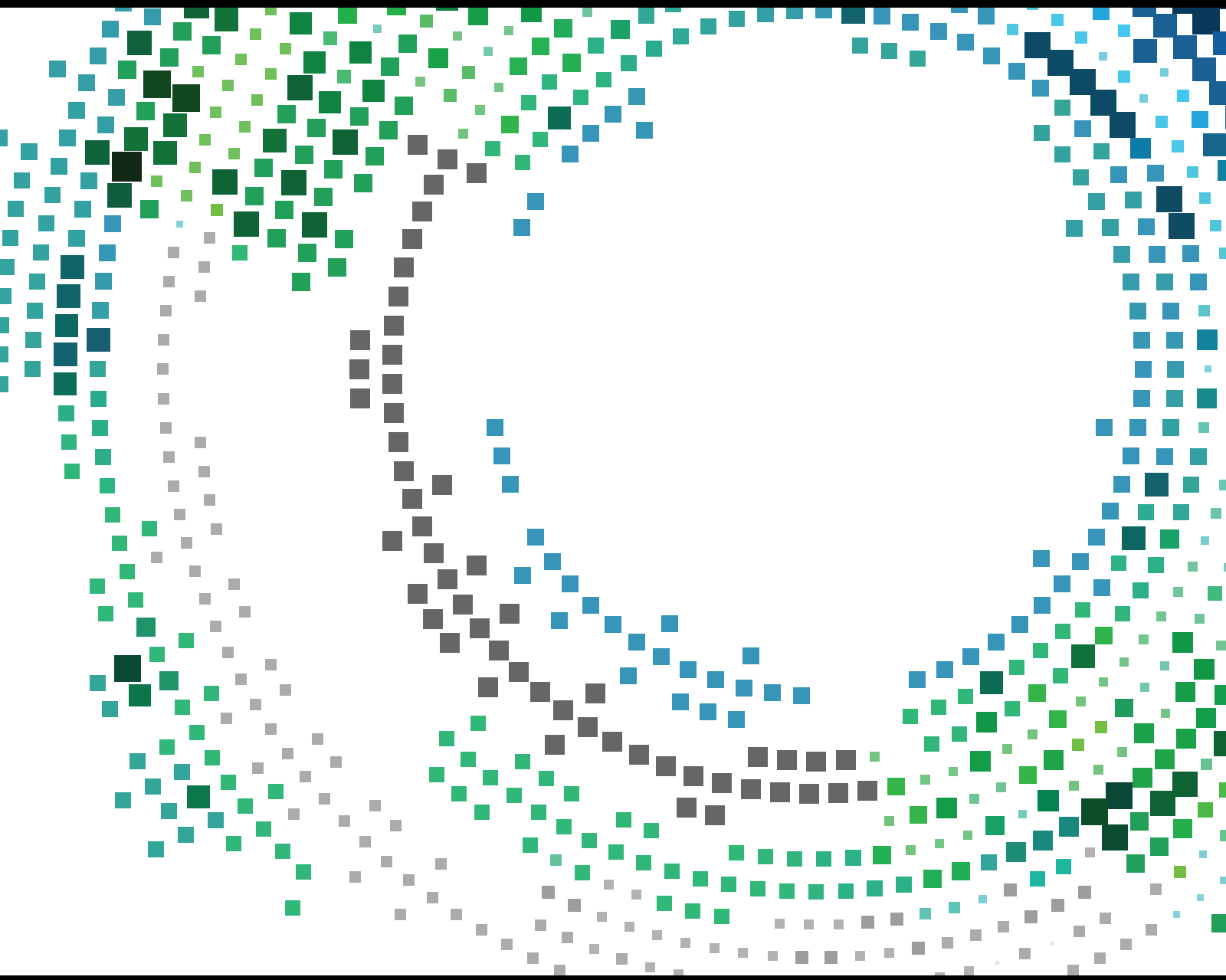
Mobile Information Systems

Design, Dimensioning, and Optimization of 4G/5G Wireless Communication Networks

Guest Editors: Mariusz Głąbowski, Haris Gacanin, Ioannis Moscholios, and Piotr Zwierzykowski



Design, Dimensioning, and Optimization of 4G/5G Wireless Communication Networks

Mobile Information Systems

Design, Dimensioning, and Optimization of 4G/5G Wireless Communication Networks

Guest Editors: Mariusz Głąbowski, Haris Gacanin, Ioannis Moscholios, and Piotr Zwierzykowski

Copyright © 2017 Hindawi Publishing Corporation. All rights reserved.

