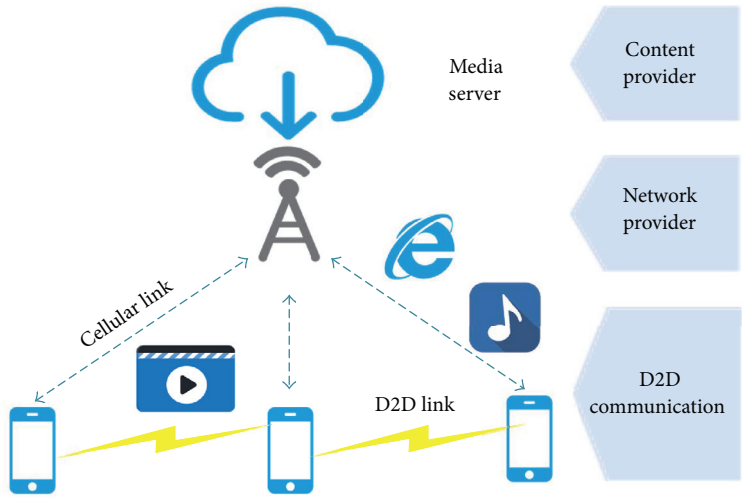


Figure 1: The architecture of media service for cellular networks.

Table 1: Different function expressions under different model of QoE functions.

Model	T1: Best Effort Service	T2: Video Service	T3: Audio Service
QoE function	$0.775\log(R) + 1.268$	$\frac{2.797 + 30 \times 0.0065 + 0.2498\ln(R)}{1 + 2.273\Theta + 7.1773(\Theta)^2}$	$\frac{2.2073\log(R)}{7.1773(\Theta)^2}$
Priority function φ	$\frac{\exp(-\lambda\pi r^2 X_i)}{Q(R(d))}$	$\frac{\exp(-\lambda\pi r^2 X_i)}{Q(R(d),\Theta(d))}$	$\frac{\exp(-\lambda\pi r^2 X_i)}{Q(R(d),\Theta(d))}$
Popularity function ϕ	$\frac{c\lvert U\rvert \exp(-\lambda \mathrm{SRC}_{m,i})}{\mathrm{SRC}_{m,i} Q(R(d))}$	$\frac{c\lvert U\rvert \exp(-\lambda \mathrm{SRC}_{m,i})}{\mathrm{SRC}_{m,i} Q(R(d),\Theta(d))}$	$\frac{c\lvert U\rvert \exp(-\lambda \mathrm{SRC}_{m,i})}{\mathrm{SRC}_{m,i} Q(R(d),\Theta(d))}$

service popularity. This assumption has been widely used [8] to describe content popularity distribution. According to the Zipf distribution, the popularity that a service m_i is requested by a user u_i is given by

$$m_i = \frac{1/i^{\gamma}}{\sum_{j=1}^{M}(1/j^{\gamma})}, \quad 1 \le i \le M,\ 1 \le j \le M, \tag{2}$$

where γ is a fixed parameter that describes the skewness of media service popularity, which defines the correlation level of user requests. High values of γ mean that most of the requests are generated from a few most popular files. For a user making a random request, m_i can be seen as the probability that the requested file is in the media service library M. If $\gamma = 0$, all media services have the same request probability, while, in the case of high values of value of γ, most of the service are requested with a low probability and there are only a few popular files.

2.2. QoE Description Model. In the cellular networks, we consider the user QoE gain generated by a media service request m from user u_i. Each user i demands media services in many various requests. Meanwhile, the different media service types are considered as an important metric to efficiently allocate resource to heterogeneous multimedia traffic. The QoE function reflects the relative satisfaction level of a user regarding the allocated resources. Due to the process of demands for different services and QoE model functions, the heterogeneous media service reflects the diversified QoE models. Since users may demand heterogeneous media services, the flexible models are required for us to account for their MOS. In this paper, we define QoE functions to characterize the users' experience for different types of media service delivery. We consider the following usage media service type for D2D applications: (T1) Best Effort Service (BES): non-real-time service, such as file download or data transmission; (T2) Video Model: HDTV signal transmission, video on demand; (T3) Audio Model: digital radio broadcasting, lossless music service [6, 7, 14]. We take advantage of the different QoE function in the previous work to quantify the user's satisfaction [15], which are shown in the Table 1.

According to previous studies [15], the media service over cellular networks is greatly impacted by the transmission rate R and PER (Packet Error Rate). Since the media service always prefers fulfilling user's demand as soon as possible, we assume that the D2D link rate R_D is higher than the cellular link rate R_c in the cell to achieve the higher user's satisfaction.

Then, the maximal achievable average data transmission rate for the downlink between the BS and UE u_i denoted by R_c, as follows:

$$R_c = B_i \log_2\left(1 + \frac{p_c d_i^{-\alpha} {h_c}^2}{N_0}\right), \tag{3}$$

where p_c means the transmission power from BS. Let B_i denote the set of all D2D communication pairs in the cell and

allocated bandwidth resource. Since D2D communication pairs share the same spectrum of the cellular uplink, we should consider the interference between all the different D2D pairs and the average transmission rate of the D2D pairs, denoted by R_D.

$$R_D = B_i \log_2 \left(1 + \frac{p_d d^{-\alpha}_{u_i,u_j} \left|h_d\right|^2}{\sum_{z \in Z} p_d d^{-\alpha}_{u_i,u_k} \left|h_d\right|^2 + N_0}\right), \quad (4)$$

where p_d is defined as the transmission power from UE. At this time, we assume that neither the transmitter nor the receiver discards packets maliciously; u_i has received the signed ACK packet to ensure that u_i successfully received the media service from the user u_j. A successful transmission probability from u_i and u_j is related to the transmission bit R_d or R_c, packet exponent Λ, and the max number of interference source Z [16]. Then, the probability of PER is given by

$$\Theta(z, R) = 1 - \left(1 - f\left(\frac{\Lambda}{R}\right)\right)^z \cdot f\left(\frac{\Lambda}{R}\right)^{(Z-1-z)}, \quad (5)$$

where

$$f\left(\frac{\Lambda}{R}\right) = e^{-\Lambda/R}\left(\frac{\Lambda}{R}\right). \quad (6)$$

2.3. Mobility Model. The user mobility pattern is modeled as a Point Poisson Process (PPP) model. A UE itself cannot know the probability of the UE it meets without the help of a BS control. Specifically, at the beginning of each time slot, each UE picks up a walking direction $\theta \in [0, 2)$ randomly and independently. The UE also chooses a constant velocity v to move during the rest of the time slot. They can be corresponded with other users during the mobility process. In this work, we assume the same mobility pattern for all users. Therefore, in each new time slot, the total users moving directions will refresh independently. Thus, the device random mobility is a Poisson distribution in the cellular network with density λ everywhere; therein λ is related to the number of users U in the cellular networks [15].

Therefore, under this three condition, users' QoE value expressions are not the same under this heterogeneous and mobile media service situation. To depict the real cellular network environment, we define $X_{m,i} = 1$ if UE u_i is in possession of service m, and otherwise. In addition, we assume that all UE has the same cache size c. The matrix $X_{i \in r} = \sum_r (X_{m,i})$ represents the state of the distributed cache under UE i's D2D radius r coverage.

Generally, we define $Q_{m,i}(\chi)$ to be the expected QoE gain generated by a request for service m from user i. Hence, we denote $Q(\chi)$ as MOS function for the media service, which represents the score of QoE. The problem of D2D resource allocation for QoE-driven media service distribution can be explicitly formulated as

$$\begin{aligned} \max \quad & Q(\chi) = \sum_{m \in M} \sum_{i \in U} P_d Q_{m,i}(\chi) X_{i \in r} + P_c Q_{m,i}(\chi) \\ \text{s.t.} \quad & \sum_{m \in M} X_{m,i} \le c, \\ & B = \sum_{i=0}^{U} Z B_i \\ & X_{m,i} \in \{0, 1\} \quad \forall i \in U \ \forall m \in M. \end{aligned} \quad (7)$$

Here, P_d and P_c are presented as the request distribution of D2D communication or cellular link, separately. Thus it can be seen that the total users QoE gain is affected by the user cache allocation, $X_{m,i}$, the media service popularity, P_c and P_d, and the bandwidth resource, Z and B_i. As a consequence, this media service problem is coupled with media content placement, user cache update and resource allocation. In the rest of the paper, we propose a distributed D2D media service scheme where each user and media service combine together to solve this issue through efficient cooperation.

3. Distributed Media Service Scheme in D2D Communication

In this section, we propose a mobile D2D data dissemination based on different QoE popularity function and priority function. In particular, we assume the UE may encounter with each other in an opportunistic way. Suppose that meetings among UE follow independent and memoryless processes. This helps us to find the optimal QoE scheme before evaluating them for these complex networks.

Due to the variable cellular network, the user's connection is based on the D2D communication radius. Meanwhile, the UE cannot clearly comprehend the status of transmit channel. The UE should attempt to achieve the most valuable media service against to the scarce resource and the limited cache storage. Therefore, when UE seeks the appropriate media service to update in the cellular network, we assume UE does not consider the interference in order to improve user satisfaction greatly. According to QoE analysis mentioned above, the media service's priority function is based on R and Θ, which is related to the Euclidean distance d. For PPP distribution in the cellular network with density λ [17], the probability that in a slot there are u devices in the D2D communication area is

$$P(u, r, \lambda) = \frac{\left(\lambda \pi r^2\right)^u}{u!} e^{-\lambda \pi r^2}. \quad (8)$$

Therefore, for a UE in the network the probability of at least another user caching to obtain the requested media service m_i within the D2D communication range and c size cache is

$$P_m^{\mathrm{D2D}} = 1 - P(u, r, \lambda) = 1 - e^{-\lambda \pi r^2 m_i c}. \quad (9)$$

(01) **Input:**
(02) User U, media service M, Initial all SRC value $\mathrm{SRC}_{m,i} = 0$
(03) Popularity function ϕ and Priority function φ
(04) **Output:**
(05) Optimal Media Service Adaptive Update Scheme
(06) **Procedure MSAUS**
(07) **if** (user i requests media service m)
(08) **if** (in the D2D transmission radius r user j)
(09) $\mathrm{SRC}_{m,i} = \mathrm{SRC}_{m,i} + 1$;
(10) **else**
(11) User j provide the media service to user i;
(12) set $\phi(\mathrm{SRC}_{m,i}) \geq 1$ and $\varphi(\mathrm{SRC}_{m,i})$
(13) m replaces the minimum priority media in i
(14) **if** ($\phi(\mathrm{SRC}_{m,i}) \geq 0$)
(15) i share m to the other users when they meets
(16) $\phi(\mathrm{SRC}_{m,i}) = \phi(\mathrm{SRC}_{m,i} - 1)$
(17) **endif**
(18) **endif**
(19) **endif**

Algorithm 1: MSAUS scheme.

Then, we have

$$E\left(Q\left(\overline{\chi}\right)\right) = \left(1 - e^{-\lambda\pi r^2 m_i c}\right) Q_d\left(R, \Theta\right) + e^{-\lambda\pi r^2 m_i c} Q_c\left(R, \Theta\right), \tag{10}$$

where $E[\cdot]$ represents the expectation of random variable systems. According to the definition of the expected gain in the meeting slot t, we can see that each user QoE is mainly composed by D2D and cellular media service. For formula (10), only m_i and $Q(R, \Theta)$ are variable for the certain networks. If we attempt to improve $E(Q(\overline{\chi}))$, we should adjust different media service distribution m_i in the network according to different service QoE function $Q(R, \Theta)$.

Therefore, the value of the total user QoE MOS value can be summed for each media service m; that is,

$$Q\left(\chi\right) = \sum_{m \in M} E\left(Q\left(\overline{\chi}\right)\right). \tag{11}$$

In the rest of this section, we attempt to design the priority and popularity function based on QoE function to design distributed media service scheme in Algorithm 1. This method might avoid the blindness of UE media service update and achieves a balance distributed service state in the network.

3.1. Expression of the Priority Function. For our proposed model, when user i requests the demand service m, the user can pick up the media service from the vicinity users or the eNB for the higher QoE MOS value. It follows that if we desire to improve the expectations of QoE satisfaction of each user, $X_{i \in r}$ should contain more abundant media service m. On the other side, for the different media service distribution in the user's cache, it may regulate status according to R and Λ. Thus, we define the heterogeneous media service priority function as

$$\begin{aligned} Q_m\left(\chi_m\right) \varphi\left(\chi_m\right) X_i &= Q_n\left(\chi_n\right) \varphi\left(\chi_n\right) X_j \\ &= \left(1 - e^{-\lambda\pi r^2 m_i c}\right) Q_d\left(R, \Theta\right). \end{aligned} \tag{12}$$

In the process of media service priority estimate, no D2D transmission behavior occurs and the service update is a spontaneous behavior, which does not affect the user's satisfaction in the other service. Therefore, in the process of predicting service priority, we assume that the D2D communication interference is not considered, which means interference index $z = 0$. However, the issue of service update caused by the decline in the quality of service within the same band will be discussed in the next section. Hence, according to (3), (4), and (5), $Q_d(R, \Theta)$ is represented as

$$\begin{aligned} R_D\left(d\right) &= B_i \log_2\left(1 + \frac{p_d d_{u_i,u_j}^{-\alpha} \left|h_d\right|^2}{N_0}\right), \\ \Theta\left(R_D\left(d\right)\right) &= 1 - \left(e^{-\Lambda / R_D(d)}\left(\frac{\Lambda}{R_D\left(d\right)}\right)\right)^{Z-1}, \\ Q_d\left(R, \Theta\right) &= Q_d\left(d\right). \end{aligned} \tag{13}$$

However, there are different QoE requirements for different media service in the Table 1, where d is the user's Euclidean distance for D2D pair under this scenario. The service in type T1, such as file downloading, is regardless of PER and MOS value decreases linearly with respect to d. T2 service, such as video streaming or conference D2D transmission is very sensitive to transmission distance d. Moreover, The marginal MOS value decrement of T3 service, such as lossless music service, becomes smaller as the distance increment. Examples of QoE functions for different type media service are shown in Figure 2.

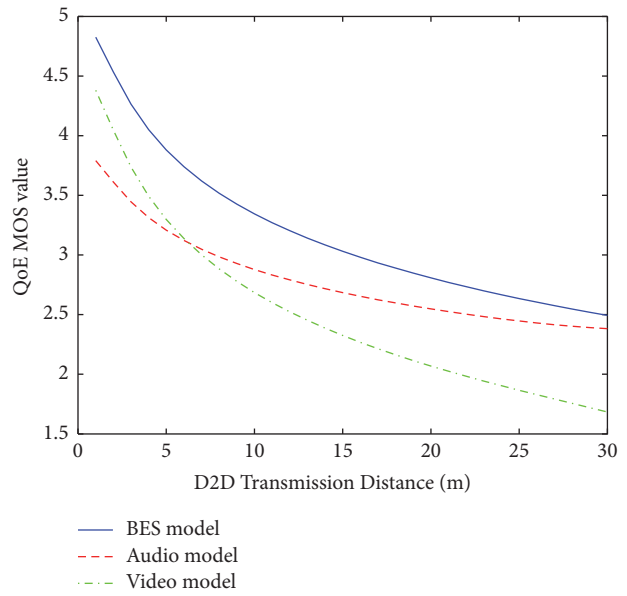


FIGURE 2: An example of QoE MOS value versus D2D distance for different type services.

Thereinto, the different types of media services obtain the same QoE MOS values $Q_c(R, \Theta)$ through the cellular link and can be considered as fixed constant. Meanwhile, it does not interact with the media service update in the cellular networks. Therefore, we mainly focus on user's satisfaction in D2D distributed scheme. Now, we obtain the probability distribution function of m's D2D radius; we use the decomposed function to model a dedicated user, that is,

$$Q(\chi) = \sum_{m \in M} E\left(Q_d\left(\overline{d}\right)\right). \tag{14}$$

Thus, we define the heterogeneous media service priority function as

$$Q_m(\chi_m)\varphi(\chi_m)X_i = Q_n(\chi_n)\varphi(\chi_n)X_j = E\left(Q\left(\widetilde{\chi}\right)\right). \tag{15}$$

This equation shows that $\partial Q/\partial X_i = E(Q_d(d))e^{-\lambda\pi r^2}\varphi(\chi_n)$ where φ is defined as

$$\varphi\left(\widetilde{\chi}\right) \propto \frac{e^{-\lambda\pi r^2 \overline{X}_i}}{Q_d(d)}. \tag{16}$$

Hence, we can set the priority function $\varphi(\widetilde{\chi})$ as the order for the u_i to rank the media service in his own cache and remove the media service with low priority; meanwhile, ensure the fairness for different media services.

3.2. Expression of the Popularity Function. We now describe the relationship between the priority function and the popularity function. Firstly, we give the expected value Service Receive Count (SRC) which is a measurement variable to $1/X_i$ for service m in UE u when a UE meets others. There is roughly a probability $c\lambda m_i$ that the media service m can be provided. Hence, we can set the UE's popularity function $\phi(|U|/X_m)$ as a first order of X_m. Meanwhile, for every user's cache, new media service always replace m with probability $X_i/c|U|$. In addition, the media service propagation is inversely proportional to the proportion and number of media services $|M|$ for all the services. In a stable steady state, the appearance of new service is equal to deleted or replaced old service. Hence, we have

$$\frac{c|U|}{X_m|M|}\phi\left(\frac{|U|}{X_m}\right) = \sum_U \frac{|U|}{X_i}\phi\left(\frac{|U|}{X_i}\right). \tag{17}$$

There are a lot of constants in the equation, and, therefore, we can write

$$\frac{1}{X_m}\phi\left(\frac{1}{X_m}\right) = \frac{1}{X_n}\phi\left(\frac{1}{X_n}\right) \quad \forall m, n \in M. \tag{18}$$

Alternatively, the scheme steady state should meet the equilibrium condition that we have

$$\varphi(x) = \frac{M}{cx}\phi\left(\frac{|U|}{x}\right) \quad \forall x > 0, \tag{19}$$

where φ is defined as shown above. It is easy to get the relationship between two functions. According to $\mathrm{SRC}_{m,i}$, the system achieves the maximum total user QoE value when the popularity function $\phi(m)$ and the priority function $\varphi(\widetilde{\chi})$ satisfy [13]

$$\phi(m) = \frac{c|U|}{|M|}\varphi\left(\frac{|U|}{m}\right). \tag{20}$$

On the basis of the relationship between the priority $\varphi(\widetilde{\chi})$ and popularity function $\phi(m)$, the popularity function can be expressed as

$$\phi\left(\mathrm{SRC}_{m,i}\right) \propto \frac{c|U|e^{-\lambda \mathrm{SRC}_{m,i}}}{\mathrm{SRC}_{m,i}Q(d)}. \tag{21}$$

Therein, $m_i|M|$ means the frequency of media service requested by users in the cellular networks. As the UE's service request is independent in the cell, we use $\mathrm{SRC}_{m,i}$ that can easily and explicitly calculate each user i's aspiration levels for media service m.

We indicate that the priority and popularity functions of each media service can be set adaptively to achieve the optimal and stable solution of the MSAUS scheme. Moreover, we summarize the expressions of the different priority and popularity functions for heterogeneous media service in the third and fourth rows of Table 1.

4. Resource Management in the Different Media Service Delivery

For different media services, the trade-off between the cellular resource and performance is also various. Therefore, the most prominent problem in D2D communication is the interference between transmission [4]. How do we achieve a trade-off between the interference control issue and updating D2D interference? Here, in this part, we will go deeply into this problem.

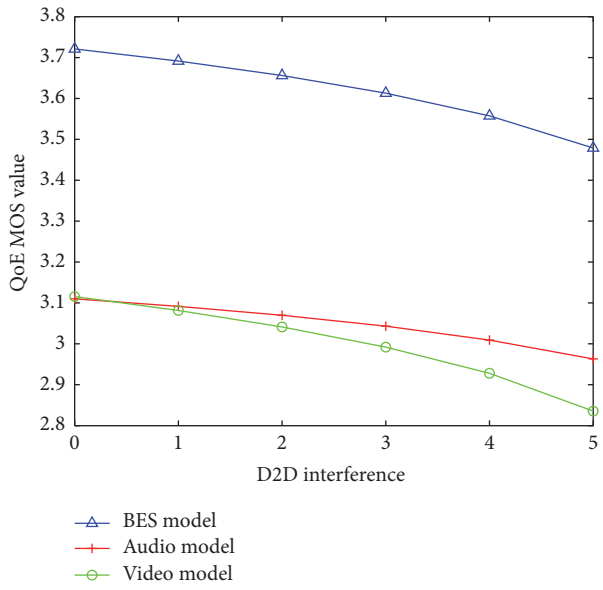


Figure 3: An example of QoE MOS value versus D2D interference z for different type services.

We give an example of QoE functions versus D2D interference for different type media service in Figure 3. In Figure 3, we can clearly perceive that the impact of D2D interference sources on the quality of multimedia services is very distinct. The increase of interference source has the greatest impact on the T3 service. Within the average increase of an interference source, the service quality will be decreased by 0.05 MOS value. Therefore, for the different effects of D2D interference on the service, the channel interference characteristics should be considered in the resources allocation B_i.

In order to ensure the transmission quality of the ongoing media service that the user requests, we propose a resource management scheme based on the media service popularity function. We call it Media Service Resource Allocation (MSRA) strategy. Note that we consider the trade-off between update cost and D2D interference to deploy limited bandwidth resource reasonably and efficiently. Each media service finds an appropriate resource allocation to render service or update it according to the popularity function ϕ.

The core point of the MSRA strategy is to find the optimal bandwidth reuse in the cellular network. Each media service and user collaborate to mark the different bandwidth and avoid the tremendous interference to reduce the ongoing quality of service. And we also define z_i to declare the B_i bandwidth allocated condition to handle the usage of the different channel state. Based on the randomness of the user initiated request m_i and uncertainty of heterogeneous services T_i, the general optimal algebraic optimization method is not suitable for this resource allocation problem in this situation. Therefore, in this paper we use the greedy algorithm to maximize the transmission performance $Q(\chi)$ of the cellular networks. The detailed operation of the MSRA is shown in Algorithm 2.

5. Numerical Results

In this section, we consider a cell network, where conventional UE is randomly distributed in the cell. Since the D2D users are usually within short distances, we adopt media service distribution model, where D2D users are uniformly distributed in a located circles and the simulation parameters are set according to [4]. The total users' QoE MOS value is used to evaluate the performance. Moreover, we compare our scheme with the LRU [5] (Least Recently Used) algorithm that is often used in the P2P networks and the typical network.

We first evaluate the performance of the MSAUS and MSRA scheme. Figure 4 compares the three schemes: the cellular network with MSAUS + MSRA, only with MSAUS, and LRU, when the subscriber number is 50, $z = 5$ and three kinds of media service ratio are 1 : 1 : 1. From Figure 4, both users' QoE gain increases with the time slot. Thus, the MSAUS and MSRA scheme can achieve almost 9 performance gains compared to the only with MSAUS scheme. Then, MSAUS + MSRA and only MSAUS scheme can achieve almost 21 and 10 performance gains comparing with LRU algorithm, respectively. Due to the rapid interaction of users, in the balance, a large number of users that uses LRU algorithm have stored many of the same high popularity media service and deleted the relatively low prevalence media services. The scarcity cache resource utilization is extremely uneven in the networks, leading to further enhancement of the quality of media service difficultly. It is shown that although the LRU algorithm is suitable for the P2P network, the algorithm is not appreciate for the dynamic network.