This is a special issue published in "Mobile Information Systems." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Editorial Board

M. Anastassopoulos, UK
C. Agostino Ardagna, Italy
J. M. Barcelo-Ordinas, Spain
Alessandro Bazzi, Italy
Paolo Bellavista, Italy
Carlos T. Calafate, Spain
María Calderon, Spain
Juan C. Cano, Spain
Salvatore Carta, Italy
Yuh-Shyan Chen, Taiwan
Massimo Condoluci, UK
Antonio de la Oliva, Spain
Jesus Fontecha, Spain
Jorge Garcia Duque, Spain
Michele Garetto, Italy
Romeo Giuliano, Italy
Javier Gozalvez, Spain
Francesco Gringoli, Italy
Peter Jung, Germany
Dik Lun Lee, Hong Kong
Sergio Mascetti, Italy
Elio Masciari, Italy
Maristella Matera, Italy
Franco Mazzenga, Italy
Eduardo Mena, Spain
Massimo Merro, Italy
Jose F. Monserrat, Spain
Francesco Palmieri, Italy
Jose Juan Pazos-Arias, Spain
Vicent Pla, Spain
Daniele Riboni, Italy
Pedro M. Ruiz, Spain
Michele Ruta, Italy
Stefania Sardellitti, Italy
Floriano Scioscia, Italy
Luis J. G. Villalba, Spain
Laurence T. Yang, Canada
Jinglan Zhang, Australia

Contents

Hindawi
Mobile Information Systems
Volume 2017, Article ID 8057275, 2 pages
http://dx.doi.org/10.1155/2017/8057275

Editorial

Design, Dimensioning, and Optimization of 4G/5G Wireless Communication Networks

Mariusz Głąbowski,[1] Haris Gacanin,[2] Ioannis Moscholios,[3] and Piotr Zwierzykowski[1]

[1]*Poznan University of Technology, Poznan, Poland*
[2]*Nokia, Antwerp, Belgium*
[3]*University of Peloponnese, Tripoli, Greece*

Correspondence should be addressed to Mariusz Głąbowski; mariusz.glabowski@put.poznan.pl

Received 22 January 2017; Accepted 23 January 2017; Published 15 March 2017

Copyright © 2017 Mariusz Głąbowski et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Designing, dimensioning, and optimization of communication networks have been an inseparable part of the development of telecommunications and information infrastructure from the very beginning of their existence. These networking problems have changed substantially over the recent years as a result of the changes in the telecommunications area towards 4G/5G wireless multiservice networks and mobile communications, as well as networks convergence. Each and every newly introduced type of a radio network technology is followed by a substantial increase in both the number and the complexity of problems that need to be resolved by theoreticians and engineers.

No matter what these developing changes may bring, the essential tasks for 4G/5G wireless communication networks remain the same: (1) to develop new technologies offering increasing radio network capacity, (2) to determine and evaluate the relationship between the quality of service (quality of experience) parameters and the parameters characterizing traffic sources (services), (3) to control and optimize the usage of radio network resources, and (4) to enhance the capabilities of data transport, transmission, and reception between end users and the core network. These tasks provide a basis for developing engineering algorithms and tools used for designing, analysis, dimensioning, and optimization of wireless systems and networks.

The objective of this special issue, containing 11 papers selected from submissions to the open call for papers, was to bring together the state-of-the-art research contributions that address challenges in 4G/5G radio networks design, dimensioning, and optimization.

In the paper "Energy Efficiency and Capacity Tradeoff in Cloud Radio Access Network of High-Speed Railways," S. Li et al. propose a predictable path loss based time domain power allocation (PPTPA) method to improve energy efficiency performance of High-Speed Railways (HSR) communication system. The authors assumed in the paper that in the future the communication system for HSR will be based on cloud radio access network. The optimization problem formulated in the paper is based on joint energy efficiency and services transmission delay constraints. The effectiveness of the proposed power allocation algorithm was validated in the paper by HSR channel measurement trace based emulation results and extensive simulation results.

The paper "Mobile Network Planning Tool Based on Data Analytics" by J. Moysen et al. describes the tool which is based on the information available in the network and allow properly deploying, configuring, and optimizing network nodes. In the tool two optimization techniques, that is, Machine Learning (ML) and Genetic Algorithms (GAs), are used. In order to evaluate the proposed tools two case studies were discussed: for the small cell deployment in a dense indoor scenario and for a detected fault in a macrocell network. A simulation-based performance analysis of the proposed tool was carried out over the Ns-3 LTE-EPC Network Simulator.

In the paper "Virtualized ANR to Manage Resources for Optimization of Neighbour Cell Lists in 5G Mobile Wireless

Networks" by Y. Shin and S. Kim the network function virtualization (NFV) for Automatic Neighbour Relation (ANR) was proposed. The author assumed that in 5G networks an increase of simultaneous handover and the change of neighbour cell lists will be very quickly and the traditional ANR function has neighbour removal function but it does not consider fast changes in the neighbour list. Therefore the new ANR-virtual network function (ANR-VNF) is proposed in the paper. The experiments for the studied strategy of ANR were executed in practice by using LTE core network (Amarisoft LTE-100 Platform).

In the paper "A Categorized Resource Sharing Mechanism for Device-to-Device Communications in Cellular Networks" by J. Chen et al., the resource sharing mechanism for device-to-device (D2D) communications is studied. In the proposed solution the D2D pairs are divided into three groups based on comparison of the minimum transmit power with the maximum transmit power of each cellular UE. The mechanism studied in the paper enables multiple D2D pairs in the second group to share the resource with cellular user equipment (UE) simultaneously and D2D pairs in the first group and the third group share resource with cellular UEs based on the transmit power minimization principle. Results of simulations conducted by the authors show that the proposed scheme allows obtaining a relatively higher network throughput and a lower transmit power of the D2D system.

The paper by P. Sroka and A. Kliks, entitled "Playing Radio Resource Management Games in Dense Wireless 5G Networks" proposes an efficient and flexible tool for interference mitigation in ultradense heterogeneous cellular 5G networks. Authors study, via simulation, various game-theory based algorithms which concentrate on the optimization of the overall base station energy consumption. Simulation results verify that the adopted game-theoretic algorithms are very promising solutions for interference mitigation outperforming the algorithm proposed for LTE-Advanced in terms of the achieved spectral efficiency.

The paper by S. Hanczewski et al. entitled "Convolution Model of a Queueing System with the cFIFO Service Discipline" proposes an approximate convolution model of a multiservice queueing system that services a mixture of independent multiservice Bernoulli-Poisson-Pascal call streams under the continuous First-In-First-Out (cFIFO) service discipline. The accuracy of the proposed convolution model is verified by simulation experiments for a number of queueing systems.

The paper by F. Alsewaidi et al. entitled "Enhancing Radio Access Network Performance over LTE-A for Machine-to-Machine Communications under Massive Access" proposes three methods to enhance RAN performance for Machine-to-Machine (M2M) communications over the LTE-A standard. The first method employs a different value for the physical Random Access Channel (RACH) configuration index to increase random access opportunities. The second method addresses a heterogeneous network by using a number of picocells to increase resources and offload control traffic from the macro base station. The third method involves aggregation points and addresses their effect on RAN performance. The results presented in the paper confirmed that the proposed methods improved RACH performance in terms of access success probability and average delay.

The paper by J. Säe and J. Lempiäinen, entitled "Maintaining Mobile Network Coverage Availability in Disturbance Scenarios," studies the case of disturbance and disaster scenarios in mobile networks. Authors perform various simulation scenarios under different network layouts with the aim of maintaining the availability of cellular networks in disturbance scenarios. Simulation results show how the mobile network availability duration can be sustained by selecting a set of evolved node B (eNB) sites to operate at a time and still maintain a reasonable service level and availability in disturbance scenarios.

In the paper "Macrocell Protection Interference Alignment in Two-Tier Downlink Heterogeneous Networks" by J. Seo et al., the MCP-IA method to limit interference problems in two-tier (picocells and macrocells) MIMO networks was elaborated. Due to the proposed solution, the additional array and diversity gains, in comparison to the results obtained on the basis of existing interference alignment solutions, can be obtained for the macro users. The performance of the proposed method was evaluated in a simulation environment, at the link level as well as at the system.