For the imbalance media service situation, our proposed scheme has achieved a better QoE gain in the cellular networks, as in Figure 5. In particular, the maximum MOS value using the proposed scheme is 231.6, whereas it is 217.3 only using MSAUS and 204.2 for the case of the LRU. This also shows that the proposed scheme, to the utmost extent, adapts to the environment which holds abundant higher QoE requested media services, like video or audio service. In view of the quality between different media services, the efficiency of the proposed algorithm is particularly prominent in the case of unbalanced media service distribution, and the problem of file popularity and scheduling of spectrum resources are of equal importance to improve the media service quality in the networks. In Figure 6, we repeat the results for different z. Note that z only has certain influence on the increasing speed of user QoE, but for the stable state z do not obtain more improvement to the total value. For the interference z caused by the media service update, different types, and different priorities have a far-reaching impact on the performance of the service. It is very important to allocate a limited frequency band, but the number of bands Z is only affected by the convergence speed of the service but has nothing to do with the convergence performance.

In Figure 7 we can clearly see that during the beginning of the system, the proposed scheme in this paper can quickly adjust the distribution of user services and improve the user satisfaction compared to the existing scheme. Moreover, in the case of different service distribution, when the popular

(01) **Input:**
(02) User U, Media service M, Popularity function ϕ
(03) Bandwidth resource $B = \sum_{i=0}^{U} z_i B_i$, $0 \le z_i \le Z$
(04) **Output:**
(05) Optimal Bandwidth Resource Allocation
(06) **Procedure MSRA**
(07) **while**
(08) **if** (user i request m from j)
(09) user i achieve B_i that contains min(z_i) interference;
(10) $z_i = z_i + 1$;
(11) **else**
(12) sort all $\phi(\mathrm{SRC}_{m,i})$ by decreasing order;
(13) **if** $(\sum_{i=0}^{n} z_i \le Zn)$
(14) $\phi_{\max}(\mathrm{SRC}_{m,i})$ allocate min(z_i) bandwidth B_i
(15) $z_i = z_i + 1$
(16) **else**
(17) B_i remain the same
(18) **endif**
(19) **endif**
(20) when user i has completed m transmission, set $z_i = z_i - 1$
(21) **endwhile**

Algorithm 2: MSRA greedy scheme.

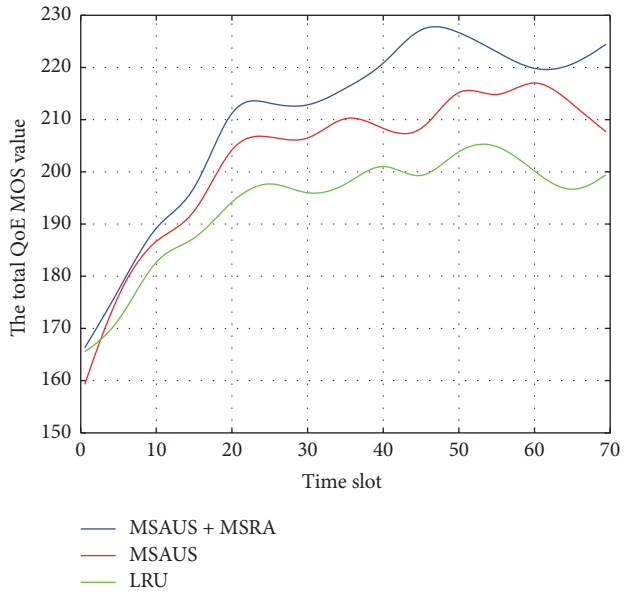


Figure 4: Performance comparison when BES : video : audio = 1 : 1 : 1.

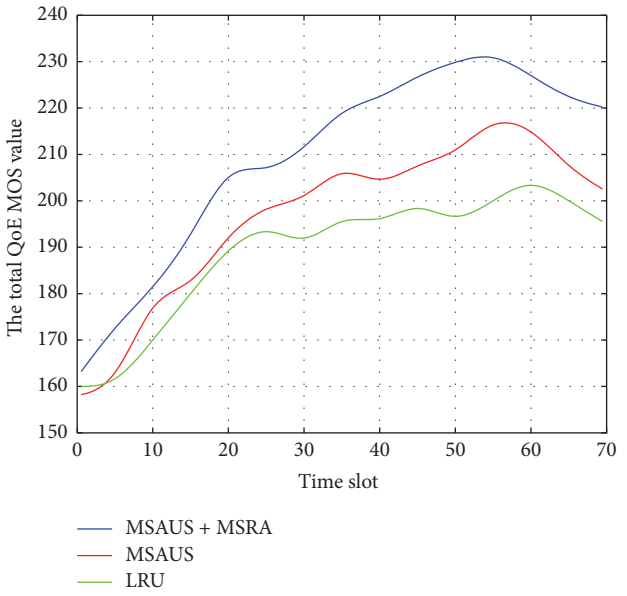


Figure 5: Performance comparison when BES : video : audio = 1 : 2 : 3.

service is concentrated on few services, the system efficiency is more obvious. When $\gamma = 0.8$, the average user's MOS value boosts 0.21 MOS in the proposed scheme comparing to LRU scheme, which is almost equal to twice as much when $\gamma = 0.4$. Figure 8 shows the media service D2D hit probability [18] for the case with two different schemes, respectively. As expected, the media service D2D hit probability increases in both schemes. Furthermore, we see the media service at the proposed scheme has much higher cache hit probability than caching at LRU scheme, as a result of service balance effect in priority function. The interaction between users makes the average MOS convergence faster; the rationality of media service storage leads to the higher D2D hit rate based on the algorithm we proposed in this paper.

6. Conclusion

In this paper, we proposed a distributed media service delivery and resource allocation scheme for D2D communication networks. Unlike the other conventional media service

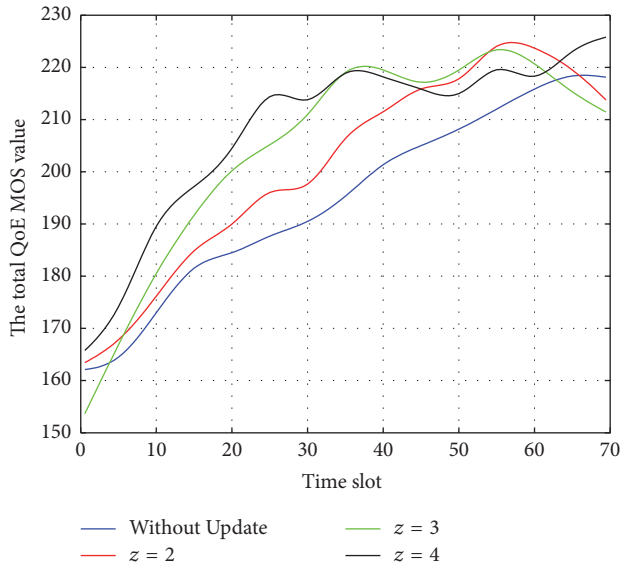


FIGURE 6: Performance comparison when z increases and BES : video : audio = 1 : 1 : 1.

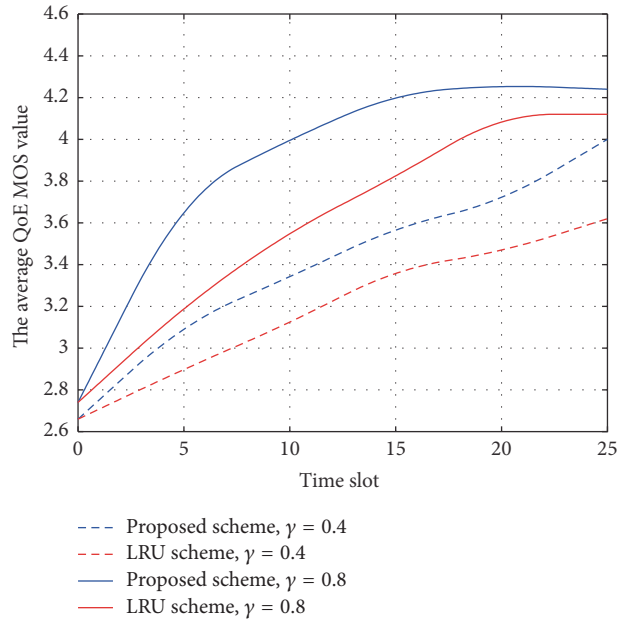


FIGURE 7: Performance comparison in the different media service popularity.

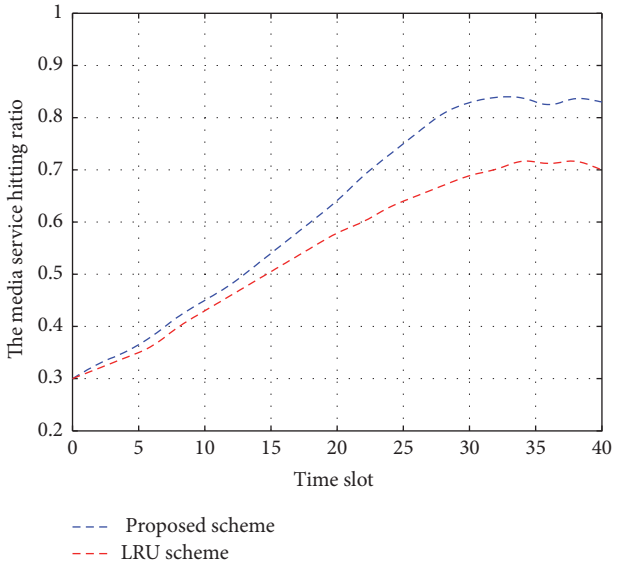


FIGURE 8: Probability of media service completion by D2D communication.

schemes or content delivery proposals, we do not focus on optimal quantity of service or throughput. Moreover, our work aims at achieving maximal total user's QoE value for dynamic cellular networks by jointly considering media service distribution and opportunistic transmission. Importantly, we combine the priority and popularity function, bandwidth allocation, and cache to achieve the goals of maximizing the total users' QoE value and solve the different content dissemination issues in D2D communication underlaying cellular networks. Our simulation results have shown that the proposed scheme achieves even somewhat QoE promotion against the other schemes.

For practical heterogeneous media service in D2D communication underlaying cellular networks, additional work needs to be excavated. For instance, we attempt to develop a media content retrieval scheme to improve the user's QoE in the heterogeneous media service system and to study an optimal media service discovery method with blind UE meeting information. In our ongoing work, we plan to address more media service transmission characteristics and study their impact on the actual D2D networks.

Notations

(u_i, U):	Mobile user, mobile user se
(t, T):	Time slot, time
(B_i, B):	Bandwidth resource
z, Z:	Interference index, the max number of interference source
m_i, M:	Media service, media service library
γ:	A fixed parameter describes the skewness of media service popularity
R_c, R_D:	Data transmission rate between D2D links or cellular links
p_c, p_d:	The transmission power from BS, the transmission power from UE
P_C, P_D:	The request distribution of D2D communication or cellular link
Λ:	Packet exponent
Θ:	Packet error rate (PER)
λ:	The density of Poisson distributed device random mobility
$X_{m,i} = (0, 1)$:	The indicator for possession of service m
c:	Cache size
$Q(\chi)$:	MOS function for the media service

ϕ: Popularity function
φ: Priority function
SRC: Service receive count
$d_{i,j}$: Users Euclidean distance for D2D pair.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is partly supported by the National Natural Science Foundation of China (61571240), the Priority Academic Program Development of Jiangsu Higher Education Institutions, the Natural Science Foundation of Jiangsu Province (BK20161517), the Qing Lan Project, the Open Research Fund of Key Lab of Broadband Wireless Communication and Sensor Network Technology (NUPT), Ministry of Education (NYKL201509), the Open Research Fund of The State Key Laboratory of Integrated Services Networks, Xidian University (ISN17-04), The Major Projects of the Natural Science Foundation of the Jiangsu Higher Education Institutions (16KJA510004), the Natural Science Foundation of Jiangsu Province (Grant no. BK20161517), and the Postdoctoral Science Foundation of China (Grant no. 2017M611881).

References

[1] A. Zhang, J. Chen, R. Hu, and Y. Qian, "SeDS: secure data sharing strategy for D2D communication in LTE-advanced networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2659–2672, 2016.

[2] D. Wu, J. Wang, R. Q. Hu, Y. Cai, and L. Zhou, "Energy-efficient resource sharing for mobile device-to-device multimedia communications," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 5, pp. 2093–2103, 2014.

[3] D. Wu, L. Zhou, Y. Cai, R. Hu, and Y. Qian, "The role of mobility for D2D communications in LTE-advanced networks: energy vs. bandwidth efficiency," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 66–71, 2014.

[4] D. Feng, L. Lu, Y.-W. Yi, G. Y. Li, G. Feng, and S. Li, "Device-to-device communications underlaying cellular networks," *IEEE Transactions on Communications*, vol. 61, no. 8, pp. 3541–3551, 2013.

[5] L. Zhou, R. Q. Hu, Y. Qian, and H.-H. Chen, "Energy-spectrum efficiency tradeoff for video streaming over mobile ad hoc networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 5, pp. 981–991, 2013.

[6] C. Xu, F. Zhao, J. Guan, H. Zhang, and G.-M. Muntean, "QoE-driven user-centric vod services in urban multihomed P2P-based vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 5, pp. 2273–2289, 2013.

[7] C. Xu, S. Jia, L. Zhong, H. Zhang, and G.-M. Muntean, "Ant-inspired mini-community-based solution for video-on-demand services in wireless mobile networks," *IEEE Transactions on Broadcasting*, vol. 60, no. 2, pp. 322–335, 2014.

[8] N. Golrezaei, P. Mansourifard, A. F. Molisch, and A. G. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 7, pp. 3665–3676, 2014.

[9] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4286–4298, 2014.

[10] N. Golrezaei, A. Molisch, A. G. Dimakis, and G. Caire, "Femto-caching and device-to-device collaboration: a new architecture for wireless video distribution," *IEEE Communications Magazine*, vol. 51, no. 4, pp. 142–149, 2013.

[11] Y. Li, Z. Wang, D. Jin, and S. Chen, "Optimal mobile content downloading in device-to-device communication underlaying cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 7, pp. 3596–3608, 2014.

[12] H. Zhu, Y. Cao, W. Wang, B. Liu, and T. Jiang, "QoE-aware resource allocation for adaptive device-to-device video streaming," *IEEE Network*, vol. 29, no. 6, pp. 6–12, 2015.

[13] L. Zhou, Y. Zhang, K. Song, W. Jing, and A. V. Vasilakos, "Distributed media services in P2P-based vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 2, pp. 692–703, 2011.

[14] L. Zhou, Z. Yang, Y. Wen, H. Wang, and M. Guizani, "Resource allocation with incomplete information for QoE-driven multimedia communications," *IEEE Transactions on Wireless Communications*, vol. 12, no. 8, pp. 3733–3745, 2013.

[15] L. Zhou, "Mobile device-to-device video distribution: theory and application," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 12, no. 3, article 38, 2016.

[16] J. Ning, S. Singh, K. Pelechrinis, B. Liu, S. V. Krishnamurthy, and R. Govindan, "Forensic analysis of packet losses in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 24, no. 4, pp. 1975–1988, 2016.

[17] J. Rao, H. Feng, C. Yang, Z. Chen, and B. Xia, "Optimal caching placement for D2D assisted wireless caching networks," in *Proceedings of the IEEE International Conference on Communications (ICC '16)*, pp. 1–6, Kuala Lumpur, Malaysia, May 2016.

[18] I. Pappalardo, G. Quer, B. D. Rao, and M. Zorzi, "Caching strategies in heterogeneous networks with D2D, small BS and macro BS communications," in *Proceedings of the IEEE International Conference on Communications (ICC '16)*, pp. 1–6, Kuala Lumpur, Malaysia, May 2016.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 3410350, 12 pages
https://doi.org/10.1155/2017/3410350

Research Article

Exploiting Delay-Aware Load Balance for Scalable 802.11 PSM in Crowd Event Environments

Yu Zhang,[1,2] Mingfei Wei,[1] Chen Cheng,[1] Xianjin Xia,[1] Tao Gu,[2] Zhigang Li,[1] and Shining Li[1]

[1]*School of Computer Science, Northwestern Polytechnical University, Xi'an, China*
[2]*School of Computer Science and IT, RMIT University, Melbourne, VIC, Australia*

Correspondence should be addressed to Yu Zhang; zhangyu@nwpu.edu.cn and Zhigang Li; lizhigang@nwpu.edu.cn

Received 19 March 2017; Accepted 1 June 2017; Published 12 July 2017

Academic Editor: Xiaoqiang Ma

This paper presents ScaPSM (i.e., Scalable Power-Saving Mode Scheduler), a design that enables scalable competing background traffic scheduling in crowd event 802.11 deployments with Power-Saving Mode (PSM) radio operation. ScaPSM prevents the packet delay proliferation of previous study, if applied in the crowd events scenario, by introducing a new strategy of *adequate competition* among multiple PSM clients to optimize overall energy saving without degrading packet delay performance. The key novelty behind ScaPSM is that it exploits delay-aware load balance to control judiciously the qualification and the number of competing PSM clients before every beacon frame's transmission, which helps to mitigate congestion at the peak period with increasing the number of PSM clients. With ScaPSM, the average packet delay is bounded and fairness among PSM clients is simultaneously achieved. ScaPSM is incrementally deployable due to only AP-side changes and does not require any modification to the 802.11 protocol or the clients. We theoretically analyze the performance of ScaPSM. Our experimental results show that the proposed design is practical, effective, and featuring with significantly improved scalability for crowd events.

1. Introduction

Energy saving for mobile devices in 802.11 networks has been a crucial issue over the last decade since Wi-Fi communication consumes a significant amount of energy. Although mobile applications have gained increasing popularity in recent years, the capacity of batteries on mobile devices grows at a much slower pace, and the limited battery life has become a bottleneck of enhancing user experience.

The IEEE 802.11 Standard [1] defines a Power-Saving Mode (PSM) for mobile devices to reduce energy consumption for Wi-Fi communication. However, PSM has become inefficient when multiple mobile clients coexist in a network. The competing background traffic among clients introduces significant delays as clients have to wait for others' transmissions, which could generate the extra energy consumption of the waiting clients.

Some recent efforts have been made to address competing background traffic scheduling in a single AP environment [2–5]. These methods optimize contention energy by isolating client traffic into different smaller time slices. However, to avoid large traffic delays, they only divide time slice within the time of one beacon interval, which produces limited number of time slices and thus leads to scalability issues for the strategies, especially when the Wi-Fi network operates in crowd event environments. A salient example is the annual Super Bowl football game in the United States, where approximately 75K attendees descend on a sports stadium for about half a day. During the 2013 Super Bowl game, 700 APs were deployed to provide a significant capacity for handling up to 30,000 simultaneous connections (i.e., with an average of 43 mobile clients accessed to each AP) [6–8].

Unlike a conventional single AP scenario, crowd event environments impose more challenges on traditional competing traffic scheduling, as summarized as follows.

(i) Large-Scale Competition. A large number of clients may simultaneously use a particular AP, and the number of communication channels is relatively limited at a typical AP cell.

(ii) More Fairness Requirement. People tend to use their Wi-Fi devices more than usual during crowd events to either share exciting live information with their friends or access the Internet, needing more fairness than they used to in other environments.

(iii) Widespread User Satisfaction. Poor performance will affect a large number of people and cause widespread user dissatisfaction.

We find that existing efforts for solving competing traffic scheduling focus much on how to save energy by eliminating contention among competing PSM clients, without considering the negative impact on packet delay due to energy conservation. However, in crowd event environments, although it is important to reduce energy cost, it is equally important to ensure users to have good packet delay performance and fairness simultaneously. Hence, we raise an intriguing question: *how to minimize energy consumption while meeting packet delay performance and ensuring fairness on the basis of scalability in crowd events?*

In this paper, we present a novel scheduler, named ScaPSM, as our first attempt to challenge the above problem. We take a fundamentally different approach *rather than reducing or even completely avoiding any contention among competing PSM clients in crowd events*; we seek to smooth the peaks that cause contention. The basic idea behind ScaPSM is to contribute a new strategy of *adequate competition* that exploits delay-aware load balance to control judiciously some competing PSM clients to contend for buffered packets and forces other PSM clients to delay their traffics to mitigate peak period congestion with packet delay deadline aware.

However, it is difficult to find the optimal adequate competition participants in order to meet both energy consumption minimization and performance requirements. We have the two following challenges to solve. First, background applications during crowd events are delay-sensitive (such as gathering group background management processes in screen-off traffic [9]). Hence, delaying client's traffic should not sacrifice the packet delay performance. The second one is that the packets which arrived at an AP usually belong to certain ongoing traffic sessions. Long traffic delays may lead to packet retransmission, which is undesirable. Hence, delaying downlink traffic must not exceed the maximum retry limit.

This paper makes the following contributions.

(i) We identify the scalability issue of competing PSM traffic in crowd event environments and formally model the delay-aware energy optimization problem in 802.11 networks, which is proved NP-hard.

(ii) We propose two algorithms (i.e., ACAA and FPSA) to determine the optimal number of competing PSM clients based on the specific properties of the problem and prove its stability.

(iii) We design a practical online scheduler, named ScaPSM, to minimize energy consumption while meeting both packet delay deadline and fairness among PSM clients in crowd event environments.

(iv) We conduct comprehensive evaluations, and the results demonstrate that, compared to NAPman and 802.11 Standard, ScaPSM achieves good energy saving with both good packet delay performance and fairness. Meanwhile the proposed algorithm achieves very close power saving to that of NAPman with reduction over 20x packet delay and ≤0.5 s traffic delays when the number of PSM clients reach 100. Our algorithm also achieves over 4x better delay fairness compared to 802.11 Standard.

The rest of the paper is organized as follows. In Section 2, we describe the system model and problem formulation. In Section 3, we present the design of ScaPSM. Performance analysis and extensive evaluation are reported in Section 4. Finally, we review the related work in Section 5 and conclude the paper in Section 6.

2. System Model and Problem Formulation

In this section, we first present the system model and then elaborate on how we handle packet delay and energy consumption while ensuring fairness. At last, we formulate the scheduling problem.

2.1. System Model. We consider a competing background traffic scheduling problem in an 802.11 deployment system during crowd events. Our system consists of one AP and m PSM clients ($m \in \mathcal{N}$). We denote $C = \{c_i \mid i \in \mathcal{N}\}$ as a set of PSM clients. For simplicity, we assume that downlink and uplink are separated, and we focus on downlink competing background traffic. In this paper, we first consider *homogeneous clients* that adopt a *static PSM* (SPSM) mechanism. Discussions on the *Adaptive PSM* (A-PSM) mechanism, another popular PSM implementation, will be left for our future studies. In addition, since scheduling the competing background traffic between PSM clients and CAM (i.e., Constant Awake Mode or high power awake mode) clients has been given a solution in [4], in this paper, we turn our attention on the competing background traffic among a large amount of PSM clients.

According to the 802.11 specification, at the beginning of each beacon interval (denoted by b_i, $i \in \mathcal{N}$), the AP notifies the PSM clients of the presence of buffered packets for them, through the Traffic Indication Map's (TIM's) field in the beacon frame. We assume that the packets of each client arrive continually over time.