The paper "LDPC Decoding on GPU for Mobile Device" by Y. Lu et al. proposed a software multicode word parallel LDPC decoder, exploiting OpenCL based graphics processing units built-in into mobile devices. Software realization of LDPC decoding ensures a possibility of dynamic change of code length, code rate, and the number of iterations, in order to tune to networks' conditions. The tests led in the experimental test-bed confirmed high performance of the proposed solution and confirms a possibility of its application for large file transmission and delay-sensitive services like video calling in mobile networks.

The paper "Latent Clustering Models for Outlier Identification in Telecom Data" by Y. Ouyang et al. deals with the problem of fast and robust identification of unexpected traffic streams in high-speed mobile networks, generated in the case of malicious attacks or technical problems. As the solution to this problem, the authors selected an approach based on clustering models and proposed the application of Gaussian Probabilistic Latent Semantic Analysis (GPLSA) and time-dependent Gaussian Mixture Models (timeGMM). The comparison of the efficiency of the methods was performed in a simulation environment.

Mariusz Głąbowski
Haris Gacanin
Ioannis Moscholios
Piotr Zwierzykowski

Hindawi
Mobile Information Systems
Volume 2017, Article ID 9643401, 9 pages
http://dx.doi.org/10.1155/2017/9643401

Research Article

Virtualized ANR to Manage Resources for Optimization of Neighbour Cell Lists in 5G Mobile Wireless Networks

Yoonsu Shin and Songkuk Kim

School of Integrated Technology, Yonsei Institute of Convergence Technology, Yonsei University, Seoul, Republic of Korea

Correspondence should be addressed to Yoonsu Shin; sinowl@yonsei.ac.kr

Received 26 August 2016; Revised 11 December 2016; Accepted 29 December 2016; Published 8 February 2017

Academic Editor: Piotr Zwierzykowski

Copyright © 2017 Yoonsu Shin and Songkuk Kim. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In future, more devices such as wearable devices will be connected to the networks. This will increase simultaneous handovers. The coverage of a cell will be small because a superhigh frequency used in 5G wireless networks does not propagate very far. This trend will increase the number of neighbour cell lists and it will accelerate the change of neighbour cell lists since the coverage of cells can be altered by the environment. Meanwhile, the ANR technology will be essential in 5G networks. Since the network environment in the future is not similar to the present, the strategy of ANR should also be different from the present. First, since practical neighbour cell lists in each cell are changed frequently and individually, it is necessary to optimize them frequently and individually. Second, since the neighbour cell lists in each cell are not changed similarly, it is necessary to operate ANR flexibly. To respond to these issues, we propose to use network function virtualization (NFV) for ANR. To evaluate the proposed strategies, we measured additional resource consumption and the latency of handover if neighbour cell lists are not optimized when UEs perform handover simultaneously. These experiments are conducted using Amarisoft LTE-100 Platform.

1. Introduction

In the IoT era, the mass use of devices such as wearable devices and wireless sensors in mobility (e.g., vehicle, personal mobility, and watch) will be connected to the networks for convenience. This will cause an increase of simultaneous handover. Thus, the handover performance will become more important for ceaseless connection and QoS. In particular, if there will be an urgent situation to receive data from networks, the handover performance will be critical. Moreover, in the 5G era, the coverage area of cells will become smaller because a superhigh frequency will be used for high throughput. The superhigh frequency cannot propagate very far. Hence, more cells are necessary to cover the same area (i.e., massive small cells). These features will increase the number of neighbour cell lists, and it will accelerate the change of neighbour cell lists because the coverage of cells can be altered by the environment, which includes reflection, diffraction, and shadowing effects. In addition, these small cells will also increase the simultaneous handover because the boundary among cells will increase. Moreover, nowadays, moving cells and small cells are usually used for data traffic rushes and radio shadow areas [1], and macrocells and massive small cells will coexist in 5G networks. In this complex network design, the configuration of neighbour cells becomes more difficult even though it is very important for handover between two neighbour cells.