2.2. Buffered Data Retrieval Model. Consider the data packets retrieval procedure between the AP and its associated clients in a beacon interval. We assume that every PSM client wakes up for beacon frame at the beginning of the beacon interval. For any client, if the corresponding TIM field in the beacon frame is set, it stays in wake mode and prepares to send a PS-Poll request frame by contending for the channel with other clients; otherwise it goes back to a low-power sleep mode to conserve power. If it wins the contention, the PSM client sends out a PS-Poll and the AP responds to it with a buffered data frame. The PSM clients remain in wake mode until the last packet is delivered, and then it goes back to sleep mode

immediately. We denote the beacon interval that a packet p arrives at the AP as $\mathscr{B}_a(p)$, and the beacon interval that p is scheduled to send to corresponding client as $\mathscr{B}_s(p)$.

We define the capacity of a beacon interval as the maximum amount of data that can be transmitted between clients and the AP in the beacon interval. Let $c(b_i)$ denote the capacity of beacon interval b_i and $v_p(b_i)$ the data transfer rate of packet p in beacon interval b_i. We have the following constraint.

$$\sum_{p\in\{p|\mathscr{B}_s(p)=b_i\}} v_p(b_i) \le c(b_i). \quad (1)$$

2.3. Packet Delay Impact. In order to reduce the energy consumption of client contentions in a beacon interval, a competing traffic scheduler can send traffic based on absolute isolation strategy on condition that there requires no change to the existing 802.11 protocol. Since the PSM clients which have not been selected to retrieve downlink data have to sleep and wait until the next beacon interval, they may suffer from long delay. This kind of packet delay may even violate certain performance bounds such as the *deadline* of a packet. To capture its performance impact, we introduce a performance cost metric $\phi_p(\cdot)$ from a packet point of view, which is exploited from [10].

For simplicity, we assume that any downlink packet can tolerate the same level of traffic delay. When the delay expectation for a buffered packet is violated, its performance may degrade significantly. This would cause bad user experience and thus a large performance cost. We take the term *deadline* as the bound of tolerable waiting delays of a packet.

Note that the packet delay is mainly caused by the MAC contention delay during the beacon interval (i.e., this time is referred to as *competing-beacon packet delay*) and the sleep delay of deferring competing for access (i.e., this time is referred to as *sleep-beacon packet delay*); we define the *performance degradation function* ϕ_p as

$$\phi_p(\text{delay}) = f_{\text{sleep}}(\text{delay}) \times f_{\text{comp}}(\text{delay}) \times \text{Len}(p), \quad (2)$$

where $\text{Len}(p)$ denotes the size of packet p. The function f_{sleep} represents the sensitivity of p to the *sleep-beacon packet delay*, and the function f_{comp} represents the sensitivity of p to *competing-beacon packet delay*. Denoting $\mathscr{B}_d(p)$ as the deadline of p, we can easily get the following property.

Property 1. Any $\phi_p(\cdot)$ should satisfy the following conditions:

(i) $\phi_p(0) = 0$.

(ii) If $d_1 < d_2$, then $\phi_p(d_1) \le \phi_p(d_2)$.

(iii) If $d_1 \le \mathscr{B}_d(p) - \mathscr{B}_a(p) < d_2$, then $\phi_p(d_1) < \phi_p(d_2)$.

The first two conditions ensure that ϕ_p captures the nondecreasing feature between the performance cost and packet delay. The third condition reflects the cost associated with the violation of deadline; that is, the user may have significantly worse experience and thus higher performance degradation cost.

Let $\mathbb{P}$ be a set of pending retrieval packets. Given $\phi_p(\cdot)$ for all the packets in $\mathbb{P}$, we can evaluate the total packet delay performance cost $\Phi(\mathbb{P}, S(\mathbb{P}))$ caused by a schedule $S(\mathbb{P})$ as $\sum_{p\in\mathbb{P}} \phi_p(\mathscr{B}_s(p) - \mathscr{B}_a(p))$. The schedule $S(\mathbb{P})$ is formulated by $S(\mathbb{P}) = \mathbb{P} \times \Gamma = \{\langle p, b_i\rangle\}$, $p \in \mathbb{P}$, $b_i \in \Gamma$, where tuple $\langle p, b_i\rangle$ signifies packet p is scheduled at the beacon interval b_i and $\Gamma = \{b_1, b_2, \ldots, b_m\}$ is a set of continuous beacon intervals during which the packets in $\mathbb{P}$ should be scheduled.

2.4. Competing PSM Clients' Fairness. Fairness is a key design objective for a competing traffic scheduler. As mentioned before, fairness should be considered in terms of both energy and delay. For energy fairness among competing PSM clients, it is ensured by the DCF scheme [11] if all clients have the same physical data rate, since the probability for each client to win channel contention is equivalent in the 802.11 Standard. Thus, in this paper, we focus on *delay fairness* whose meaning is that each client should receive fair opportunity to be scheduled by the AP before the corresponding deadline, irrespective of the number of clients.

Before formally providing a delay fairness metric, we first introduce three definitions about delay fairness among PSM clients as follows.

Definition 2 (delay of packet). Define the time of a buffered packet $d_p = b_s(p) - b_a(p)$ to denote its residential time or delay time at the AP.

Definition 3 (delay of client). The delay time D_i of a client c_i is the longest delay time among all buffered packets of c_i. Let p_{ij}^b and n_i^b represent client c_i's jth packet buffered in the AP and the number of buffered packets of c_i at the bth beacon interval, respectively; then $D_i = \max_{j=1}^{n_i^b} d_p(ij)$.

Definition 4 (delay fairness among competing clients). Given a time period Γ and the set of competing clients, one calls this kind of relation among these competing clients delay fairness if the delays of all competing clients are equal; that is, $D_i = D_j$, $\forall i, j \in \mathcal{N}$ and $i \ne j$.

It is worth noting that the definition of delay fairness above is strict. In the future, we can relax this strict definition and allow different clients having different delay tolerances based on specific application traffic and user's preference. For simplicity in this paper, we intensively consider the delay fairness model defined above.

To measure how well a competing traffic scheduler satisfies the delay fairness, we use the following *Relative Delay Fairness Bound* as a delay fairness metric based on [12].

Definition 5 (relative delay fairness bound). Let $C(\Gamma)$ be the set of clients that are delayed in a given time period Γ. Let ω_i be the weight of client c_i. We use RDFB to stand for Relative Delay Fairness Bound, which is defined as

$$\text{RDFB} = \sup_{i,j\in C(\Gamma)} \left| \frac{D_i(\Gamma)}{\omega_i} - \frac{D_j(\Gamma)}{\omega_j} \right|. \quad (3)$$

RDFB bounds the gap of delays experienced by any two clients in any given time period. Intuitively, the smaller the

gap is, the fairer the scheduler achieves. One of our objectives is to design a competing traffic scheduler with small RDFB.

2.5. Competing Traffic Energy Consumption. An 802.11 radio with PSM operation typically has three basic states: ACTIVE (i.e., TX/RX), IDLE, and SLEEP. We denote the corresponding radio power as P_A (or $P_{\text{tx}}/P_{\text{rx}}$), P_I, and P_S, respectively. As aforementioned, a PSM client wakes up for beacon frame at the beginning of a beacon interval. Each client decides to enter into one of the three states based on the TIM bit settings and results of channel contentions. Specifically, if a client's TIM field is set and it wins the contention, it will enter the high power ACTIVE state to download its data packets by means of PS-Poll → DATA → ACK frame sequences. While if a client's TIM field is set but it fails the contention, it will enter an IDLE state until it succeeds in accessing the wireless channel. Hence, the power consumption in IDLE state is much higher than that in SLEEP state and slightly lower than that in ACTIVE state. We assume that the energy consumption in SLEEP state is negligible since the radio is powered off. For simplicity, we neglect the energy consumption of changing client's radio states.

The competing traffic energy consumption is composed of the energy consumption in ACTIVE state and that in IDLE state. We first estimate the energy consumption in ACTIVE state. For any given size of data units, the data transmission energy depends on the product of two factors: the transmission power and the time taken to transmit all the data bits. Let $t_r(p)$ denote the average time allocated to a client that has one pending packet to retrieve. Based on [13], when there is no PS-Poll collision or transmission corruption, $t_r(p)$ can be expressed as

$$t_r(p) = \frac{CW_{\min}}{2} + T_{\text{pspoll}} + 2\text{SIFS} + T_{\text{data}} + T_{\text{ack}} + \text{DIFS}, \tag{4}$$

where $CW_{\min}$ presents the average back-off time for clients to send a PS-Poll frame, T_{pspoll} represents the time taken by a client to send a PS-Poll frame, T_{data} represents the time taken by the AP to send a DATA frame to the client, T_{ack} represents the time taken by a client to send an ACK frame to the AP, and both SIFS duration and DIFS duration are constants defined by 802.11 protocol.

Let $E_{\text{trans}}(p)$ denote the average packet transmission energy consumed during $t_r(p)$. Then it can be computed as

$$E_{\text{trans}}(p) = t_r(p) P_A. \tag{5}$$

In a given beacon interval b_i, we define three sets C_s^b, C_c^b, and C_f^b to denote the set of clients whose TIM field was set by the AP, the set of clients that successfully retrieved all buffered packets, the set of clients that only retrieved part, instead of all, of the buffered packets, respectively. We use N_c^b to stand for the total number of buffered packets that was successfully received by the clients during beacon interval b_i. Let r_i^b denote the number of client i's buffered packets that still remain at the AP; then we can compute N_c^b as

$$N_c^b = \sum_{i \in C_c^b} n_i^b + \sum_{i \in C_f^b} \left(n_i^b - r_i^b\right). \tag{6}$$

Let $E_{\text{trans}}(b_i)$ denote the total packet transmission energy consumed during a beacon interval b_i; then it can be expressed as

$$E_{\text{trans}}(b_i) = \sum_{p \in \mathbb{P}_c^b} E_{\text{trans}}(p). \tag{7}$$

Thus, during time period Γ, given $\mathbb{P}$ and a schedule $S(\mathbb{P})$, the total packet transmission energy can be estimated as

$$\widetilde{E}_{\text{tran}}(\mathbb{P}, S(\mathbb{P}), \Gamma) = \sum_{p \in \mathbb{P}} E_{\text{trans}}(b_i). \tag{8}$$

We now compute the energy consumption in ILDE state. Let us define three variables $M_s^b \in \mathcal{N}$, $M_c^b \in \mathcal{N}$, and $M_f^b \in \mathcal{N}$ to denote the number of clients in C_s^b, C_c^b, and C_f^b, respectively. According to the Buffered Data Retrieval Model, we can have $M_s^b = M_c^b + M_f^b$. As mentioned before, when only one PSM client wins the contention and is retrieving data frames from the AP, the transmission can be successfully performed. During this transmission time, the rest of clients whose TIM field was set should stay in IDLE state and consume idle power (i.e., P_I). Hence, let $E_{\text{idle}}(p)$ denote the idle energy consumed during an average packet transmission time $t_r(p)$; then it can be computed as

$$E_{\text{idle}}(p) = \left(M_s^b - 1\right) t_r(p) P_I. \tag{9}$$

Note that a background PSM client may quit contentions when it is indicated that no more frames are pending at the AP. It is difficult to determine the number of contending clients in the network on each round of contention. For simplicity, we assume that there are always M_s^b contenders in a given beacon interval b_i.

Let $E_{\text{idle}}(b_i)$ denote the total idle energy consumed by all the contending clients during b_i; then it can be expressed as

$$E_{\text{idle}}(b_i) = \sum_{p \in \mathbb{P}_c^b} E_{\text{idle}}(p). \tag{10}$$

Thus, under the schedule of $S(\mathbb{P})$, the total idle energy during the transmission of P can be estimated as

$$\widetilde{E}_{\text{idle}}(\mathbb{P}, S(\mathbb{P}), \Gamma) = \sum_{p \in \mathbb{P}} E_{\text{idle}}(b_i). \tag{11}$$

2.6. Problem Formulation. Our objective is to find a schedule $S(\mathbb{P})$ that can minimize the total energy consumption for transmitting buffered data in $\mathbb{P}$ without the packet delay performance degradation. It is constrained below an upper bound (denoted by $\widetilde{\Phi}$) during a given time period Γ. That is,

$$\sum_{p \in \mathbb{P}} \phi_p \left(\mathscr{B}_s(p) - \mathscr{B}_a(p)\right) \leq \widetilde{\Phi}. \tag{12}$$

A higher performance bound suggests a longer tolerable delay. Based on (8) and (11), during a given time period Γ, the total energy consumption of all PSM clients can be calculated as $E(\mathbb{P}, S(\mathbb{P}), \Gamma) = \widetilde{E}_{\text{tran}}(\mathbb{P}, S(\mathbb{P}), \Gamma) + \widetilde{E}_{\text{idle}}(\mathbb{P}, S(\mathbb{P}), \Gamma)$.

In order to formulate the optimization problem, we first need to introduce variables $x_{i,j} \in \{0, 1\}$. If $x_{i,j} = 1$, it represents that the i packet in $\mathbb{P}$ (denoted by p_i) is scheduled at the beacon interval b_j; otherwise it is not scheduled.

Then we model the problem as below.

$$\min \quad E(\mathbb{P}, S(\mathbb{P}), \Gamma) \tag{13}$$

$$\text{s.t.} \quad \sum_{j=1}^{m} x_{i,j} = 1, \quad i = 1, 2, \ldots, n \tag{14}$$

$$\sum_{i=1}^{n} x_{i,j} \cdot \text{Len}(p_i) \le c(b_j), \quad j = 1, 2, \ldots, m \tag{15}$$

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{m} j \cdot x_{i,j} - \mathscr{B}_a(p_i) \right) \le \widetilde{\Phi}, \tag{16}$$

where (14) stands for the fact that the ith packet should be scheduled at only one determined beacon interval b_i during the time period of Γ; capacity constraint (15) and delay cost constraint (16) are corresponding to (1) and (12), respectively.

It is worth noting that variables $x_{i,j}$ can only be set to integer one or zero. Hence, the problem modeled in (13)–(16) is an integer programming problem. It is *NP-hard* proved by [14]. Thus, we try to look for an approximate solution, instead of finding the optimal solution.

3. Scheduling Analysis and Algorithms

In this section, we first introduce our scheduling analysis to build a new mathematical model for making clear our target. After that, we design an online scheduler, named ScaPSM (i.e., Scalable Power-Saving Mode Scheduler), aiming to be implemented for real deployment. ScaPSM accounts for the packet delay performance and also ensures fairness among multiple PSM clients as the number of competing clients increases.

3.1. Delay-Aware Energy Scheduling Analysis. The proposed model in (13)–(16) demands that all traffic information in future time window Γ must be available. However, this assumption is limited in practical scenarios since future traffic information of clients cannot be perceived (only historical and present traffic information can be accessed by AP).

We present ScaPSM which does not require any future information of clients. It makes AP track the progress of each client's buffer and adaptively schedules traffic for them. Specifically, ScaPSM makes scheduling decisions in each beacon interval to obtain *adequate competition*, that is, optimizing both the energy and delay performance while ensuring fairness.

Generally, the delay sensitivity of a packet is reflected by its *deadline*. As aforementioned in Section 2.3, large *sleep-beacon packet delay* may even lead to violate the *deadline* of a packet. Therefore, in order to design a delay-aware scheduler, we require any packet to be delivered before a delay upper bound L; that is, $d_p = \mathscr{B}_s(p) - \mathscr{B}_a(p) \le L$.

At any given beacon interval b_i, in order to get *adequate competition* among multiple PSM clients to optimize the overall energy saving without degrading packet delay performance, we need to choose the clients with minimum total energy consumption for data delivering and leave others to stay in sleep state. Therefore, *the key problem we are facing is: given any beacon interval b, how to determine the set* C_s^b *defined in Section 2.5*?

Note that the PSM clients that have not been selected to be in wake mode should wait for the next beacon interval to transmit their PS-Poll requests. This implies that if C_s^b has a small number of clients, the average sleep-beacon packet delay significantly increases. On the other hand, if C_s^b has a large number of clients, these clients may consume significantly power owing to severe contention among themselves.

Therefore, we must judiciously control the qualification and number of competing PSM clients before every beacon frame's transmission. Specifically, in order to determine the set C_s^b, on the one hand, we need to avoid excessive clients to be scheduled at the same beacon interval; on the other hand, the packet delay of clients d_p should not be larger than L. This is essentially a load balance problem at the time range from b_i to b_{i+L-1}. We model the load balance problem as a *min–max of the number of participant problems*.

Before formal formulation, we first introduce the definition of client's remaining time.

Definition 6 (remaining time of client). The remaining time $DL(i)$ of a client c_i is the difference between the packet delay upper bound and the delay of client c_i. Given the packet delay upper bound (L) and the delay of client (D_i), then $DL(i) = L - D_i$.

To take into account the delay performance in our scheduling, according to the remaining time of a client, we first classify the PSM clients into L groups $\mathbb{G} = \{G_1, G_2, \ldots, G_L\}$, with each group $G_j = \{c_i \mid c_i \in \{DL(i) = j - 1\}\}$ $(1 \le j \le L)$. Let $M_j = |G_j|$ represent the number of clients in group G_j. In order to mitigate the congestion at the competing background traffic peak, some clients in G_i may be shifted to G_j $(i > j)$ at earlier beacon intervals for being scheduled. We use a variable M_{ij} to stand for the number of the shifted clients. The variable M_{ij} should meet the scheduling deadline constraint which is expressed by

$$M_{ij} = 0, \quad \text{if } (i \le j). \tag{17}$$

We define k_i to denote the number of clients being arranged to be scheduled at beacon interval of group G_i. k_i can be computed by

$$k_i = M_i + \sum_{j=1}^{L} M_{ji} - \sum_{l=1}^{L} M_{il}, \tag{18}$$

where $\sum_{j=1}^{L} M_{ji}$ stands for the number of clients which has been shifted into group G_i and $\sum_{l=1}^{L} M_{il}$ represents the number of clients which has been shifted out of group G_i.

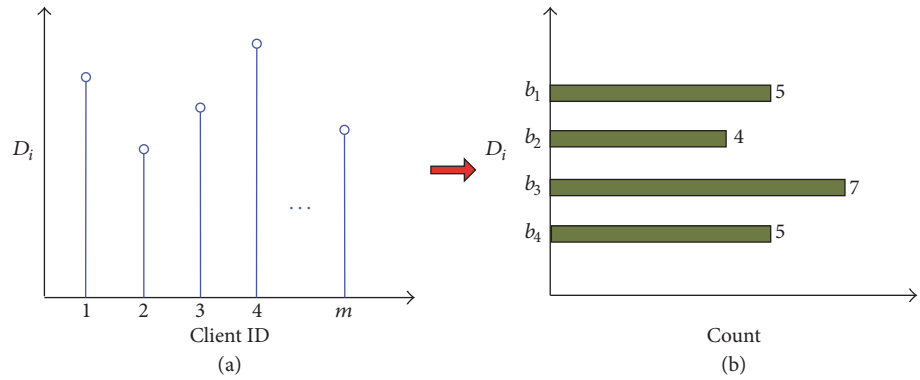


FIGURE 1: Organize clients into L groups.

Therefore, we model the min–max of the number of participant problems as follows.

$$\begin{aligned} \min \quad & \max_{i=1}^{L} k_i \\ \text{subject to} \quad & \text{Constraints (17) and (18).} \end{aligned} \tag{19}$$

Note that the scheduling of PSM client's traffic delivery is controlled by setting TIM bit in a beacon frame. We perform the competing background traffic scheduling for PSM clients at each beacon interval by two steps. In the first step, we determine the number of competing clients (denoted by k_i). In the second step, we select k_i right clients to be scheduled, with the goal of delay fairness.

3.2. Adequate Competition Assignment Algorithm. We start by computing k_i through a load balancing water-filling framework. The basic idea is described as follows.

Consider the L beacon intervals (denoted by $b_1, \ldots, b_L$) starting from the current beacon interval b_i. We use $\mathcal{U}_t$ to denote an *adequate competition set sequence*. Formally, $\mathcal{U}_t = \langle k_1, \ldots, k_L \rangle$, where k_i denotes the number of clients being arranged to be scheduled at b_i. As shown in Figure 1, we organize the clients into L groups, with each group $G_j = \{c_i \mid c_i \in \{DL(i) = j - 1\}\}$ $(1 \le j \le L)$. We initially arrange the clients in G_j to be scheduled at b_j (i.e., set $k_j = M_j$). We consider k_i as the water level of beacon interval b_i and attempt to shift clients from high water-level beacon intervals to low-level ones until water levels of the L beacon intervals can finally reach a *stable state*, just like water flowing. Specifically, restricted by traffic deadlines, we only allow water in b_i flowing forwardly to b_j's $(j < i)$. In this way, if too many clients are arranged to be scheduled at b_i, we will reschedule some of them to be in wake mode at earlier beacon interval for load balancing. Before proceeding, we formally define the *stable state* in L beacon interval as follows.

Definition 7 (stable state). Given an adequate competition scheduling arrangement $\mathcal{U}_t = \langle k_1, \ldots, k_L \rangle$, one says the L beacon intervals are in a stable state if $k_i \ge k_j$ is satisfied for any two beacon intervals b_i and b_j $(1 \le i < j \le L)$. In this case, one calls $\mathcal{U}_t$ a *stable adequate competition scheduling*.

Let $\mathbb{U}$ be the complete set of stable adequate competition scheduling arrangement. We have the following lemma and corollary.

Lemma 8. *Given $\mathcal{U}_t^* = \langle k_1^*, \ldots, k_L^* \rangle$, if $k_1^* = \max_{i=1}^{L}\{k_i^*\}$ and $k_1^* \le k_1$ is satisfied for any $\mathcal{U}_t \in \mathbb{U}$, $\mathcal{U}_t = \langle k_1, \ldots, k_L \rangle$, then $\mathcal{U}_t^*$ is the optimal adequate competition scheduling arrangement.*

Proof. We prove this lemma by using contradiction. We assume $\mathcal{U}_t^*$ is not optimal and denote the real optimal arrangement by $\mathcal{U}_t'$. $\mathcal{U}_t'$ does not satisfy $(k_1' = \max_{i=1}^{L}\{k_i'\} \wedge k_1' \le k_1)$ for any $\mathcal{U}_t \in \mathbb{U}$. There exist two cases.

Case A. If $k_1' \ne \max_{i=1}^{L}\{k_i'\}$, suppose $k_j' = \max_{i=1}^{L}\{k_i'\}$. We have $k_j' > k_{j-1}'$. In this case, by shifting $(k_j' - k_{j-1}')/2$ clients from b_j to b_{j-1}, we can produce another arrangement $\mathcal{U}_t''$ with $k_j'' = k_{j-1}'' = (k_j' + k_{j-1}')/2$. Since $k_j'' < k_j'$, $\mathcal{U}_t''$ is a better arrangement. This contradicts the assumption that $\mathcal{U}_t'$ is optimal.

Case B. If $(k_1' = \max_{i=1}^{L}\{k_i'\} \wedge k_1' > k_1)$ for some $\mathcal{U}_t \in \mathbb{U}$, since $k_1 = \max_{i=1}^{L}\{k_i\}$ under $\mathcal{U}_t$, $\mathcal{U}_t$ is a better arrangement than $\mathcal{U}_t'$. This, again, yields the contradiction.

This completes the proof. □

Corollary 9. *Given $\mathcal{U}_t^* = \langle k_1^*, \ldots, k_L^* \rangle$, if $\mathcal{U}_t^* \in \mathbb{U}$ and $k_1^* \le k_1$ is satisfied for any $\mathcal{U}_t \in \mathbb{U}$, then $\mathcal{U}_t^*$ is optimal.*

Corollary 9 suggests finding optimal arrangements within $\mathbb{U}$. In what follows, we will elaborate our algorithm to compute the optimal adequate competition scheduling arrangement.

The algorithm starts from the initial arrangement of $\mathcal{U}_t^1 = \langle M_1, \ldots, M_L \rangle$. It incrementally performs *stabilizing operations* from b_1 to b_L as follows: based on $\mathcal{U}_t^1$, we attempt to change water levels in b_1 and b_2 (i.e., k_1 and k_2) into a stable state, with minimal increase in k_1. And this produces $\mathcal{U}_t^2$. Next, based on $\mathcal{U}_t^2$, we further stabilize water levels among b_1, b_2, and b_3. Proceeding as above, it will finally cover b_L and produce a stable arrangement $(\mathcal{U}_t^L)$ with minimum k_1.

We describe one step of the operations in detail. Generally, we will produce $\mathcal{U}_t^i$ from $\mathcal{U}_t^{i-1}$. Note that beacon intervals

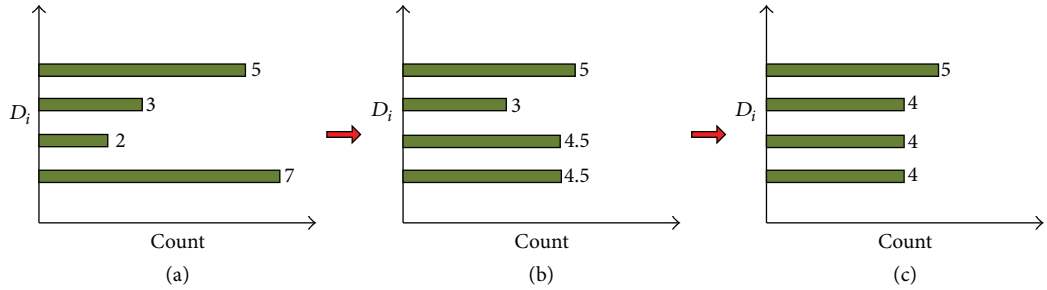


Figure 2: Illustration of stabilizing operations.

$b_1, \ldots, b_{i-1}$ are already in a stable state under $\mathcal{U}_t^{i-1}$. We want to push such stable state further to beacon interval b_i. In case that $k_i > k_{i-1}$ ($k_i = M_i$ under $\mathcal{U}_t^{i-1}$), we need to shift "water" from b_i to earlier beacon intervals for load balancing. This is achieved by the *stabilizing operations*, which require detailed discussions below.

We conduct stabilizing operations to determine the amount of water (M_{ij}) flowing from b_i to each b_j ($1 \le j < i$). We compare k_i with k_{i-1}. If $k_i \le k_{i-1}$, beacon intervals from b_1 to b_i are already in a stable state, and we do not need any operations. If $k_i > k_{i-1}$, we first try to balance water levels in b_{i-1} and b_i. We let $(k_i - k_{i-1})/2$ units of water flow from b_i to b_{i-1} (i.e., $M_{i,i-1} = (k_i - k_{i-1})/2$) and get $k_i' = k_{i-1}' = (k_i + k_{i-1})/2$. Next, we go back to check if the increase of k_{i-1} will break the stable state in $b_1, \ldots, b_{i-1}$. If $k_{i-1}' \le k_{i-2}$, the above operations successfully produce a stable state in beacon intervals from b_1 to b_i. Otherwise, we reattempt to balance water levels in the beacon intervals ranging from b_{i-2} to b_i. In this case, we will produce $k_{i-2}' = k_{i-1}' = k_i' = \overline{K}_3$, where $\overline{K}_3 = \sum_{j=i-2}^{i} k_j/3$. And hence, $M_{lb} = \overline{K}_3 - k_j$, $i-2 \le j < i$. We continue the process until the water levels in $b_1, \ldots, b_i$ become stable.

We illustrate the above operations by an example. As shown in Figure 2, $\mathcal{U}_t^3 = \langle 5, 3, 2, 7 \rangle$. We compute $\mathcal{U}_t^4$ from $\mathcal{U}_t^3$. Since $k_3 < k_4$, we balance water levels in b_3 and b_4 and get $k_3' = k_4' = 4.5$ (see Figure 2(b)). As $k_2 < k_3'$, we further balance water levels in b_2, b_3, and b_4, which produces $k_2' = k_3' = k_4' = 4$ (see Figure 2(c)). In this case, as $k_1 > k_2'$, the 4 beacon intervals reach a stable state. We obtain $\mathcal{U}_t^4 = \langle 5, 4, 4, 4 \rangle$. The amount of water flows from b_4 to b_3 and b_2 are $M_{4,3} = 2$ and $M_{4,2} = 1$, respectively.

Now, we give the algorithm in Algorithm 1. It is clear that it needs to be executed L^2 loops. We note that L is a constant given in advance. Hence, the computation complexity of the algorithm is $O(0)$. As the algorithm ensures k_1 to be increased minimally during each stabilizing operation, according to Corollary 9, the computed adequate competition scheduling arrangement is optimal.

3.3. Fair Participant Selection Algorithm. Based on the results produced by Algorithm 1, we move on to address the scheduling problem by selecting the right k competing clients to be scheduled. The main idea of Algorithm 2 is that scheduler can fairly select the right competing clients by computing dynamic priority weight ω_i for each client c_i. Note that, in Algorithm 1, k_i's and M_{ij}'s may not be integers. Hence, we derive k as $k = \lceil k_1 \rceil$. We select the k clients as follows: (i) all clients in G_1 must be scheduled; (ii) for clients in G_i ($i > 1$), we select the $\lfloor M_{i1} \rfloor$ clients with highest weight $\omega_i = n_i^b/DL(i)$ to ensure delay fairness scheduling defined in Section 2.4. Our strategy is described in Algorithm 2.

Input: M_i $(1 \le i \le L)$.
Output: $\mathcal{U}_t = \langle k_1, \ldots, k_L \rangle$, and M_{ij} $(1 \le i \le L,\ 1 \le j < i)$.
(1) $k_i \leftarrow M_i$, $(i = 1, \ldots, L)$.
(2) **for** $i = 2$ to L **do**
(3) $\quad k_i' \leftarrow k_i$.
(4) $\quad$ **for** $j = i - 1$ to 1 **do**
(5) $\quad\quad$ **if** $k_j \ge k_{j+1}'$ **then**
(6) $\quad\quad\quad$ let $k_l \leftarrow k_l'$, $(l = j+1, \ldots, i)$, and beak the loop.
(7) $\quad\quad$ **end if**
(8) $\quad\quad$ $k_l' \leftarrow \sum_{l=j}^{i} k_l/(i - j + 1)$, $(l = j, \ldots, i)$.
(9) $\quad\quad$ $M_{il} \leftarrow k_l' - k_l$, $(l = j, \ldots, i-1)$.
(10) $\quad$ **end for**
(11) **end for**

Algorithm 1: Adequate competition assignment algorithm (ACAA).

Input: k_1, M_{i1} $(1 < i \le L)$.
Output: $\mathcal{W}_t$ the set of clients scheduled at beacon interval b.
(1) $k \leftarrow \lceil k_1 \rceil$.
(2) count $\leftarrow |G_1| + \sum_{i=2}^{L} \lfloor M_{i1} \rfloor$.
(3) Compute $\omega_i = n_i^b/DL(i)$ for each $c_i \in \mathcal{G}$.
(4) Add all clients in G_1 to $\mathcal{W}_b$.
(5) **for** $i = 2$ to L **do**
(6) $\quad$ Add the $\lfloor M_{i1} \rfloor$ clients with highest ω_i in G to $\mathcal{W}_b$.
(7) **end for**
(8) Select (k – count) clients with highest ω_i from the all clients in $\mathcal{G}$, and add them to $\mathcal{W}_b$.

Algorithm 2: Fair participant selection algorithm (FPSA).

4. Performance Evaluations

In this section, we evaluate the performance of ScaPSM through stability analysis and simulations.

4.1. Stability Analysis. To demonstrate the stability of ScaPSM, we need to show two proofs. First, we should show that there exists one (equilibrium) state at which, once hit, the system will stay forever. Second, we should show that the system will move to the equilibrium state eventually regardless of its initial or current state.

Theorem 10. *The equilibrium state is delay-aware overall energy-optimal, and it achieves min–max of the number of participants at any beacon interval.*

Clearly, in the equilibrium state, the scheduler in the AP completes an adequate competition scheduling arrangement procedure by using Algorithm 1. Then, the statement is true according to Lemma 8.

4.2. Methodology and Simulation Setup. We compare ScaPSM with both 802.11 Standard [1] and NAPman [4]. Using an absolute isolation strategy, NAPman can create no contention wireless channel access for PSM clients. Thus, it may be regarded as an optimal energy saving scheduler for competing background traffic.

Three metrics are used for performance evaluation, that is, *energy consumption*, *packet delay,* and *delay fairness*. Through our analysis, ScaPSM can achieve all three desirable properties. Specifically, when PSM clients increase, ScaPSM should optimize overall energy consumption without degrading packet delay while ensuring fairness among PSM clients.

First, we start with controlled PSM client traffic in order to highlight various aspects of ScaPSM. This controlled traffic is intended to represent the worst case for various scheduling strategies since there is a packet in every beacon interval for a PSM client. Second, to compare ScaPSM with NAPman and 802.11 Standard in crowd events, we configure the server to send packets with a random interval ranging from 10 ms to 300 ms (based upon the SIGCOMM'08 traces [15], primarily Web traffic). In a typical crowd event scenario, according to the report in [6], each AP is associated with 107 clients on average, where up to 43 clients may simultaneously access Wi-Fi. To simulate such highly-competitive environment, we employ one AP and up to 100 clients in our OMNet++ simulation. We configure parameters with $T_b = 100$ ms, $L = 500$ ms, where T_b is the duration of one beacon interval. Each data point is an average over 20 independent runs.

4.3. Impact of Number of PSM Clients on Power Consumption. We first investigate the impact of the number of PSM clients on power consumption of one single client and all clients. The results of controlled traffic and trace-driven traffic are shown in Figures 3(a)-3(b) and Figures 3(c)-3(d), respectively.

In Figures 3(a)-3(b), we can see that as the number of PSM clients increases, both ScaPSM and 802.11 Standard consume higher power than NAPman (i.e., an approach with no contention) in the cases of both one single client and all clients. However, the power consumption of ScaPSM increases much slower than that of 802.11 Standard. This is easy to understand because the total number of competing clients of 802.11 Standard is equivalent to the total number of PSM clients if all clients have buffered packets in the AP. Since the adequate competition strategy of ScaPSM can control the number of competing clients at any beacon interval, the power consumption caused by contention reduces correspondingly. Moreover, the power consumption of ScaPSM is very close to that of NAPman. In Figures 3(c)-3(d), we see that, under the trace-driven traffic, with the number of PSM clients increasing, the power drawn by one single client and all clients with ScaPSM is still much lower than that of 802.11 Standard. The results are similar to that of the controlled traffic configuration.

4.4. Impact of Number of PSM Clients on Packet Delay. We design experiments to study the impact of the number of PSM clients on packet delay performance of our algorithms. The results of controlled traffic and trace-driven traffic are shown in Figures 4(a)-4(b) and Figures 4(c)-4(d), respectively.

Figure 4(a) shows that the packet delay of ScaPSM is almost unchanged as the number of PSM clients increases. The adequate competition assignment algorithm of ScaPSM achieves good performance. In contrast, NAPman performs the worst. The average packet delay of no contention becomes very high as the number of PSM clients increases. When the number of PSM clients is 40, the average packet delay of no contention is around seven times as large as that of ScaPSM. Due to the large difference between no contention and our algorithm, the logarithmic function is used to have a fine-grained view, as shown in Figure 4(b). The result shows that the absolute isolation strategy of NAPman optimizes energy consumption at the cost of large packet delays. This is because in order to eliminate power consumption caused by contentions, NAPman will generate a large amount of *sleep-beacon packet delay*. Since ScaPSM is delay-aware, the average packet delay is bounded by deadline L. As shown in Figure 4(b), the packet delay of 802.11 Standard is the lowest when the number of PSM clients increases. This is because ScaPSM has to delay some clients' traffic by the right duration to mitigate traffic congestion. As a result, ScaPSM produces small *sleep-beacon packet delays*. Figures 4(c)-4(d) show that, under trace-driven traffic, ScaPSM and NAPman exhibit similar performance. This shows that ScaPSM achieves good scalability.

4.5. Impact of Number of PSM Clients on Delay Fairness. Finally, we confirm experimentally that ScaPSM provides good delay fairness, irrespective of the number of PSM clients. The results of controlled traffic and trace-driven traffic are shown in Figures 5(a) and 5(b), respectively. As the absolute isolation strategy of NAPman can be considered as a round-robin scheme, we exclude NAPman from the delay fairness comparison.

From Figure 5(a), we can see that ScaPSM has a better delay fairness performance than 802.11 Standard, no matter under controlled traffic or trace-driven traffic. The RDFB value of ScaPSM remains almost a small constant. In contrast, the RDFB value of 802.11 Standard increases 4x larger than that of ScaPSM when the maximum number of clients changes from 50 to 100. This is because ScaPSM is delay-aware, and the difference between the weights ω_i and ω_j of any two clients is constrained. In contrast, 802.11 Standard

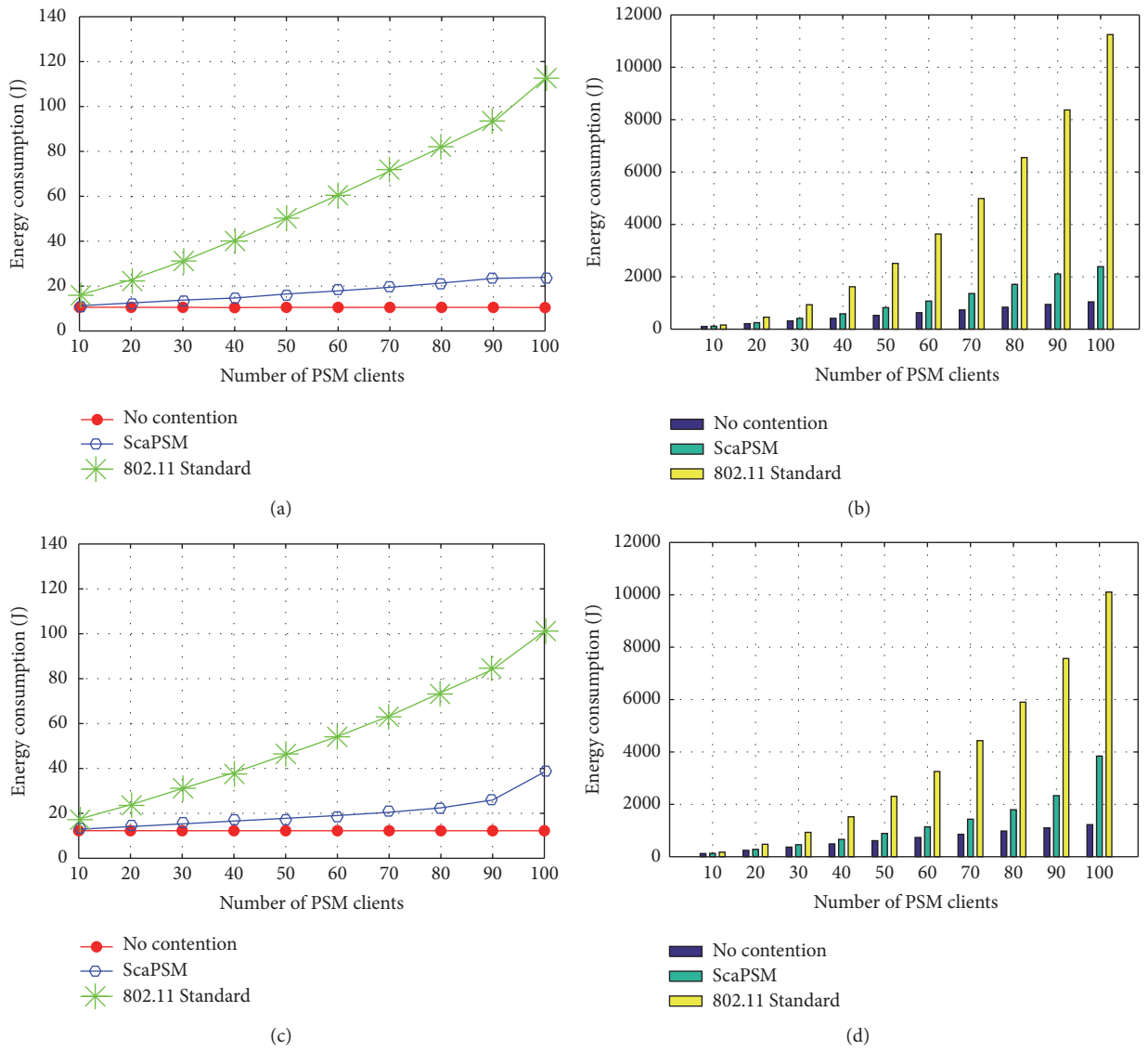


FIGURE 3: (a) Power drawn by a static PSM versus number of PSM clients under controlled traffic; (b) overall power drawn by all static PSMs versus number of PSM clients under controlled traffic; (c) power drawn by a static PSM versus number of PSM clients under trace-driven traffic; (d) overall power drawn by all static PSMs versus number of PSM clients under trace-driven traffic.

has no concept of delay deadline; therefore, its difference between any two clients' weights is large. Similar results are also observed under trace-driven traffic, as shown in Figure 5(b).

5. Related Work

5.1. Crowd Event Scenario. Network communication in crowd events has recently attracted much research attentions. Shafiq et al. [16] and Erman and Ramakrishnan [7] take the first step to study traffic characteristics in the crowd event scenario. Although they do not propose strategies for performance improvement, their work provides crucial insights into the design of ScaPSM. There are a few techniques, such as WiFox [8] and AMuSe [17], being proposed to improve system throughput in dense AP/client environments. However, they do not address the energy issues of 802.11 network. Our work essentially fills the gap.

5.2. Contention Avoidance Scheduling. This issue has been extensively studied in [2–5] to save energy. Time slicing is used in [2] to make each PSM client's packets delivered only in its appointed time slice to save power and reduce the effect of background traffic. In LAWS [3], the AP advertises a subset of PSM clients in the beacon and clients use information in the beacons to determine their polling sequence to avoid client contention. However, these solutions require modifications to the 802.11 Standard, and thus, changes to both mobile clients and APs are unavoidable. SOFA [5] maximizes the total sleep time of all clients. However, SOFA assumes that AP has full control of its downlink traffic which is often limited in the 802.11 Standard since the AP shares the

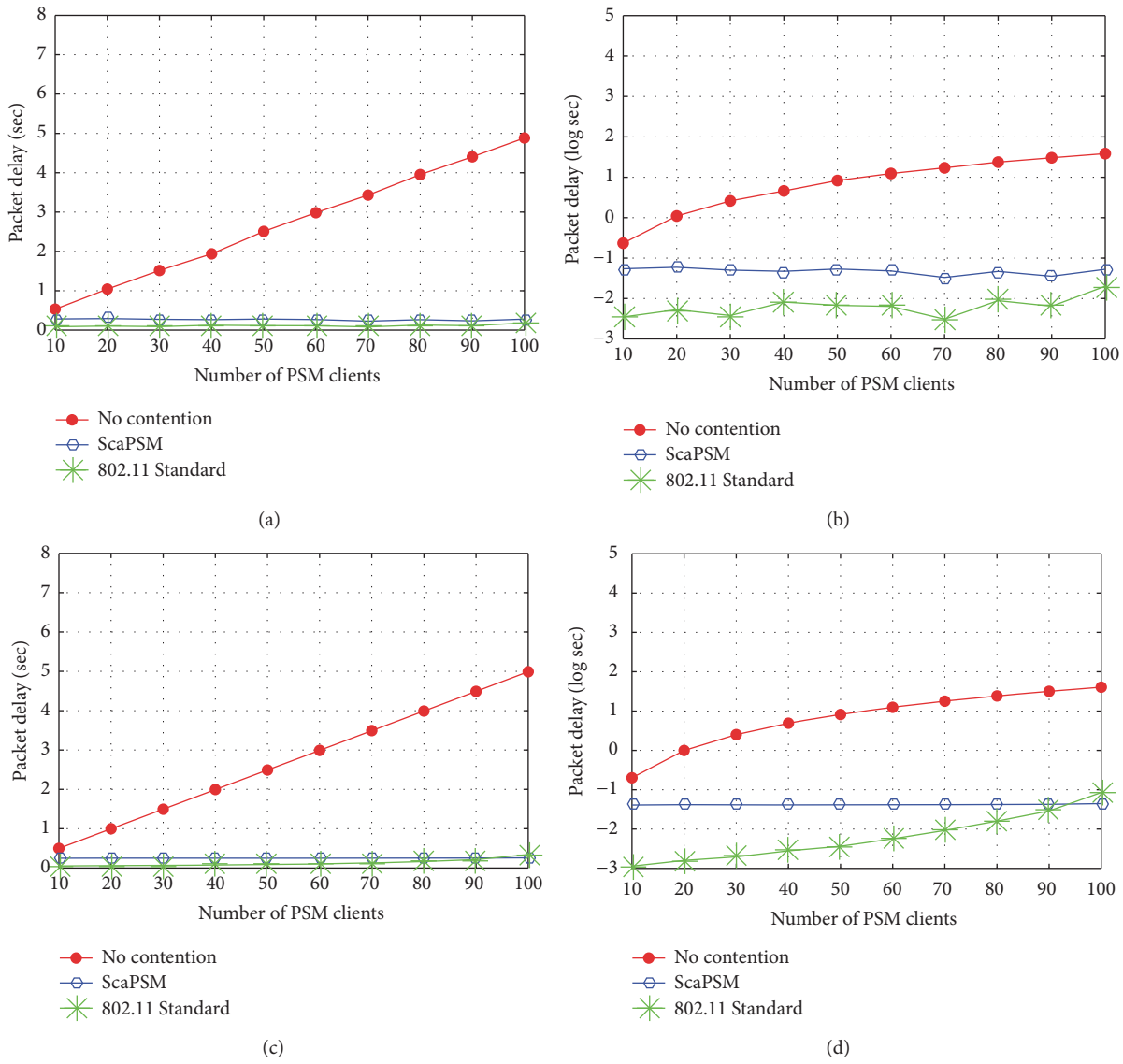


FIGURE 4: (a) Packet delay versus number of PSM clients under controlled traffic; (b) packet delay versus number of PSM clients by logarithmic function under controlled traffic; (c) packet delay versus number of PSM clients under trace-driven traffic; (d) packet delay versus number of PSM clients by logarithmic function under trace-driven traffic.

channel access with associated clients equally. NAPman [4] implements a new energy-aware fair scheduling algorithm at AP to minimize Wi-Fi radio wake-up time and eliminate unnecessary retransmissions in the presence of competing traffic. Further, NAPman leverages AP virtualization to make different PSM clients wake up at staggered time intervals, so that these clients can monopolize wireless channel and receive TIM separately. However, previous efforts generally consider how to save energy by reducing or even eliminating contention among competing PSM clients and also ignore the negative impact on packet delay performance due to energy conservation. Moreover, all these solutions are not for crowd events scenario where scalability is considered as a major issue. SleepWell [18] coordinates the activity circles of multiple APs to allow clients to sleep longer, and therefore this technique may be complementary to ScaPSM.

5.3. Beacon Management Method for WLAN. Lee et al. [13] propose a beacon management scheme that restricts the number of nodes in wake mode in each beacon interval with the maximum number of packets to be delivered according to the transmission duration. However, they only consider the power efficiency when network congestion occurs, instead of the whole views, such as the trade-off between energy consumption and delay performance. EDP [19] provides an analytical model for energy consumption and packet delay in highly congested 802.11 networks and proposes a power-saving strategy to determine the number of PSM clients in wake mode that balances energy consumption and packet delay. However, EDP does not consider fairness issues among PSM clients. Moreover, these schemes are not for crowd event scenarios where a large amount of PSM clients with the delay-aware requirement simultaneously compete for access.

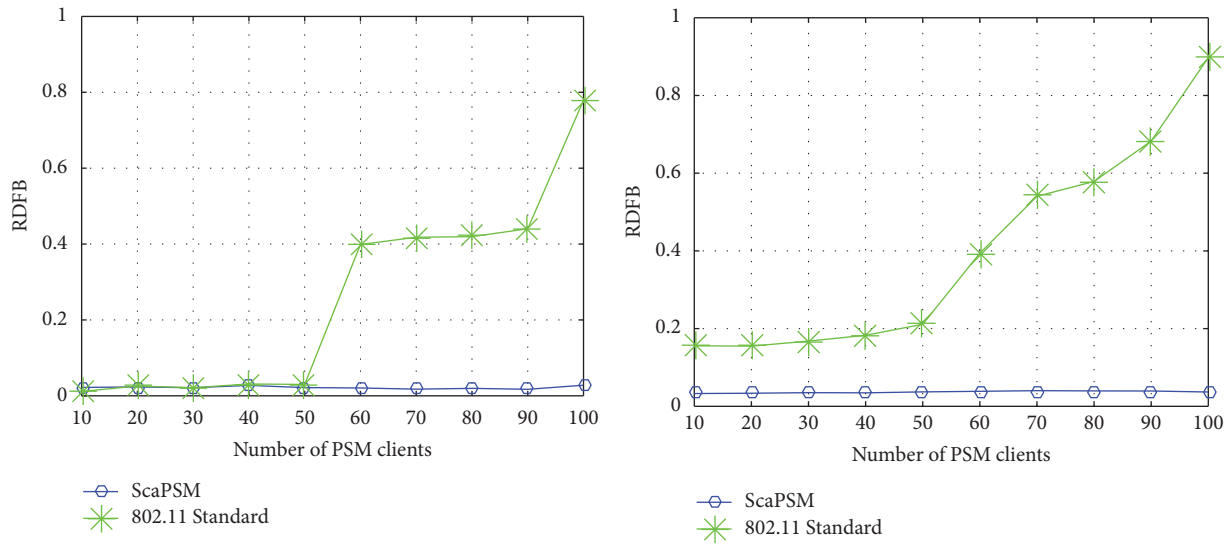


FIGURE 5: (a) Delay fairness versus number of PSM clients under controlled traffic; (b) delay fairness versus number of PSM clients under trace-driven traffic.

6. Conclusions

In this paper, we address the energy issues of mobile devices in 802.11 networks for crowd events. We propose an online competing background traffic scheduling algorithm to improve the client energy efficiency, while ensuring the packet delay performance. Different from existing work, we formulate the delay performance degradation problem and build a comprehensive metric to capture the impact of delay performance and delay fairness. Our evaluation results demonstrate the effectiveness of our proposed schemes in achieving better performance over existing work. We further validate the high energy saving of our proposed scheduling algorithm with increasing the number of PSM clients through controlled and trace-driven simulations. In our future work, we will investigate the impact of application traffic and heterogeneous mobile devices for crowd events.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] LAN/MAN Standards Committee, S. Committee, and I. Computer, *Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, vol. 2012, 2012.

[2] Y. He and R. Yuan, "A novel scheduled power saving mechanism for 802.11 wireless LANs," *IEEE Transactions on Mobile Computing*, vol. 8, no. 10, pp. 1368–1383, 2009.

[3] H.-P. Lin, S.-C. Huang, and R.-H. Jan, "A power-saving scheduling for infrastructure-mode 802.11 wireless LANs," *Computer Communications*, vol. 29, no. 17, pp. 3483–3492, 2006.

[4] E. Rozner, V. Navda, R. Ramjee, and S. Rayanchu, "NAPman: network-assisted power management for WiFi devices," in *Proceedings of the 8th Annual International Conference on Mobile Systems, Applications and Services (MobiSys '10)*, pp. 91–105, San Francisco, Calif, USA, June 2010.

[5] Z. Zeng, Y. Gao, and P. R. Kumar, "SOFA: A sleep-optimal fair-attention scheduler for the power-saving mode of WLANs," in *Proceedings of the 31st International Conference on Distributed Computing Systems, ICDCS 2011*, pp. 87–98, July 2011.

[6] Super bowl plans to handle 30,000 wi-fi users at once and sniff out 'rogue devices'. http://arstechnica.com/information-technology/2013/02/super-bowl-plans-to-handle-30000-wi-fi-users-at-once-and-sniff-out-rogue-devices/, 2013.

[7] J. Erman and K. K. Ramakrishnan, "Understanding the super-sized traffic of the super bowl," in *Proceedings of the ACM IMC'13*, pp. 353–359, 2013.

[8] A. Gupta, J. Min, and I. Rhee, "WiFox: scaling WiFi performance for large audience environments," in *Proceedings of the 8th ACM International Conference on Emerging Networking EXperiments and Technologies (CoNEXT '12)*, pp. 217–228, ACM, December 2012.

[9] J. Huang, F. Qian, Z. M. Mao, S. Sen, and O. Spatscheck, "Screen-off traffic characterization and optimization in 3G/4G networks," in *Proceedings of the 2012 ACM Internet Measurement Conference, IMC 2012*, pp. 357–363, November 2012.

[10] Y. Cui, S. Xiao, X. Wang, M. Li, H. Wang, and Z. Lai, "Performance-aware energy optimization on mobile devices in cellular network," in *Proceedings of the 33rd IEEE Conference on Computer Communications, IEEE INFOCOM 2014*, pp. 1123–1131, May 2014.

[11] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, 2000.

[12] W. Wang, B. Liang, and B. Li, "Low complexity multi-resource fair queueing with bounded delay," in *Proceedings of the 33rd IEEE Conference on Computer Communications, IEEE INFOCOM 2014*, pp. 1914–1922, May 2014.

[13] J. R. Lee, S. W. Kwon, and D. H. Cho, "A new beacon management method in case of congestion in wireless lans," in *Proceedings of the IEEE VTC'05*, pp. 12–15, 2005.

[14] R. M. Karp, "Reducibility among combinatorial problems," *Complexity of Computer Computations*, pp. 85–103, 1972.

[15] A. Schulman, D. Levin, and N. Spring, 2008, Crawdad data set umd/sigcomm2008.

[16] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang, "A first look at cellular network performance during crowded events," in *Proceedings of the 2013 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS 2013*, pp. 17–28, June 2013.

[17] Y. Bejerano, J. Ferragut, K. Guo et al. et al., "Scalable WiFi multicast services for very large groups," in *Proceedings of the ICNP'13*, pp. 1–12, 2013.

[18] J. Manweiler and R. R. Choudhury, "Avoiding the rush hours: WiFi energy management via traffic isolation," in *Proceedings of the ACM MobiSys'11*, pp. 253–266, 2011.

[19] D. Jung, R. Kim, and H. Lim, "Power-saving strategy for balancing energy and delay performance in WLANs," *Computer Communications*, vol. 50, pp. 3–9, 2014.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 2539673, 8 pages
https://doi.org/10.1155/2017/2539673

Research Article

Privacy-Preserving Meter Report Protocol of Isolated Smart Grid Devices

Zhiwei Wang and Hao Xie

School of Computer Sciences, Nanjing University of Posts and Telecommunications, Nanjing, China

Correspondence should be addressed to Zhiwei Wang; zhwwang@njupt.edu.cn

Received 17 January 2017; Revised 24 April 2017; Accepted 4 May 2017; Published 6 June 2017

Academic Editor: Zhe Yang

Smart grid aims to improve the reliability, efficiency, and security of the traditional grid, which allows two-way transmission and efficiency-driven response. However, a main concern of this new technique is that the fine-grained metering data may leak the personal privacy information of the customers. Thus, the data aggregation mechanism for privacy protection is required for the meter report protocol in smart grid. In this paper, we propose an efficient privacy-preserving meter report protocol for the isolated smart grid devices. Our protocol consists of an encryption scheme with additively homomorphic property and a linearly homomorphic signature scheme, where the linearly homomorphic signature scheme is suitable for privacy-preserving data aggregation. We also provide security analysis of our protocol in the context of some typical attacks in smart grid. The implementation of our protocol on the Intel Edison platform shows that our protocol is efficient enough for the physical constrained devices, like smart meters.

1. Introduction

While the swift advances in smart grid are triggering radical innovations in this field, today's power grid is widely different from the traditional grid [1–4]. Traditional grid has the characteristic of centralized one-way transmission, which only transmits electricity from the generation plants to customers. Smart grid is featured with intelligent transmission (decentralized two-way transmission) and distribution networks, which combines the traditional grid and the new information processing technologies. On the one hand, smart grid integrates more green energies such as solar and wind power into energy supply; on the other hand, it improves the reliability, security, and efficiency of electric system by two-way communication of consumption data and other electric system's operations. In general, smart grid can realize the intelligent electricity generation, resource allocation, and dynamic pricing.

In this system, smart grid devices such as smart meters play an important role for collecting the power usage data and the status data. Such data are generated by some plug-in monitor sensors. In general, the smart grid data communication network can be divided into four layers [5] as Figure 1 shows. Various sensors and other smart grid devices consisting of a home area network are the first layer. Then, the smart meters and a neighborhood gateway which form a neighborhood area network are the second layer. Furthermore, all the neighborhood gateways connecting each other consist of the third layer network. Moreover, the forth layer network is a high speed public network through fiber gateways which is responsible for transfer all the data to the data center in electricity service provider (ESP).

However, not all smart grid devices are connected to the smart grid data communication network, due to the network outrage or opt-out agreement between the customers and the ESP. According to the utility-scale smart meter deployments report [6] published by Electric Innovation at Edison Foundation, the smart meters only cover 43% US homes. Some smart grid devices are located sparsely and far away from the data center of ESP. Thus, it would be a heavy cost to extend the smart grid data communication network for covering such isolated smart grid devices. Moreover, some in-network smart grid devices also will be disconnected from the smart grid network due to the natural disasters such as tornado and

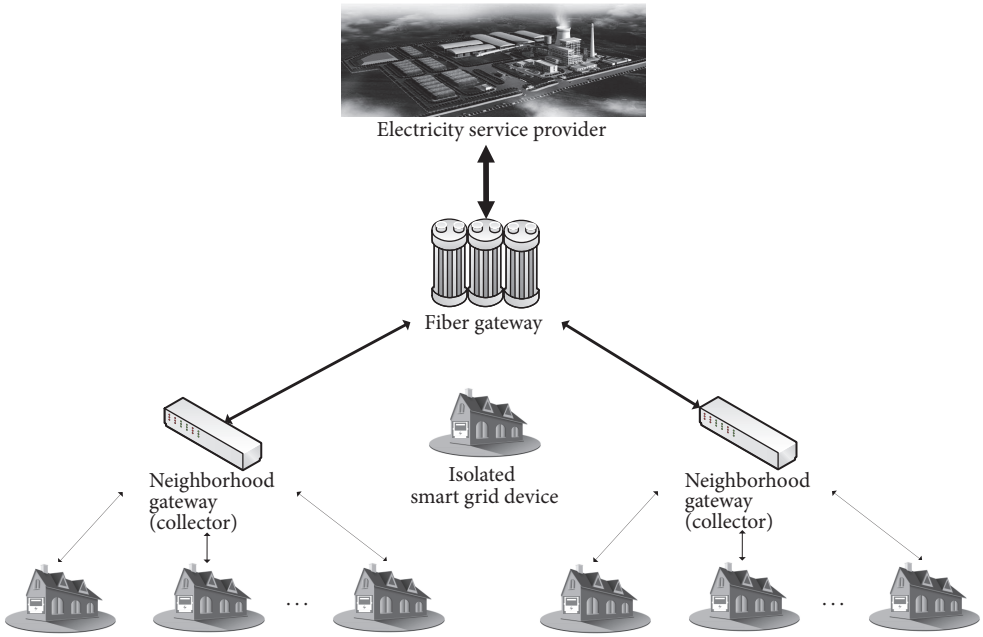


Figure 1: Smart grid data communication network.

earthquake. Thus, for such isolated smart grid devices, the ESP may send a worker to the location of them and read the power usage data by using the handhold smart meter reader.

In general, several protocols are used in smart grid communication network [7], for the propose of authentication, power allocation, meter reporting, and so on. The meter report protocol is used to calculate the total monthly power consumption data for each individual customers. For the isolated smart grid device, a smart reader device should be used as a bridge between the ESP and it as Figure 2 shows. Although the smart reader device needs to read the smart meter more frequently for monitoring the energy supply, the ESP only needs to obtain the total long-term consumption data for the energy forecast.

Up to now, several privacy-protection aggregation schemes have been proposed. Li et al. [8] constructed an incremental aggregation scheme based on a virtual aggregation tree which relies on the topology of network. Garcia and Jacobs [7] proposed an aggregation scheme combined with additive secret sharing. Lu et al. [9] proposed an efficient privacy-preserving scheme for multidimensional data structure. The three schemes are all based on Pallier's homomorphic encryption technology. Fan et al. [10] proposed data aggregations scheme based on the subgroup indistinguishability assumption. All the above aggregation schemes are designed for the in-network smart grid devices, and they are used to aggregate individual usage date from different customers. For the isolate smart grid devices, Sha et al. [5] proposed a secure and efficient authentication protocol, but their meter report protocol did not provide a data aggregation mechanism for privacy-preserving. For the isolated smart grid devices, there exists the same drawback as in-network devices that fine-grained power usage data may leak the personal privacy information [11, 12]. If a corrupted worker in the ESP can obtain the fine-grained power usage data, then he can analyze the daily activities of the customer. Thus, a secure data aggregation mechanism for privacy protection is also required for isolated smart grid devices. The fine-grained power usage data should be protected in the reader device and cannot be leaked to anyone else.

This paper aims to propose an efficient privacy-preserving meter report protocol for the isolate smart grid devices. The protocol not only contains an additively homomorphic encryption scheme used to aggregate the encrypted data but also includes a linearly homomorphic signature scheme [13, 14] for protection against unintentional errors and altering messages in malicious. Furthermore, both the isolated smart grid devices and the reader devices have only restricted resources, and thus both the encryption and signature schemes should provide the high performance in terms of efficiency.

The contributions of this paper can be listed as follows: (1) We propose an encryption scheme with additively homomorphic property to aggregate the encrypted metering data. To be compatible with the data aggregation, we also propose a linearly homomorphic signature scheme which is used to sign the ciphertext of metering data. The signatures will be aggregated along with the ciphertexts stored in the reader device. This allows the ESP to verify the correctness of aggregated result by checking the aggregation signature. (2) We provide a security analysis to our meter report protocol in context of several typical attacks in smart grid. (3) To evaluate the appropriacy of our meter report protocol for the resource-constrained devices, we implement our protocol on the Intel Edison platform which is a development system for Internet of Things (IoT) devices.

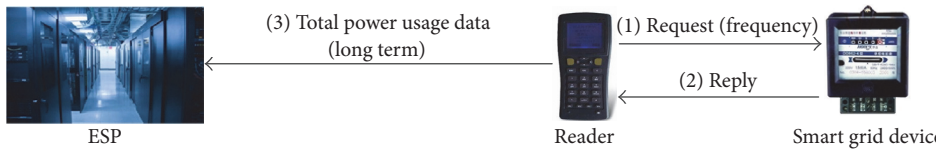


Figure 2: Meter report of isolated smart grid devices.

Organization. Related mathematical concepts to our construction and proofs are reviewed in Section 2. The privacy-preserving meter report protocol for isolated smart grid devices is proposed in Section 3. We analyze our protocol against several typical attacks in Section 4. Section 5 discusses the performance of our protocol on the platform of MacBook Pro and Edison. Finally, we conclude our paper in Section 6.

2. Preliminary

In this section, we review related mathematical concepts for our construction and proofs.

Assuming that G and G_T are two cyclic groups with the prime order p, we define $e : G \times G \to G_T$ to be the bilinear map as it has the following properties:

(1) Bilinear: $\forall g_1, g_2 \in G$, $a_1, a_2 \in Z_p$, $e(g_1^{a_1}, g_2^{a_2}) = e(g_1, g_2)^{a_1 a_2}$.

(2) Nondegenerate: $\exists g \in G, e(g, g) \neq 1$.

(3) Efficient computability: there exists an efficient algorithm to compute $e(g_1, g_2)$ for all $g_1, g_2 \in G$.

We define the q-strong Diffie-Hellman (q-SDH) assumption over G as follows.

Definition 1 (q-SDH assumption). Let $\text{Gen}(1^\iota)$ be a group generation algorithm that takes a security parameter ι as input and outputs a description of a prime order group $\Theta = \{p, G, G_T, e\}$. The q-SDH assumption over group G states that, for any probabilistic polynomial-time (PPT) attackers, given a tuple $(g, g^\beta, g^{\beta^2}, \ldots, g^{\beta^q})$ for randomly chosen $\beta \xrightarrow{R} Z_p$ and $g \xrightarrow{R} G$, the advantage for obtaining a solution $(\gamma, g^{1/(\beta+\gamma)})$ is negligible in ι, where $\gamma \in Z_p$.

Next, we define two composite order groups (G', G'_T) with order $N = pq$, where p and q are distinct large primes. Thus, G is a product of two groups $G' = G'_p \times G'_q$, and their orders are p and q, respectively. In essence, the subgroup indistinguishability assumption is that an element in group G' is computationally indistinguishable from a random element in G'_p or G'_q. Let g' be a generator of G'. We define a nongenerate and efficiently computable bilinear map $e : G' \times G' \to G'_T$ over G' and G'_T. The subgroup indistinguishability assumption [15] can be described as follows.

Definition 2 (subgroup indistinguishability assumption). Let $\text{Gen}(1^\iota)$ be a group generation algorithm that takes a security parameter ι as input and outputs a description of a multiplicative group $\Psi = \{p, q, G', G'_T, e'\}$, where $G' = G'_p \times G'_q$.

The subgroup indistinguishability assumption over group G' states that, for any PPT attackers, the advantage

$$\begin{aligned}&\text{Adv}_{\mathbf{A}}(\iota)\\ &= \left|\Pr\left[\mathbf{A}(\Psi, x) = 1; x \longleftarrow_R G'\right] - \Pr\left[\mathbf{A}(\Psi, x^q) = 1\right]\right|,\\ &\text{Adv}_{\mathbf{A}}(\iota)\\ &= \left|\Pr\left[\mathbf{A}(\Psi, x) = 1; x \longleftarrow_R G'\right] - \Pr\left[\mathbf{A}(\Psi, x^p) = 1\right]\right|\end{aligned} \tag{1}$$

is negligible in ι.

3. Design of Meter Report Protocol

3.1. System Model. There are three parties including electricity service provider (ESP), reader, and isolated smart grid device in the system model of the proposed protocol. The ESP and the isolated smart grid device should setup their public/secret key pairs and other public information. When the reader tries to frequently collect the encrypted metering data from the isolated smart grid device, several attacks may be possible. Firstly, an attacker may listen to the communications between the reader and the isolated smart grid device to obtain the metering data or alter the messages. Secondly, a corrupted reader may be used to obtain the power usage data. Thirdly, a corrupted reader may provide an incorrect total power usage data to the ESP. Finally, a fake ESP worker may analyze the power usage data with fine granularity to identify the daily activities of the customer.

In the meter report model as Figure 2 shows, the reader needs to much more frequently read from the smart grid device for monitoring the energy supply. Each time the reader reads, the smart grid device encrypts its metering data with a random number and signs it before he sends it to the reader. After a long term, the ESP can only obtain the total power usage data of the customer.

3.2. Construction. The proposed protocol consists of four phases, which will be described in detail as follows. Some notations can be defined here.

(i) $H : \{0, 1\}^* \to Z_N^*$ is a one-way hash function.

(ii) t is the tag of currently regular period.

(iii) ID_{esp} is the identity information of electricity service provider.

(iv) r_i is the ith random number chosen by smart grid device.

(v) r_0 is the sum of random numbers $\sum r_i$.

(vi) α is the secret key of isolated smart grid device.

(vii) Y is the public key of isolated smart grid device.

(1) Setup Phase

(i) ESP: the ESP randomly chooses two distinct large primes (p, q) and computes the RSA parameter $N = pq$ (example initiation: let $\mathbf{P}$, p, and q be distinct large primes such that $\mathbf{P} = 2pq + 1$. Obviously, $Z_{\mathbf{p}}^*$ is a quadratic residue group with order $N = pq$. $Z_{\mathbf{p}}^*$ can be denoted as $Z_{\mathbf{p}}^* = G_p \times G_q$, where G_p and G_q are both prime order cyclic groups. Gonzalez et al. proved that the subgroup decision assumption over $Z_{\mathbf{p}}^*$ holds if the factoring problem over N is hard). It generates u in group G with order N and produces a generator g of the subgroup G_q. Then, it computes $h = u^q$, which is an element in subgroup G_p. Finally, the ESP publishes the public parameters $\{N, u, h, g, \mathrm{ID}_{\mathrm{esp}}\}$ where $\mathrm{ID}_{\mathrm{esp}}$ is its identity information and keeps $\{p, q\}$ as the secret information.

(ii) Isolated smart grid device: the isolated smart grid device randomly chooses $\alpha \in Z_N^*$ as its secret key and publishes the public key $Y = g^{\alpha}$. Then, let $\mathrm{ID}_{\mathrm{esp}}$ denote the identity of ESP who is the customer's energy supplier.

(2) Reading Phase

(i) Isolated smart grid device: when the reader needs to read the metering data m_i for the ith time in a long term, the isolated smart grid device chooses $r_i \in Z_N^*$ randomly and computes a ciphertext $\mathrm{CT}_i = g^{m_i} h^{r_i}$. We assume that reader reads the metering data n times during such a long term. There is a limitation that $r_0 = \sum_{i=1}^{n} r_i$ should not be a large number. Then, the smart grid device computes a signature

$$\begin{aligned} \sigma_1 &= g^{1/(\alpha + H(\mathrm{ID}_{\mathrm{esp}} \| t))}, \\ \sigma_{2i} &= \left(\mathrm{CT}_i\right)^{1/(\alpha + H(\mathrm{ID}_{\mathrm{esp}} \| t))}, \end{aligned} \tag{2}$$

where t is the tag of currently regular period. Finally, it sends $\{\mathrm{CT}_i, (\sigma_1, \sigma_{2i})\}$ to the reader.

(ii) Reader: after receiving $\{\mathrm{CT}_i, (\sigma_1, \sigma_{2i})\}$, the reader verifies identity of its ESP and the currently long term by checking $e(\sigma_1, Y \cdot g^{H(\mathrm{ID}_{\mathrm{esp}} \| t)}) = e(g, g)$. Here, the reader verifies the smart grid device's first signature component to assure that who is its ESP and to avoid that the customer will make payments for an improper ESP. If the signature σ_1 is true, then the reader stores $\{\mathrm{CT}_i, (\sigma_1, \sigma_{2i})\}$.

(3) Aggregation Phase

(i) Isolated smart grid device: at the end of a long term, the isolated smart grid device encrypts r_0 as $\mathrm{CT}_{r_0} = g^{r_0} h^s$ with a random number $s \in Z_N^*$ and sends it to the reader.

(ii) Reader: after receiving $\mathrm{CT}_{r_0} = g^{r_0} h^s$, the reader needs to aggregate the total power usage data of the isolated smart grid device. We assume that the reader has read the smart grid device n times during this long term, and thus n ciphertext/signature pairs $\{\mathrm{CT}_i, (\sigma_1, \sigma_{2i})\}_{i \in [1,n]}$ have been stored in the reader. Then, the reader computes $\mathrm{CT} = \prod_{i=1}^{n} \mathrm{CT}_i$ and $\sigma_2 = \prod_{i=1}^{n} \sigma_{2i}$, and reports $\{\mathrm{CT}, (\sigma_1, \sigma_2), \mathrm{CT}_{r_0}\}$ to the ESP.

(4) Decryption and Verification Phase

(i) ESP: when the ESP receives $\{\mathrm{CT}, (\sigma_1, \sigma_2), \mathrm{CT}_{r_0}\}$, it firstly verifies its identity information and the currently long term by checking $e(\sigma_1, Y \cdot g^{H(\mathrm{ID}_{\mathrm{esp}} \| t)}) = e(g, g)$ and then computes $W = \mathrm{CT}_{r_0}^p = (g^p)^{r_0}$ and $\widetilde{g} = g^p$. Since r_0 is not a large number, the ESP can compute the discrete log of W on the base of $\widetilde{g}$ by using Pollard's lambda method [16] in polynomial time. Then, the ESP computes $V = \mathrm{CT} \cdot h^{-r_0} = g^{\sum_{i=1}^{n} m_i}$. Since the total power usage data $M = \sum_{i=1}^{n} m_i$ is also not a large number, the ESP can compute the discrete log of V on the base of g. Finally, the ESP computes $\varrho = \sigma_2^p$ and verifies σ_2 by checking $e(\varrho, Y \cdot g^{H(\mathrm{ID}_{\mathrm{esp}} \| t)}) = e(\widetilde{g}, g)^M$.

The correctness of the above formulas can be depicted as follows.

Authentication of Its ESP

$$\begin{aligned} &e\left(\sigma_1, Y \cdot g^{H(\mathrm{ID}_{\mathrm{esp}} \| t)}\right) \\ &\quad = e\left(g^{1/(\alpha + H(\mathrm{ID}_{\mathrm{esp}} \| t))}, g^{\alpha} \cdot g^{H(\mathrm{ID}_{\mathrm{esp}} \| t)}\right) = e\left(g, g\right). \end{aligned} \tag{3}$$

Ciphertext Decryption

$$\begin{aligned} W &= \mathrm{CT}_{r_0}^p = g^{r_0 \cdot p} \cdot h^{s \cdot p} = g^{r_0 \cdot p} \cdot u^{s \cdot p \cdot q} = \left(g^p\right)^{r_0} \cdot 1 \\ &= \widetilde{g}^{r_0}, \\ V &= \mathrm{CT} \cdot h^{-r_0} = g^{\sum_{i=1}^{n} m_i} \cdot h^{\sum_{i=1}^{n} r_i + (-r_0)} = g^{\sum_{i=1}^{n} m_i} \cdot h^0 \\ &= g^{\sum_{i=1}^{n} m_i}. \end{aligned} \tag{4}$$

Aggregate Signature Verification

$$\begin{aligned} \varrho = \sigma_2^p &= \left(\left(g^{\sum_{i=1}^{n} m_i} \cdot h^{\sum_{i=1}^{n} r_i}\right)^{1/(\alpha + H(\mathrm{ID}_{\mathrm{esp}} \| t))}\right)^p \\ &= \left(\left(g^p\right)^{\sum_{i=1}^{n} m_i} \cdot u^{\sum_{i=1}^{n} r_i \cdot p \cdot q}\right)^{1/(\alpha + H(\mathrm{ID}_{\mathrm{esp}} \| t))} \\ &= \left(\left(g^p\right)^{\sum_{i=1}^{n} m_i} \cdot 1\right)^{1/(\alpha + H(\mathrm{ID}_{\mathrm{esp}} \| t))} \\ &= \left(\widetilde{g}^{\sum_{i=1}^{n} m_i}\right)^{1/(\alpha + H(\mathrm{ID}_{\mathrm{esp}} \| t))}. \end{aligned} \tag{5}$$

Thus,

$$e\left(\varrho, Y \cdot g^{H(\mathrm{ID}_{\mathrm{esp}}\|t)}\right) = e\left(\left(\widetilde{g}^{\sum_{i=1}^{n} m_i}\right)^{1/(\alpha+H(\mathrm{ID}_{\mathrm{esp}}\|t))}, g^{\alpha+H(\mathrm{ID}_{\mathrm{esp}}\|t)}\right) = e\left(\widetilde{g}, g\right)^{M}. \quad (6)$$

4. Security Analysis

Our privacy-preserving meter report protocol is proposed not only to prevent the unauthorized parties to read or alter the metering data from the isolated smart grid devices, but also to securely aggregate the fine-grained power usage data in a long term. Here, we show the security properties of our scheme in context of six typical attacks in smart grid.

4.1. Against External Attack. The external attackers can eavesdrop on the communication channels to obtain the unauthorized information. In our protocol, all the metering data are encrypted, which provide strong protection to the external attackers. The proof of Theorem A.2 in Appendix shows that our encryption scheme satisfies the CPA secure under the subgroup indistinguishability assumption. The external attackers also cannot alter a metering data of the isolated smart grid device, since they cannot forge a valid signature. Theorems A.4 and A.5 in Appendix show that our linearly homomorphic signature schemes are unforgeable under the q-SDH assumption and Boneh and Boyen signature.

4.2. Against Smart Grid Device Attack. A smart grid device attack is that a fake smart grid device aims to mimic a legitimate device. In our design, we use the signature technology to prevent a fake smart grid device from authenticating with the reader and ESP. Moreover, a fake smart grid device may want to let the customer to pay for an improper ESP, but our design can also avoid this situation, since the first component of linearly homomorphic signature is a signature of the proper ESP's identity, and its unforgeable security is under Boneh and Boyen signature (the security proof of Theorem A.4 can be seen in Appendix).

4.3. Against Internal (Reader) Attack. An attacker may use a lost legitimate reader to obtain the unauthorized information or maliciously alter total the power usage data of a smart grid device, which is called the internal (reader) attack. In reading phase, the legitimate reader only can verify the signature of device's identity. But the power usage data m_i cannot be recovered from the ciphertext $\mathrm{CT}_i = g^{m_i} h^{r_i}$, since the reader cannot get the ESP's secret key (p, q). In aggregation phase, the reader also cannot decrypts CT_{r_0} to get r_0 and obtains the total power usage data. On the other hand, the linearly homomorphic signature and the encryption of r_0 prevent the reader from altering the total power usage data, since it does not know the secret key α of the isolated smart grid device. The unforgeability of our linearly homomorphic signature scheme has been proved by Theorems A.4 and A.5. The properties of linearly homomorphic signature also protect the correctness and integrity of the total power usage data.

4.4. Against Internal (ESP) Attack. We assume that the legitimate workers of ESP make the malicious attacks. After receiving the ciphertext/signature pair $\{\mathrm{CT}, (\sigma_1, \sigma_2), \mathrm{CT}_{r_0}\}$ from the reader, the ESP can compute $V = \mathrm{CT} \cdot h^{-r_0} = g^{\sum_{i=1}^{n} m_i}$ to recover the total power usage data. However, the ESP cannot decrypt the individual metering data m_i from CT and r_0, since it does not know each corresponding random number r_i.

4.5. Against Man-In-The-Middle Attack. A Man-In-The-Middle attacker aims to mimic the right person to fool one side by using the information from another side. In reader-device and ESP-device authentication, a public key based linearly homomorphic signature scheme is used to authenticate the device's identity and the ciphertexts. It provides the strong defense for the Man-In-The-Middle attacks, since the attacker cannot convince the reader and ESP to accept its public key.

4.6. Against Replay Attack. If an attacker obtains the information between the communication of two sides, then he intercepts the communication and replays the information maliciously, which is called replay attack. In our designing, we use the tag of currently term t to prevent the replay attack from different terms. If the attacker wants to modify t in device's signature for the replay attack, then he should get the device's secret key α. However, it is almost impossible to guess the device's secret key. If an attacker wants to make replay attack in the same period, then it should modify r_0 in ciphertext CT_{r_0} that is also impossible.

5. Performance Analysis

Let P denote the pairing computation cost, E denote the exponent cost, and Mu denote the point multiplication. Table 1 shows the computational complexity of our protocol.

Following the theoretical analysis, we test our scheme on two different platforms, where one is a normal personal computer, and the other is a resource-constrained device. We implement our protocol in C with the pairing based cryptography (PBC) library [17] for the underlying arithmetic and pairing operations. We use the Type-A curves as defined in PBC library for the implementation, since the Type-A curves offers the highest efficiency among all the three types of curves.

The first test machine is MacBook Pro with Intel core i5 CPU (2.5 GHz) running Os X 10.9.3, which RAM is 4 GB. The second test machine is Intel Edison development platform, which is designed to rapidly prototype and produce Internet of Things (IoT) products. Since the isolated smart grid device and reader device are usually resource-constrained devices, we test our protocol on this platform. We use Edison platform with a dual-core, dual-threaded Intel Atom CPU at 500 MHz and 1 GB RAM, running Yocto Linux v1.6.

Table 2 shows the time cost of reading phase for smart grid device and reader. We compute the average value on 100 randomized runs. The time cost of isolated smart grid device is about 0.43 seconds, if our protocol is run over the Edison platform. For the reader, it needs 0.42 seconds to verify the signature, while the protocol is run over the Edison platform. In aggregation phase, the time cost of isolated device is about

Table 1: Computational complexity of our protocol.

Computational complexity	ESP	Reader	Isolated device
Reading phase	Null	$1P + 1\text{Mu} + 1E$	$1\text{Mu} + 4E$
Aggregation phase	Null	$2n\text{Mu}$	$1\text{Mu} + 2E$
Decryption and verification phase	$3P + 3\text{Mu} + 6E$	Null	Null

Table 2: Time cost in reading phase.

Platform	MacBook	Edison
Smart grid device	0.02 s	0.43 s
Reader	0.016 s	0.42 s

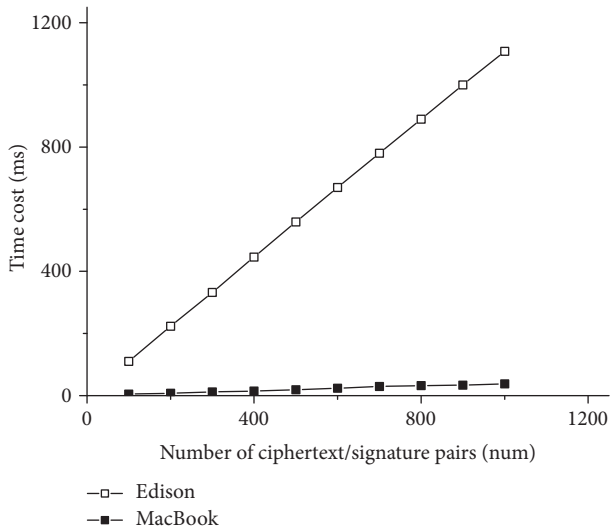


Figure 3: Time Cost of reader in aggregation phase.

1.5 milliseconds on the Edison platform, while it needs about 0.06 milliseconds over MacBook Pro. Figure 3 shows the time cost of reader in aggregation phase. We can see that the time consuming of reader is increased by the number of ciphertext/signature pairs to be aggregated. The time cost of decryption for the ESP is about 77 milliseconds. Although the total power usage data M is increased by the number of individual consumption data m_i, the computation of the discrete log of V is very slightly raised.

6. Conclusion

In practical, the fine-grained individual power consumption data may leak the personal privacy information of the users. Thus, in order to protect the personal privacy, data aggregation mechanism should be designed in the meter report protocol. In this paper, we propose an efficient privacy-preserving meter report protocol for the isolated smart grid devices, which consists of an encryption scheme with additively homomorphic property and a linearly homomorphic signature scheme. To prevent unauthorized seeing the intermediate metering data, the metering data should be encrypted by using the encryption scheme with additively homomorphic property and aggregated using such a property. Besides the encryption scheme, a linearly homomorphic signature scheme which is compatible with data aggregation is also designed in our protocol for verifying the correctness and integrity of the aggregation result. We give security analysis to our protocol in context of six typical attacks in smart grid. The implementation of our protocol on the Edison platform shows that our protocol is efficient enough for the resource-constrained devices.

Appendix

Here, we provide the security proofs to the encryption scheme and the linearly homomorphic signature scheme used in the proposed meter report protocol.

Definition A.1. A public key encryption scheme is CPA secure, if for all the advantage of any PPT attacker A in the following game is negligible in the security parameter ι.

Setup. The challenger obtains the public/secret key pair (pk, sk) by running Setup(1^ι) and sends pk to the attacker, where the public key includes a message space $\mathbf{M}$ and a ciphertext space Ξ. The challenger sets $\text{Enc}(pk, m)$ as an encryption algorithm.

Challenge. The attacker sends two messages m_0 and m_1 with the same length to the challenger. Then, the challenger responds the challenge ciphertext $\text{CT} = \text{Enc}(pk, m_b)$ under a random bit b.

Output. The attacker outputs its guess b' to b. If $b' = b$, then the attacker wins the game.

Theorem A.2. *If the subgroup indistinguishability assumption holds on G, then the above encryption scheme is CPA secure.*

Proof. We assume that there exists an attacker **A** which can break the above encryption scheme with nonnegligible probability $\epsilon(\iota)$ and a challenger **C** that takes an instance of subgroup indistinguishability assumption. We will prove the theorem by an interaction game between **A** and **C**.

Setup. The challenger **C** is given an instance (N, G, G_T, x) of subgroup indistinguishability assumption and generates a generator $g \in G$. Then, it sends the public parameters (N, G, G_T, x, g) to the attacker **A**.

Challenge. **A** chooses two messages m_0 and m_1 with the same length and then sends them to **C**. **C** chooses a random number $r^* \in Z_N^*$ and returns the challenger ciphertext $\text{CT}^* = g^{m_b} x^{r^*}$, where b is a random bit.

Output. **A** outputs its guess b' to b, and if $b' = b$, then **A** wins the game and **C** outputs that "x is uniformly in G_p". Otherwise, **C** outputs that "x is uniformly in G".

If x is uniformly in G, then the challenge ciphertext CT is randomly in G_T, which is independent of b. Thus, $\Pr[b' = b] = 1/2$ in this case. However, if x is uniformly in G_p, then $\Pr[b' = b] = 1/2 + \epsilon(\iota)$ in this case since **A** can break the above encryption scheme with the probability of $\epsilon(\iota)$. The probability difference of these two cases is $\epsilon(\iota)$, which is nonnegligible in our assumption. But it contradicts that the subgroup indistinguishability assumption is hard. Thus, our assumption is not correct, and the encryption scheme is CPA secure. □

Our linearly homomorphic signature scheme is based on Boneh and Boyen signature [18], which has been proved strongly unforgeability against a weak attacker under the q-SDH assumption. Here, we will firstly provide the security definition of linearly homomorphic signature.

Definition A.3. An linearly homomorphic signature scheme is simply unforgeable, if for all the advantage of any PPT attacker **A** in the following game is negligible in the security parameter ι.

Setup. The challenger obtains the public/secret key pair (pk, sk) by running Setup(1^ι) and sends pk to the attacker, where the public key includes a message space **M** and a signature space Σ. The challenger sets Sign as the signing algorithm and Verify as the verification algorithm.

Queries. The attacker sends a random number $x \in \{0,1\}^*$ and a message $m \in \mathbf{M}$ to the challenger for a signature query. Then, if m is the first query for x, the challenger randomly chooses a tag $\tau_x \in Z_N^*$ and gives it to the attacker. Otherwise, the challenger looks up the previously chosen τ_x. The challenger then returns the signature $\sigma \leftarrow \text{Sign}(sk, \tau_x, m)$. This query can be repeated for a polynomial times; however there is a restriction that at most n message can be queried for one tag τ_x. We let V_x denote the set of elements m queried for x.

Output. The attacker outputs a tag $\tau^* \in Z_N^*$, a message $m^* \in \mathbf{M}$, and a signature $\sigma^* \in \Sigma$. The attacker wins if $\text{Verify}(pk, \tau^*, m^*, \sigma^*) = 1$ and satisfies one of the following conditions (the type 2 forgery can be split into 2 subtypes):

Type 1: $\tau^* \neq \tau_x$ for all x queried by attacker (a type 1 forgery).

Type 2: $\tau^* = \tau_x$ for one pair of x, and $m^* \neq \sum_{i=1}^{n} m_i$, where $m_i \in V_x$ (a type 2 forgery).

Type 2(a): the first element σ_1^* of signature (σ_1^*, σ_2^*) output by the attacker is *not* equal to the signature $(\sigma_1, \sigma_2) \leftarrow \text{Sign}(sk, \tau_x, m)$ computed by the challenger.

Type 2(b): the first element σ_1^* of signature (σ_1^*, σ_2^*) output by the attacker equals the signature $(\sigma_1, \sigma_2) \leftarrow \text{Sign}(sk, \tau_x, m)$ computed by the challenger.

The advantage of the attacker is the probability that the attacker wins the game.

We can show that type 1 and type 2(a) forgery in our linearly homomorphic signature scheme will lead to a forgery of the underlying Boneh and Boyen (BB) signature.

Theorem A.4. *Our linearly homomorphic signature scheme is secure against type 1 and type 2(a) forgeries, if BB signature is strong unforgeable against a weak attacker.*

Proof.
Sketch. The challenger simulates the public key of our scheme by using the public key of BB signature and the element $h = g^\delta$. For responding the signature query on m in our scheme, the challenger queries τ to the challenger of BB signature and obtains σ_1. Then, the challenger returns $(\sigma_1, \sigma_2 = \sigma_1^{m+\delta \cdot r})$ for a random number $r \in Z_N^*$. Finally, if the attacker of our scheme outputs a valid forgery (τ^*, m^*, σ^*), then the first component of σ^* is a valid forgery of BB signature. □

Theorem A.5. *Our linearly homomorphic signature scheme is secure against type 2(b) forgeries, if q-SDH assumption holds.*

Proof.
Sketch. The challenger of our scheme takes as input an instance $(g, g^\beta, g^{\beta^2}, \ldots, g^{\beta^q})$ of the q-SDH assumption and forms the polynomial $P(x) = \prod_{i=1}^{q}(x + \tau_i) \in Z_N[t]$ for the q distinct tags $\tau_1, \ldots, \tau_q$ queried by the attacker. Let $P_l(x) = \prod_{i \neq l}(x + \tau_i)$ and $l^* \in \{1, \ldots, q\}$ randomly chosen by the challenger. Then, the challenger constructs $X = g^{P(\alpha)}$, $Y = g^{P_{l^*}(\alpha)}$, and $h = g^\delta$, which can be used to respond to the signature queries from the attacker. Finally, when the attacker returns the forged signature $\sigma^* = (\sigma_1^*, \sigma_2^*)$ on m^* and τ^*, the challenger computes $z = \sigma_2^* / X^{(m^* + \delta \cdot r^*) \cdot 1/(\alpha + \tau^*)}$. If the forged signature is valid, then $z = Y^{1/(\alpha+\tau^*)}$.

Let $P_{l^*}(t)/(t + \tau^*) = Q(t) + c/(t + \tau^*)$, where $Q(t)$ is a polynomial over Z_N^*. Thus, $z = g^{P^*(\alpha)/(\alpha+\tau^*)} = g^{b \cdot (Q(\alpha) + c/(\alpha+\tau^*))}$. Then, $(\tau^*, (z/g^{Q(\alpha)})^{1/c})$ is a solution to the q-SDH assumption. □

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] F. Li, W. Qiao, H. Sun et al., "Smart transmission grid: vision and framework," in *Proceedings of IEEE Transactions on Smart Grid*, vol. 1, pp. 168–177, 2010.

[2] D. Niyato, L. Xiao, and P. Wang, "Machine-to-machine communications for home energy management system in smart grid," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 53–59, 2011.

[3] Z. M. Fadlullah, M. M. Fouda, N. Kato, A. Takeuchi, N. Iwasaki, and Y. Nozaki, "Toward intelligent machine-to-machine communications in smart grid," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 60–65, 2011.

[4] H. Liang, B. J. Choi, W. Zhuang, and X. Shen, "Towards optimal energy store-carry-and-deliver for PHEVs via V2G

system," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM '12)*, pp. 1674–1682, Orlando, Fla, USA, March 2012.

[5] K. Sha, N. Alatrash, and Z. Wang, "A secure and efficient framework to read isolated smart grid devices," *IEEE Transactions on Smart Grid*, 2016.

[6] I. E. I., "Utility-scale smart meter deployments: building block of the evolving power grid," IEI Smart Meter Update, sep 2014.

[7] F. D. Garcia and B. Jacobs, "Privacy-friendly energy-metering via homomorphic encryption," in *Proceedings of 6th International conference Security and Trust Management*, vol. 6710, pp. 226–238, Springer, Berlin, Germany, 2011.

[8] F. Li, B. Luo, and P. Liu, "Secure information aggregation for smart grids using homomorphic encryption," in *Proceedings of the 1st IEEE International Conference on Smart Grid Communications (Smart Grid Comm '10)*, pp. 327–332, IEEE, Gaithersburg, Md, USA, October 2010.

[9] R. Lu, X. Liang, X. Li, X. Lin, and X. Shen, "EPPA: an efficient and privacy-preserving aggregation scheme for secure smart grid communications," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 9, pp. 1621–1631, 2012.

[10] C.-I. Fan, S.-Y. Huang, and Y.-L. Lai, "Privacy-enhanced data aggregation scheme against internal attackers in smart grid," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 1, pp. 666–675, 2014.

[11] Z. Fu, K. Ren, J. Shu, X. Sun, and F. Huang, "Enabling personalized search over encrypted outsourced data with efficiency improvement," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 9, pp. 2546–2559, 2015.

[12] T. Ma, J. Zhou, M. Tang et al., "Social network and tag sources based augmenting collaborativerecommender system," *IEICE Transactions on Information and Systems*, vol. E98-D, no. 4, pp. 902–910, 2015.

[13] D. Mandell Freeman, "Improved security for linearly homomorphic signatures: a generic framework," in *Public Key Cryptography – PKC 2012*, pp. 697–714, Springer, Heidelberg, Berlin, Germany, 2012.

[14] Z. Wang, G. Sun, and D. Chen, "A new definition of homomorphic signature for identity management in mobile cloud computing," *Journal of Computer and System Sciences*, vol. 80, no. 3, pp. 546–553, 2014.

[15] Z. Brakerski and S. Goldwasser, "Circular and leakage resilient public-key encryption under subgroup indistinguishability," in *Advances in Cryptology – CRYPTO 2010*, Rabin. and T., Eds., vol. 6223, pp. 1–20, Springer, Heidelberg, Berlin, Germany, 2010.

[16] E. Teske, "Computing discrete logarithms in arithmetic progressions," http://citeseerx.ist.psu.edu/.

[17] B. Lynn, "The pairing-based cryptography (pbc) library," http://crypto.stanford.edu/pbc.

[18] D. Boneh and X. Boyen, "Short signatures without random oracles and the SDH assumption in bilinear groups," *Journal of Cryptology*, vol. 21, no. 2, pp. 149–177, 2008.

Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 4943691, 10 pages
https://doi.org/10.1155/2017/4943691

Research Article

Det-WiFi: A Multihop TDMA MAC Implementation for Industrial Deterministic Applications Based on Commodity 802.11 Hardware

Yujun Cheng, Dong Yang, and Huachun Zhou

School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China

Correspondence should be addressed to Dong Yang; dyang@bjtu.edu.cn

Received 13 December 2016; Revised 4 March 2017; Accepted 28 March 2017; Published 16 April 2017

Academic Editor: Feng Wang

Wireless control system for industrial automation has been gaining increasing popularity in recent years thanks to their ease of deployment and the low cost of their components. However, traditional low sample rate industrial wireless sensor networks cannot support high-speed application, while high-speed IEEE 802.11 networks are not designed for real-time application and not able to provide deterministic feature. Thus, in this paper, we propose Det-WiFi, a real-time TDMA MAC implementation for high-speed multihop industrial application. It is able to support high-speed applications and provide deterministic network features since it combines the advantages of high-speed IEEE802.11 physical layer and a software Time Division Multiple Access (TDMA) based MAC layer. We implement Det-WiFi on commercial off-the-shelf hardware and compare the deterministic performance between 802.11s and Det-WiFi under the real industrial environment, which is full of field devices and industrial equipment. We changed the hop number and the packet payload size in each experiment, and all of the results show that Det-WiFi has better deterministic performance.

1. Introduction

In recent years, wireless communication has gained significant importance in industrial automation. A great amount of industrial applications adopts wireless network control systems [1–5] thanks to their ease of deployment and the low cost of their components compared with the wired control systems. Most of industrial applications require assured maximum end-to-end delivery latency and reliable transmission, namely, the deterministic feature of the system. Besides, the sampling rate of different applications is varied and some applications require the network to provide extremely high transfer speed. Thus, both of the transmission rate and deterministic feature of the network should be considered when designing a control system.

Several standards are dedicated for manufacturing automation and process automation, such as WirelessHART [6], ISA100.11a [7], and WIA-PA [8]. They are based on the IEEE 802.15.4-2006 [9] standard and have already been applied. Their main characteristic is the use of the TDMA based MAC protocol to enable more reliable and real-time communication. However, the transfer rate defined in IEEE 802.15.4 is up to 250 kbps, which cannot provide high enough sampling rate for high-speed application. On the other hand, IEEE 802.11 [10] is able to support high-speed communication (up to 150 Mbps in 802.11n). Besides, the IEEE 802.11s [11] amendment, which aims to provide support for multihop networks including both static topologies and ad hoc networks, seems hopeful to be applied in multihop industrial application. However, it still adopts carrier sense multiple access with collision avoidance (CSMA/CA) based MAC layer, which does not provide any time delay guarantee on data delivery and is not feasible for most of industrial automation application.

To address this problem, a huge amount of literature has been proposed. The proposed schemes can be classified into two groups. The first group aims to optimize the conventional 802.11 MAC layer [12–15]. Some default

parameters in 802.11 MAC, such as contention window and backoff mechanism, are optimized to meet the industrial application demand. However, the uncertain delay brought by CSMA/CA still cannot be avoided. The second group substitutes the TDMA MAC layer for the conventional 802.11 MAC. Without the CSMA/CA mechanism, a better deterministic performance can be achieved. This paper belongs to this group, and, thus, this group of solutions will be reviewed below.

Many TDMA based MAC protocols have been previously proposed based on commodity 802.11 hardware. SoftMAC [16] initiates an implementation to create a precise control over the wireless transmission and reception. Based on Atheros MadWifi driver, the author proposes several properties of commodity 802.11 hardware that should be disabled to make it a flexible platform. Soft-TDMAC [17] is a multihop TDMA protocol which is said to realize microsecond synchronization. The real-time performance of Soft-TDMAC is great in testbed result due to the centralized network structure and TDMA based MAC layer. However, it is not designed for control applications. RT-WiFi [18] aims to provide high sampling rate and deterministic timing guarantee on packet delivery in wireless control systems. It is also based on a TDMA data link layer combined with IEEE 802.11 physical layer. But it can only deploy in completely connected networks; that is, it cannot support any kind of multihop application.

In this paper, we provide a detailed presentation of the Det-WiFi, which is able to provide support for high-speed multihop industrial deterministic application. It adopts a centralized network architecture and uses time division multiplexing access strategy to provide end-to-end delay guarantee of data. The main contributions of this paper can be summarized as follows:

(1) We designed Det-WiFi, a new MAC protocol for high-speed multihop industrial deterministic application. The detailed MAC features are considered to make sure the system can be used in practice. Compared with the existed protocols, Det-WiFi is capable of enabling real-time and high-speed communication simultaneously.

(2) We implemented Det-WiFi based on commodity 802.11 hardware, which is ubiquitous and relatively cheap. With several driver modifications, Det-WiFi is both easy and affordable to apply to an industrial control system.

(3) We set up a testbed under the real industrial environment to validate the performance of Det-WiFi. We tested and analyzed both Det-WiFi and 802.11s protocol in detail. All of the results show that Det-WiFi has better deterministic performance compared with 802.11s.

The rest of the paper is organized as follows. Section 2 presents the system design of Det-WiFi. Section 3 describes the Det-WiFi implementation details. Section 4 explains the testbed configuration and experiment results. Section 5 concludes the paper.

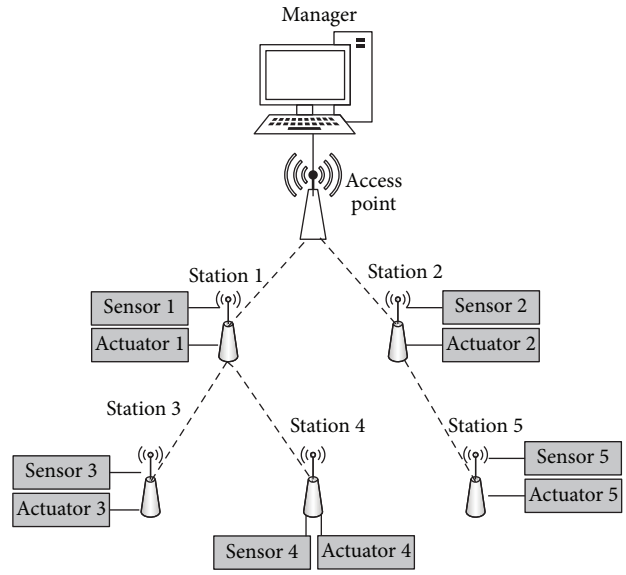


Figure 1: Det-WiFi network structure.

2. Det-WiFi Protocol Design

Det-WiFi is a wireless multihop protocol, which aims to provide high-speed, real-time, and reliable communication for a wide variety of industrial real-time applications. In this section, we present the design details of Det-WiFi protocol.

Figure 1 shows a typical Det-WiFi network structure for industrial real-time control system. The network structure is composed of four parts: manager, access point (AP), stations, and actuators and sensors. The manager is the brain of the network, and it is located at the top of the structure. When a station joins the network, manager receives information from the station, including station address, parent station address, and receive signal strength indicator (RSSI). Thus, the manager holds a lot of information of the stations and has the capability to control the entire network using the information. The AP is a bridge between the manager and stations. It is responsible for delivering all of the uplink or downlink messages to manager or stations. The AP and the stations form a multihop network; each station is attached with one actuator and one sensor. Stations control the actuators according to the management information from AP and monitor the data from sensors. Since the sensors generate a big amount of data and almost all of the traffic in the network requires deterministic guarantee, it is challenging to design a scheme to fulfill such stringent requirements, especially for a multihop network solution.

To address this problem, Det-WiFi adopts TDMA scheme instead of distributed coordination function (DCF) in regular Wi-Fi to ensure the reliability and real-time capability of the network. DCF is a mechanism based on carrier sense multiple access with collision avoidance (CSMA/CA). It is a general media access method, but it does not seem practicable for stringent real-time communication due to the undetermined time delay introduced by channel contention and backoff mechanism. Hence, DCF is unable to provide end-to-end

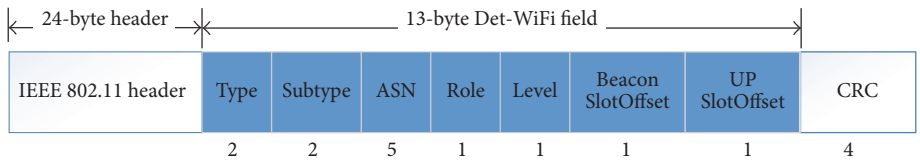


Figure 2: The beacon frame format.

delay guarantee and is not appropriate for deterministic communication. In TDMA scheme, in contrast, communication resource is divided into time-slots and assigned to stations. Due to its contention-free and predictable time delay, it is feasible for deterministic communication. Moreover, Det-WiFi employs centralized network architecture and manager is responsible for controlling the entire network, such as time-slot allocation and scheduling, sending routing information to stations. With the control of the manager and the TDMA scheme, Det-WiFi is feasible for the industrial real-time applications.

2.1. Network Joining Process. The network joining process of Det-WiFi consists of two steps: AP joining process and stations joining process.

Det-WiFi begins to work when the manager boots up. At the start, if AP detects the manager online, it will prepare for the network joining process. It sends the *JOIN_REQUEST* frame to the manager, which contains join information such as its MAC address and hop number. These pieces of information is stored locally by the manager after it receives the *JOIN_REQUEST* frame from AP. After the join information is completely stored, a frame called *JOIN_REPLY* is sent back to AP from manager. *JOIN_REPLY* frame contains control information and resource allocation information. The most important information is time-slot allocation information, including beacon time-slot offset and uplink time-slot offset. Manager utilizes the values of time-slot offsets to allocate beacon and uplink time-slot to AP, respectively. According to these pieces of information, AP knows which time-slot it should occupy and adjusts itself to the appropriate transmission state. At last, AP sends a frame back called *JOIN_ACK* to make the manager confirm its joining status. If the manager receives *JOIN_ACK* frame correctly, AP is considered in joining state, and the three-way handshake joining process is completed.

After the AP joins the network, the stations should prepare for joining process. At this time, AP is going to broadcast the beacon frames (Figure 2) periodically, which contains ASN (absolute slot number), beacon slot offset, role, level, and up slot offset. ASN is used to record the global slot number; level is the maximum child stations that AP can control; role shows the role of the beacon sender (AP or station); beacon slot offset shows which beacon slot the AP occupies; and up slot offset shows which up slot is allocated to AP. If any station hears the beacon, it is going to choose AP as its neighbor and parent according to the beacon frame. Actually, one station may hear several beacons from different neighbors, so it will select one as its parent. After parent selection is finished, it could send *JOIN_REQUEST* frame to the manager through the AP for joining the network. The joining procedure of station is similar to the procedure of AP; the only difference is that it sends frames to the manager through the AP rather than by itself directly. It is worth mentioning that AP plays the role of bridge between manager and stations in station joining process. It just forwards the joining frames as well as the time-slot allocation information. AP does not generate packets or do any scheduling stuff. Once this station joins the network, it broadcasts its own beacon frames periodically. Other stations which hear the beacons follow similar steps to join the network until all the stations join the network. Since then, the deployment of Det-WiFi is completed and it is ready to collect sensor data and control the actuators.

2.2. Frame and Time-Slot Design. The basic frame structure is comprised of an 802.11 header and the Det-WiFi field. The IEEE 802.11 header is used to control basic 802.11 network features, and the function of Det-WiFi network is controlled by Det-WiFi field actually. There are three main frame types in Det-WiFi, including beacon frame, data frame, and management frame. The frames are distinguished by type and subtype field.

To avoid collision, data transmission procedure in TDMA is divided into time-slots. Only one station can access the channel at the same time, and each station only accesses the channel in the time-slot which is allocated for it. In Det-WiFi, time-slot allocation for one station is completed when the station receives the *JOIN_REPLY* from manager (forwarded by AP) in joining procedure (Section 2.1), and the time-slot information is recorded locally in the time-slot table of the station. Because our goal is to design and implement a basic architecture of a deterministic wireless network, a detailed discussion of the various methods of allocating time-slot and scheduling time-slot are beyond the scope of this paper.

In Det-WiFi, A superframe is formed by an infinite cycling sequence of all the allocated time-slots. Time-slots are circulated as superframes go by. Generally, there are three main types of time-slots in a superframe: beacon slot, transmitting (Tx) slot, and receiving (Rx) slot. AP and stations broadcast beacons in beacon slot, and data is transmitted and received in transmitting and receiving slot, respectively. The structure of the three main types of time-slots are shown in Figure 3.

Beacon slot is used to broadcast beacon frames; thus no acknowledgement (ACK) is required; other frames are sent in the Tx slot and need ACK to confirm whether the frame transmits to the target successfully; after a station has received a data frame in Rx slot, it should send an ACK back to confirm its receiving state. Because of the difference among the three

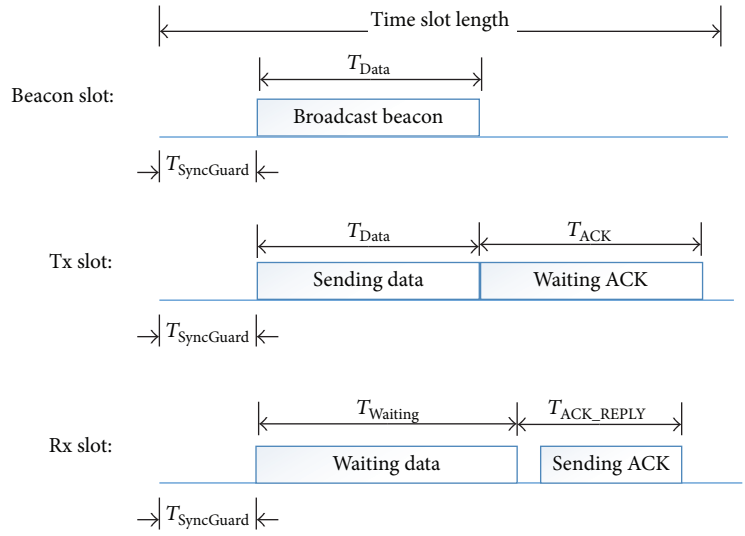


Figure 3: Structure of three main slot types.

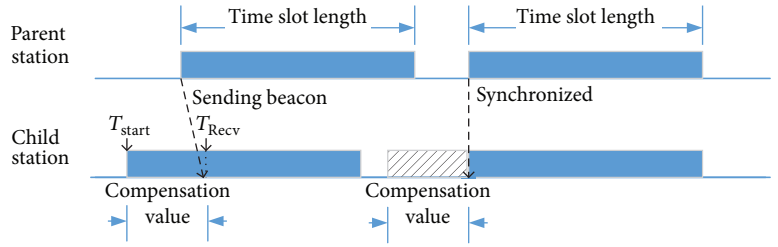


Figure 4: Time-slot synchronization procedure.

types of slots, the designs of these time-slots are not the same. The slot sizes of three types of slots can be simply expressed (corresponding to Figure 3) as follows:

$$T_{BeaconSize} \geq T_{SyncGuard} + T_{Data}, \quad (1)$$

$$T_{TxSize} \geq T_{SyncGurad} + T_{Data} + T_{ACK}, \quad (2)$$

$$T_{RxSize} \geq T_{SyncGurad} + T_{Waiting} + T_{ACK\_REP}, \quad (3)$$

where $T_{SyncGurad}$ is an interval to make the station be able to tolerate slightly synchronization inaccuracy. T_{Data} is the delay in data transmission. It is caused by both Linux kernel delay and radio transmission delay. T_{ACK} is the ACK transmission time, and it is similar to T_{Data}. $T_{Waiting}$ is the time in which receiver side waits for the frame incoming. The measured values of these parameters are shown in Section 3.3.

2.3. Synchronization. Accurate synchronization is the foundation of TDMA mechanism. In Det-WiFi network time system, the local clock of AP is used as the network reference clock. Every station should keep synchronization directly or indirectly with the AP. The process of synchronization is comprised of two parts: ASN synchronization and time-slot synchronization.

As we mentioned in Section 2.1, ASN is used to record the global slot number. Every time-slot passes by, ASN adds one, and it keeps increasing since the network boots up. It is contained in every beacon frame. When a parent station broadcasts its beacons, the children stations read the ASN information and synchronize its local ASN with the parent station. It is important because synchronized ASN ensures the AP and all the stations are in the same time-slot state. Besides ASN synchronization, time-slot synchronization is also finished when the children stations receive beacons from their parent (Figure 4). When children stations receive beacons, receiving time T_{recv} is recorded. Besides, every start time of time-slot T_{start} is recorded by stations. If transmission delay is ignored, calibration value for children stations is

$$T_{calib} = T_{recv} - T_{start} + T_c, \quad (4)$$

where T_c is the compensation for the Linux kernel delay, and, luckily, the delay is quite stable. In our test, this delay is 10 μs $\pm$ 2. Thus, when next slot is coming, children stations can synchronize with the parent station using a changed slot size:

$$T_{next} = T_{timeslot} + T_{calib}. \quad (5)$$

After that, children stations should repeat the time-slot procedure synchronization constantly to maintain the synchronization state. Instead of beacon frames, any frame exchanged with parent station, no matter data frame or

beacon frame, can be used for synchronization, based on the same synchronization mechanism.

In order to reduce the impact of synchronization error, in Det-WiFi, synchronization guard time (Section 2.2) is set at the start of one time-slot. The guard time can improve the error tolerance in the synchronization procedure. However, there is another issue called variation accumulation, which means the synchronization variation can accumulate across the multihop network. To address this problem, Det-WiFi limits the hop number. The maximum hop number which Det-WiFi is able to support is:

$$N_{\max} = \frac{T_{\text{SyncGuard}}}{2 \times T_{\text{SyncVar}}}, \tag{6}$$

where $N_{\max}$ is the maximum hop number, $T_{\text{SyncGuard}}$ is the guard time which is set to 100 μs in Det-WiFi, and T_{SyncVar} is the synchronization variation which is 2 μs. Thus, we can figure out that the supported maximum hop number is 25 hops. In practice, the hop number can be set to 10 to avoid any unknown synchronization variation, which is still enough for many industrial applications.

3. Implementation

In this section, we describe the implementation details of Det-WiFi from three parts: system architecture, driver modification, and timers.

3.1. System Architecture. As we mentioned in Section 3, the network structure of Det-WiFi is comprised of manager, AP, stations, and attached sensors and actuators. Sensors and actuators are varied in different industrial applications, and, therefore, we only reserve an interface for them and use a data generator to emulate their work process in testbed experiments. Because we focus on the deterministic feature of Det-WiFi, we only simply implement basic function of the manager. It just stores the information from stations and distributes a fixed time-slot table to stations after stations join. AP and stations adopt the same hardware and network stack, and they just run in different work mode. In the following paragraph, we will focus on the implementation of stations and AP.

The system architecture of station (or AP) is shown in Figure 5. To achieve deterministic performance, some default features of hardware need to be modified. Thus, in our system, we use the AR9287 chipset-based commodity 802.11 b/g/n hardware. It uses ath9k hardware driver module on Linux system, which is open source and easy to modify. On the top of ath9k, mac80211 is a framework that provides standard IEEE802.11 MAC related functionality. We use Ubuntu 14.04 as the operating system, and the kernel version is 3.14.57. Considering the compatibility issues, we just modified the two modules slightly. We will talk about the detailed modification steps in Section 3.2.

Based on the two kernel modules, Det-WiFi is comprised of three components: packet queues, task scheduler, and system state container (SSC). There are two packet queues in Det-WiFi: sending queue (Tx queue) and receiving queue

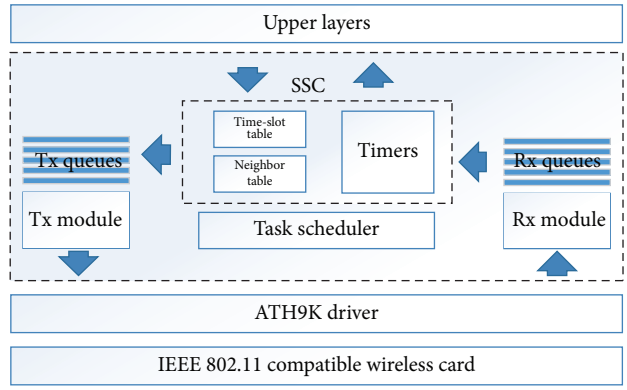


Figure 5: Det-WiFi system architecture.

(Rx queue). When packets have been prepared to send, they are put in the sending queue, waiting for the appropriate time-slot, and then are sent to the driver to transmit. Similarly, when packets are received from lower layer, they are stored in the receiving queue. Task scheduler is used to schedule the tasks and control the behavior of the Det-WiFi, including sending beacons, slots cycling, and network joining. These tasks are distinguished by priority level according to the urgency of the tasks, and the task scheduler will execute high-priority task first rather than the low-priority task. SSC consists of time-slot table, neighbor table, and timers: time-slot table records the time-slot cycling sequence, which is acquired from the manager when it joins the network; neighbor table is used to store the neighbors information, which is advertised in the neighbors beacons; timers are responsible for maintaining the time information of Det-WiFi; and most of the tasks are triggered by several timers. The detailed information of timers will be shown in Section 3.3.

Due to the compatible MAC design, the original network upper layers are well-supported. Many applications based on standard UDP/TCP can work without any problem. Moreover, some of real-time multihop applications which cannot be deployed before are able to deploy now because of the deterministic network features.

3.2. Driver Modification. In order to realize multihop deterministic characteristics, some default features of driver should be modified. Luckily, ath9k is an open source driver, and many MAC features of network interface card are able to be changed. We mainly focus on two parts: frame format modification and disabling CSMA mechanism.

It is hard to reuse 802.11 frame format directly due to the topology and structure difference between Det-WiFi and 802.11. Thus, most of the default control and management frames of IEEE 802.11 are abandoned, and only data frame is remained. The control and management information is contained in the payload (Det-WiFi field) of the default 802.11 data frame. However, to send the modified frame properly and successfully, several places of the IEEE 802.11 header need to be changed: the frame should be claimed broadcast, and then the sending process will not be bothered by the default MAC address; the fragment bit is set to zero to tell the receiver

it is not a fragment frame; every source address field is filled with self-defined address instead of the MAC address of NIC. These modifications of frame format bring much convenience to the multihop communication and ensure the frame still can be recognized by the driver.

Several CSMA mechanisms have negative influence on deterministic feature, including virtual carrier sense (NAV) PHY Clear Channel Assessment (CCA) and transmission backoff. Thus, these mechanisms should be disabled. The AR9287 chipset provides a diagnostic register which has several special functions. We set *AR_DIAG_IGNORE_VIRT_CS* and *AR_DIAG_FORCE_RX_CLEAR* flags to disable the NAV and force a CCA, respectively. In order to disable the transmission backoff, the contention windows CWmin and CWmax are set to zero in the initializing process of driver queues. It is worth mentioning that there are four queues in the driver which are known as *AC_BE*, *AC_BK*, *AC_VI*, and *AC_VO*. These correspond to the four priorities of traffic in enhanced distributed channel access (EDCA) which is proposed in IEEE 802.11e [19] amendment. To ensure all frames transmit in order, only *AC_VO* queue is retained, and we map all frames to that queue. As for the transmission rate, the default minstrel rate control algorithm is disabled, and we adopt a fixed transmission rate 54 Mbps. However, when a fixed transmission rate is used, the driver will send the frame at the lowest rate (1 Mbps in 802.11g). The reason is that the driver will modify the transmission rate automatically when the frame is claimed broadcast or has no ACK policy. After modified, the NIC is able to send frame at the fixed rate of 54 Mbps.

3.3. Timers. In order to realize microsecond precision timing function, timers based on Linux kernel jiffies cannot be employed, as they only provide millisecond granularity. Instead, we adopt high resolution timer (hrtimer) [20, 21] for timing tasks. Hrtimer is able to provide submicrosecond timing precision, and it can fully satisfy the timing demand.

There are several timers based on hrtimers in Det-WiFi in charge of triggering several tasks. The most important timer called mtimer is to keep time-slot generating. When the mtimer is triggered, a new time-slot starts. One Tx slot is comprised of transmission procedure and receiving ACK procedure (discussed in Section 2.2). Thus the transmission delay

$$T_{\text{Data}} = T_{\text{RadioDelay}} + T_{\text{KernelDelay}}, \tag{7}$$

where $T_{\text{RadioDelay}}$ is the radio transmission delay and $T_{\text{KernelDelay}}$ is the Linux kernel delay. If the data is sent at the rate of 54 Mbps and the frame length is 524 bytes (500 bytes for the payload and 24 bytes for the 802.11 header), the radio transmission delay is

$$T_{\text{RadioDelay}} = \frac{(500 + 24) \times 8}{54 \times 10^6} = 77.6\,\mu\text{s}. \tag{8}$$

However, we measured the total transmission delay from the time when a frame is just handed to driver to the time when the frame is sent out successfully; the average delay is 236 μs. From (7), we can figure out that the average Linux kernel delay is 158.4 μs. Then we changed the frame length to 224 bytes and 34 bytes, and the average kernel delay is almost unchanged. We measured that the ACK delay is 167 μs in the same way, which is consistent with expected values. If we set the synchronization guard time to 150 μs, we can calculate the time-slot size from (2):

$$T_{\text{TxSize}} \geq 150\,\mu\text{s} + 236\,\mu\text{s} + 167\,\mu\text{s} = 553\,\mu\text{s}. \tag{9}$$

The size of beacon slot and Rx slot is smaller; thus (9) can be used to determine the system slot size and the mtimer timing value.

Figure 6: Testbed in real industrial environment.

4. Performance Evaluation

To evaluate the deterministic multihop features of Det-WiFi, we set up a testbed under the real industrial environment (Figure 6), which is full of field devices and industrial equipment. We compare the performance between Det-WiFi and 802.11s in this scenario. Open80211s [22] is an open-source implementation of the ratified IEEE 802.11s wireless mesh standard. Moreover, it is well-supported in the ath9k driver. Therefore, we intend to test the performance of 802.11s based on Open80211s. There are three PCs in our experiment, all of them are equipped Atheros AR9287 IEEE 802.11 compatible NIC. Every device installed a UDP packets generator program, and it generates packets at fixed intervals (5 ms) to emulate the sampling process. The frequency of NIC is set to 2.462 Ghz corresponding channel 11 of IEEE 802.11 b/g/n. We set the time-slot length 600 μs, which is longer than 553 μs to reserve enough time for time-slot internal guard, and the time-slot table is fixed for the experiment.

In our experiment, we focus on the transmission delay of MAC layer and packets loss ratio, which show the deterministic performance of the network. Among the delay metrics, the maximum delay and delay standard deviation are relatively appropriate to reflect the deterministic feature of the network. Besides, the mean delay is also a key metric. To measure the packet loss ratio, we set two counters to record the sent out packets and received packets, respectively. Packet loss is calculated as the subtraction of the number of sent-out packets from the number of received packets. To measure the latency of the network, we start a timer when a frame is

Table 1: Comparison of latency and packet loss ratio between Det-WiFi and 802.11s in real industrial environment.

Payload	Scenarios	Max latency (μs)		Mean latency (μs)		Latency standard deviation (μs)		Packet loss ratio	
		Det-WiFi	802.11s	Det-WiFi	802.11s	Det-WiFi	802.11s	Det-WiFi	802.11s
50 bytes	Two-hop	1321	31626	1084	1303	25.08	388.3	0.22%	0%
	Four-hop	2718	16076	2147	2371	55.25	631.1	0.56%	0%
200 bytes	Two-hop	1303	13906	1119	1285	29.87	301.6	0.20%	0%
	Four-hop	2702	53695	2284	2517	40.73	894.5	0.59%	0%
500 bytes	Two-hop	1394	21200	1142	1225	29.13	363.7	0.22%	0%
	Four-hop	2861	27300	2361	2505	50.06	768.9	0.58%	0%

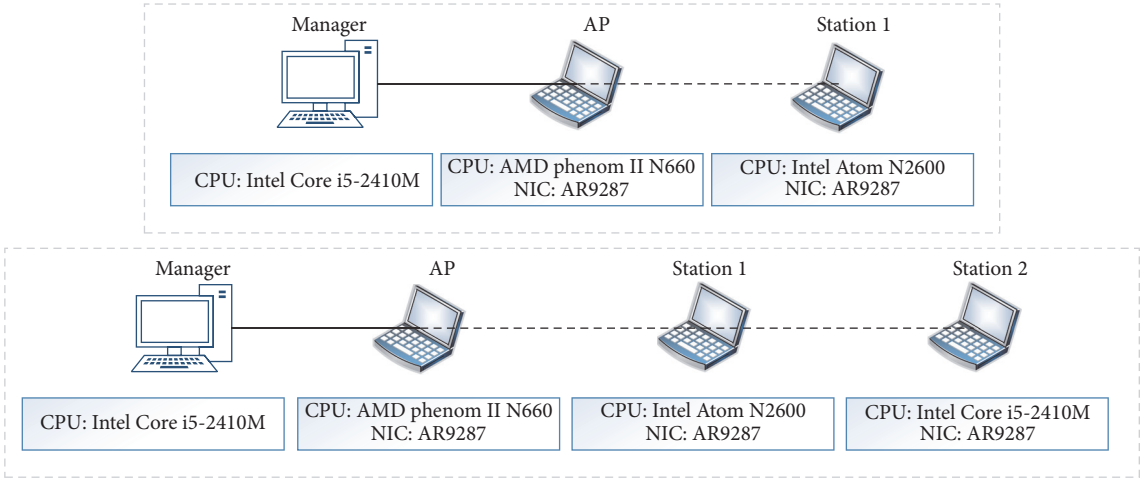
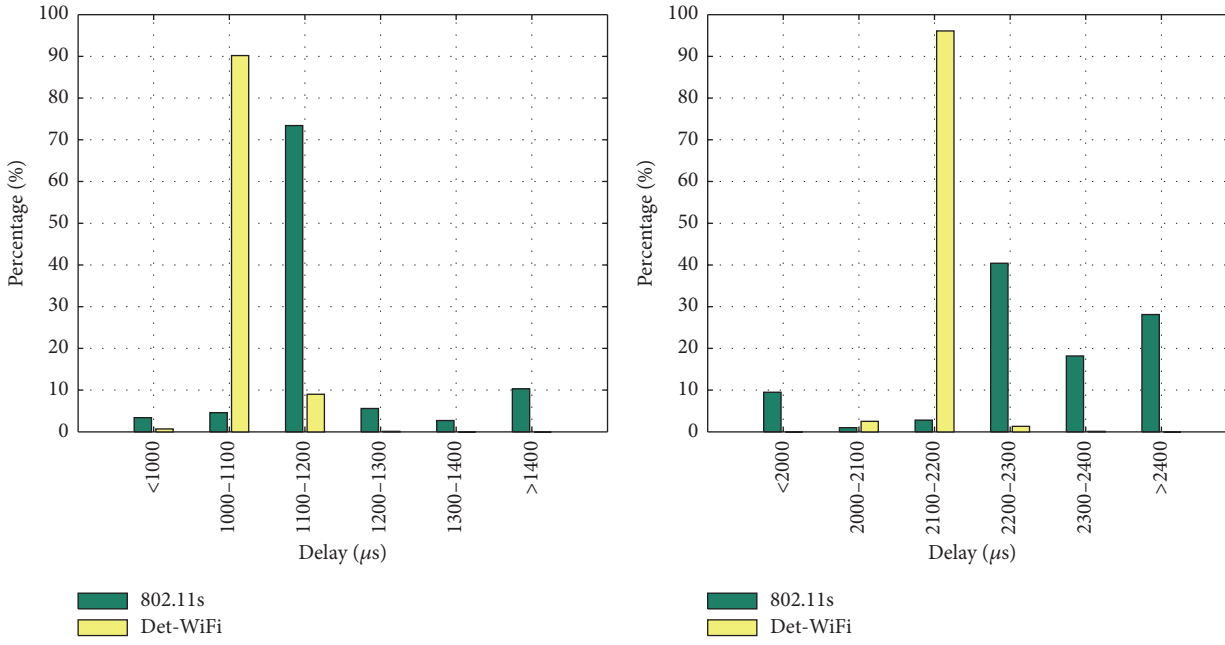


Figure 7: Two test topologies' setting.

ready to transmit at the MAC layer. As the frame arrives at the destination station, it sends back the same length frame at once. At last, we measure the time between the frame start time and the response frame arrival time, both at the MAC layer. Based on this measuring method, we compare the latency and packet loss ratio between Det-WiFi and 802.11s in the scenario of two-hop and four-hop networks, respectively. The two test topologies are shown in Figure 7. In each scenario, we transmit three kinds of data packets in different size (50 B, 200 B, and 500 B payload) to emulate different sampling data. Each experiment runs for 10 minutes. Besides, each experiment is conducted for five times to ensure the reliability of test data, and the result is the average of the five.

The test results are shown in Table 1. We compare the mean, max, standard deviation of MAC latency, and packet loss ratio between 802.11s and Det-WiFi. When adopting different packet size, other metrics in both two-hop and four-hop scenarios are almost unchanged except the mean latency. For the Det-WiFi, the mean latency is slightly increasing with the increase of packet size, because the larger packet needs more time to transmit when the transmission rate is fixed. Since the other metrics are not influenced by the payload size, we take the 500 B payload size condition as an example to validate the deterministic performance of Det-WiFi. The latency standard deviation in 802.11s network is up to 12.5 times to that of Det-WiFi in both scenarios, and the max latency in 802.11s is 15 times and 9.5 times to that of Det-WiFi in two-hop and four-hop experiments, respectively. The high latency standard deviation and max latency in 802.11s network should be attributed to the random backoff mechanism, which makes the time latency uncertain. However, we observe there are 0.22% and 0.58% packet loss in Det-WiFi, respectively. This is due to the slight interference traffic introduced by other 802.11 based devices in the industrial field, and we confirmed it by monitoring the channel using Wireshark. The packet loss is totally acceptable compared with the high latency standard deviation in the 802.11s network. The same conclusion can be drawn from the 50 B and 200 B payload size condition.

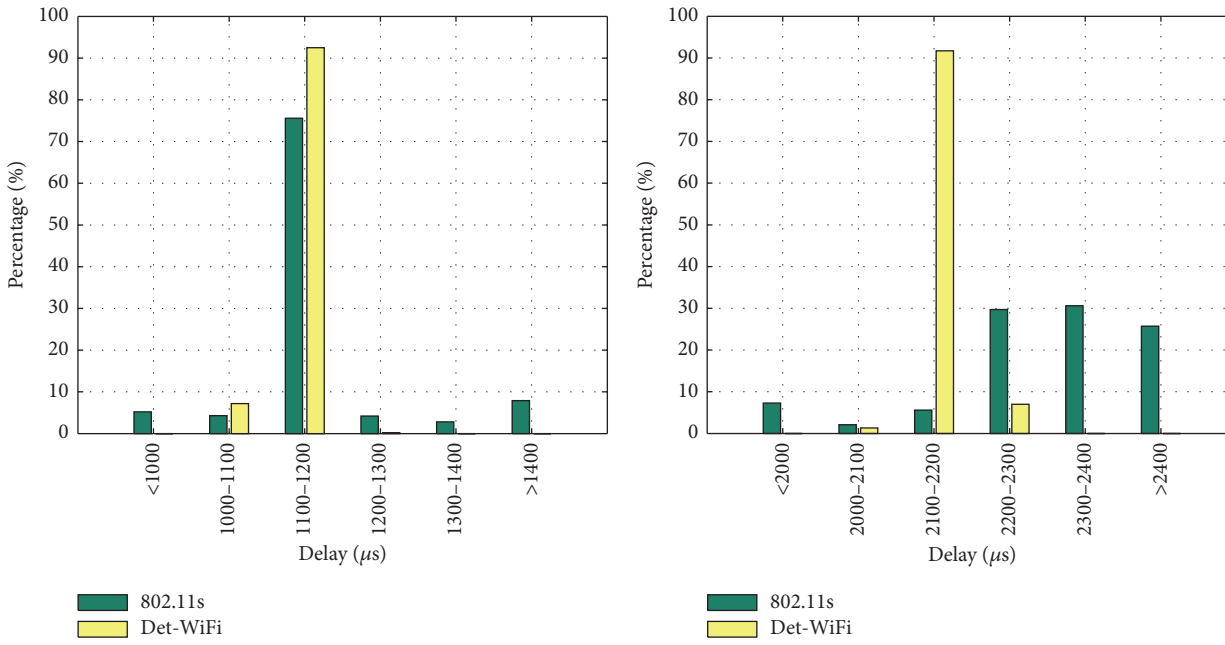
We also plot histograms (Figures 8–10) to illustrate the deterministic performance of the 802.11s and Det-WiFi more intuitively. Considering the packet payload size only has small influence on the mean latency, we still take the 500 B payload size condition (Figure 10) for instance. In two-hop scenario (Figure 10(a)), over 95% of packets are delivered in the interval of 1100–1200 μs and almost 100% of packets arrive in 1200 μs in Det-WiFi. On the contrary, the latency in 802.11s distributes separately. Only 86.2% of packets arrive in 1200 μs and 4.7% of packets are delivered over 1400 μs. In the four-hop scenario (Figure 10(b)), the deterministic performance is



(a) The delay distribution in two-hop experiment

(b) The delay distribution in four-hop experiment

Figure 8: The delay distribution in 50 B payload size condition.



(a) The delay distribution in two-hop experiment

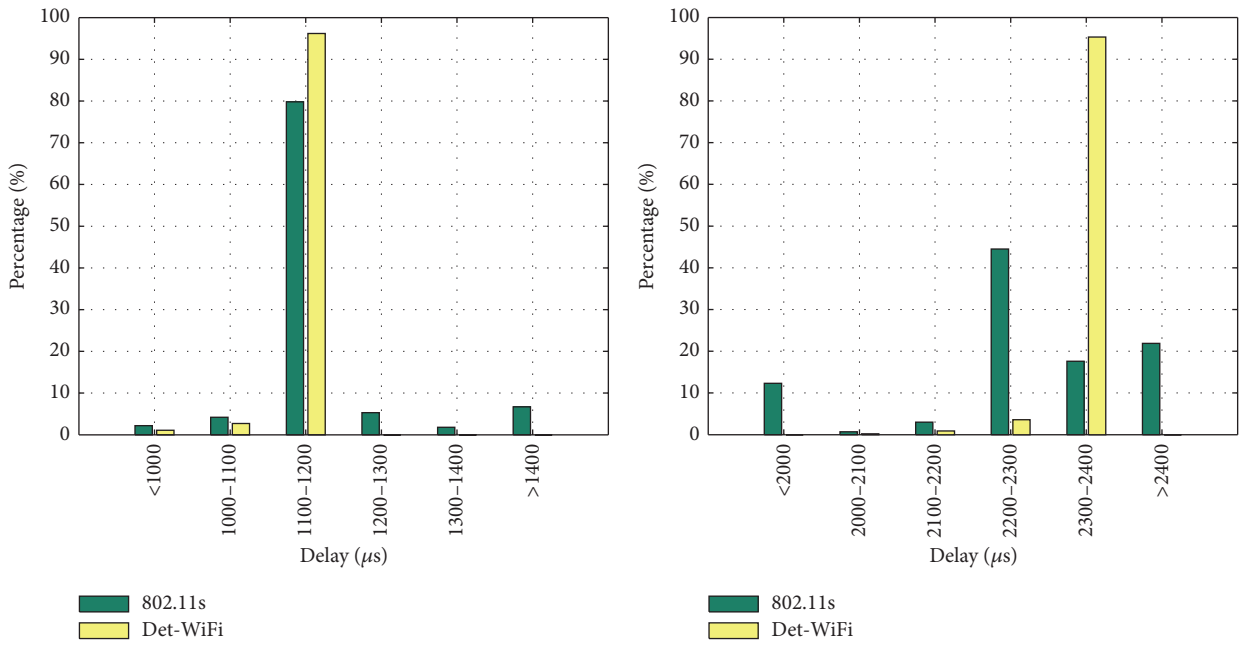
(b) The delay distribution in four-hop experiment

Figure 9: The delay distribution in 200 B payload size condition.

even worse in 802.11s network because more hops bring more channel contention and backoff, while the latency of Det-WiFi is still quite stable. In 802.11s network, more than 20% of packets are delivered over 2400 μs and the latency variance is even greater than the two-hop scenario. In the 50 B and 200 B payload size condition (Figures 8 and 9), the histograms show similar results.

5. Conclusion

In this paper, we propose the Det-WiFi, which aims to provide support for high-speed multihop industrial deterministic demand application. It is based on the physical layer of IEEE 802.11, adopts a centralized network architecture, and uses TDMA strategy instead of CSMA/CA mechanism. Besides,



(a) The delay distribution in two-hop experiment

(b) The delay distribution in four-hop experiment

Figure 10: The delay distribution in 500 B payload size condition.

Det-WiFi is implemented based on the commercial IEEE 802.11 hardware with a few modifications, which makes it have good compatibility. To validate the performance of Det-WiFi, we set up testbed under the real industrial environment and compared the deterministic performance between 802.11s network and Det-WiFi. The time latency and packet loss ratio are chosen to reflect the deterministic feature of the two. In the 500 B payload size condition, the latency standard deviation in 802.11s network is up to 12.5 times that of Det-WiFi, and the max latency in 802.11s is 15 times and 9.5 times that of Det-WiFi in two-hop and four-hop experiments, respectively. We also repeat the experiment in 50 B and 200 B payload size condition; the test results show that Det-WiFi has better deterministic performance compared with the 802.11s network.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Electronic Information Technology Innovation and Cultivation (Grant no. Z171100001217004), Fundamental Research Funds for the Central Universities (Grants nos. 2015JBM006 and 2015YJS011), and National 863 Program of China (Grant no. 2015AA016103).

References

[1] M. Magno, D. Boyle, D. Brunelli, B. O'Flynn, E. Popovici, and L. Benini, "Extended wireless monitoring through intelligent hybrid energy supply," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 4, pp. 1871–1881, 2014.

[2] O. Kreibich, J. Neuzil, and R. Smid, "Quality-based multiple-sensor fusion in an industrial wireless sensor network for MCM," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 9, pp. 4903–4911, 2014.

[3] S. X. Ding, P. Zhang, S. Yin, and E. L. Ding, "An integrated design framework of fault-tolerant wireless networked control systems for industrial automatic control applications," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 1, pp. 462–471, 2013.

[4] M. Chen, "Reconfiguration of sustainable thermoelectric generation using wireless sensor network," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 6, pp. 2776–2783, 2014.

[5] X. Chen, R. Q. Hu, G. Wu, and Q. C. Li, "Tradeoff between energy efficiency and spectral efficiency in a delay constrained wireless system," *Wireless Communications and Mobile Computing*, vol. 15, no. 15, pp. 1945–1956, 2015.

[6] J. Song, S. Han, A. K. Mok et al., "WirelessHART: applying wireless technology in real-time industrial process control," in *Proceedings of the 14th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS '08)*, pp. 377–386, April 2008.

[7] ISA100, http://www.isa.org/isa100.

[8] IEC 62601:2015, https://webstore.iec.ch/publication/23902/.

[9] IEEE 802.15 WPAN Task Group 4, http://www.ieee802.org/15/pub/TG4.html.

[10] "IEEE 802.11 Wireless Local Area Networks," http://www.ieee802.org/11/.

[11] IEEE 802.11-TASK GROUPS, http://www.ieee802.org/11/Reports/tgs update.htm.

[12] F. Tramarin, S. Vitturi, M. Luvisotto, and A. Zanella, "On the use of ieee 802.11n for industrial communications," *IEEE*

Transactions on Industrial Informatics, vol. 12, no. 5, pp. 1877–1886, 2016.

[13] G. Tian, S. Camtepe, and Y.-C. Tian, "A deadline-constrained 802.11 MAC protocol with QoS differentiation for soft real-time control," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 2, pp. 544–554, 2016.

[14] F. Babich, M. Comisso, M. D'Orlando, and A. Dorni, "Deployment of a reliable 802.11e experimental setup for throughput measurements," *Wireless Communications and Mobile Computing*, vol. 12, no. 10, pp. 910–923, 2012.

[15] D. K. Lam, K. Yamaguchi, Y. Shinozaki et al., "A fast industrial WLAN protocol and its MAC implementation for factory communication systems," in *Proceedings of the 20th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA '15)*, September 2015.

[16] M. Neufeld, J. Fifield, C. Doerr, A. Sheth, and D. Grunwald, "Softmac—flexible wireless research platform," in *Proceedings of the 4th Workshop on Hot Topics in Networks HotNets-IV*, 2005.

[17] P. Djukic and P. Mohapatra, "Soft-TDMAC: a software-based 802.11 overlay TDMA MAC with microsecond synchronization," *IEEE Transactions on Mobile Computing*, vol. 11, no. 3, pp. 478–491, 2012.

[18] Y.-H. Wei, Q. Leng, S. Han, A. K. Mok, W. Zhang, and M. Tomizuka, "RT-WiFi: real-time high-speed communication protocol for wireless cyber-physical control applications," in *Proceedings of the IEEE 34th Real-Time Systems Symposium (RTSS '13)*, pp. 140–149, December 2013.

[19] 802.11-1997 - IEEE Standard for Information Technology-Telecommunications and Information Exchange Between Systems-Local and Metropolitan Area Networks-Specific Requirements-Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications.

[20] T. Gleixner and D. Niehaus, "Hrtimers and beyond: transforming the linux time subsystems," in *Proceedings of the Ottawa Linux Symposium (OLS '06)*, Ottawa, Canada, 2006.

[21] W. Torfs and C. Blondia, "TDMA on commercial of-the-shelf hardware: fact and fiction revealed," *AEU—International Journal of Electronics and Communications*, vol. 69, no. 5, pp. 800–813, 2015.

[22] Open80211s project, https://github.com/o11s/open80211s.