Meanwhile, it is necessary to inquire assumption that an increase of neighbour cell lists will accelerate the change of neighbour cell lists. First, practical neighbour cell lists, according to the strength of the signal from other cells, are not fixed. This means that a strong signal from far-off cells can be received in the serving cell, and it propagates sufficiently far because radio signal strength is affected by environmental factors such as temperature and humidity [2–5]. This effect is also due to shadowing, fast fading, reflection, and so on [6, 7]. For example, authors in [6] emulated the propagation effects by increasing the standard deviation of the lognormal shadowing in the macrocells. In this simulation, the small cell had just one single neighbour macrocell as a neighbour cell list at the region with the no shadowing (i.e., a small cell was in a macrocell), but four neighbour cells were detected in a

small cell with the shadowing of 12 dB standard deviation. If the standard deviation would increase, the number of neighbour cell lists would increase. On the contrary, if the standard deviation would decrease, the number of neighbour cell lists would again decrease. In this simulation, the number of neighbour cell lists in a macrocell was varied from six to about 20, and the number of neighbour cell lists in a small cell was varied from one to about 10 according to shadowing standard deviation. Therefore, if there are many small cells in a macrocell, it is expected that the number of neighbour cell lists in a cell increases more and the neighbour cell lists in a cell change more frequently.

Due to massive small cells in 5G networks, automatic neighbour relation (ANR) technology, which automatically detects and configures neighbour cells, will be essential and become more important in 5G networks because manual configuration and optimization of neighbour cell lists in each individual small cell will become more costly and difficult. ANR detects a new cell's information including physical cell ID (PCI), E-UTRAN cell global identifier (ECGI), and the IP address from operations, administration, and maintenance (OAM) in order to execute handover to a new cell when a serving cell does not know the new cell and the serving cell receives the strong signal of the new cell. This handover method is called UE-Triggered ANR with OAM Support [8]. In other handover methods, ANR receives only the IP address from OAM and the other information from mobile phones [9]. Basically, ANR makes it possible for enhanced NodeB (eNodeB) to detect the neighbour cells on the basis of UE measurements [10]. The UE eventually sends Measurement Report messages, including PCI, to eNodeB when the UE gets a stranger signal than that from serving eNodeB, or periodically transfers this message [11].

Finally, since the network environment in the future (i.e., an increase of simultaneous handovers and a frequent change of neighbour cell lists) will not be the same as the present, the strategy of ANR should also be changed as we needed a new strategy of self-organizing networks in the past due to enterprise femtocells [12], and the strategies to be considered are discussed as below. First, it is necessary to optimize the neighbour cell lists frequently and individually because practical neighbour cell lists for each cell are changed frequently and individually due to the change of environment and the redundant neighbour cell lists could be burdensome to perform handover quickly. Although traditional ANR function has neighbour removal function [13], it does not consider this future condition (i.e., frequent change of practical neighbour cell lists). Recently, ANR algorithm for this complex network environment is researched [14], but it is only for overreached scenario. To apply the change of a cell's coverage according to a natural phenomenon, it is necessary to check the signal frequently and individual for optimizing neighbour cell lists. Second, it needs to operate ANR flexibly because the neighbour cell lists in some cells are rarely modified while those in other cells are frequently modified. For example, when a great number of people gather for a festival, game, or national holiday, it is necessary to operate new small cells or moving cells for a while. Additionally, if a tall building is built or if it is rainy or sunny, the cell coverage is altered, and this can also change the neighbour cell lists. Since this trend continues and even accelerates, these strategies will become more essential in the future.

For these strategies, we propose using network function virtualization (NFV) [15] for operating ANR. That is, ANR-virtual network function (ANR-VNF), in which network entry is virtualized and has ANR functionality, can be deployed and extended frequently and flexibly. For instance, in the case where the deviation of a floating population is large, such as that at a stadium, it is more efficient and inexpensive to operate moving cells than fixed cells for data offload. Also, in the case where the fluctuation of shadowing standard deviation is large due to weather, more resources are needed for optimizing neighbour cell lists quickly. In any case, in order to configure and optimize neighbour cell lists more quickly, it is necessary to extend the capacity of ANR functionality.

The remainder of this paper is organized as follows. Section 2 provides background on self-organizing networks (SON), ANR, NFV, and so on. Section 3 adduces the experiments for necessity of these strategies and expected hazard. Section 4 explains the proposed methods such as ANR-VNF in detail. Finally, Section 5 gives the conclusions.

2. Background and Related Work

2.1. Self-Organizing Networks (SON). Operating radio networks is a challenging task, especially in cellular mobile communication systems due to their latent complexity. This complexity arises from the number of network elements and interconnections between their configurations. In a heterogeneous network, it is difficult to handle the variety of technologies and their precise operational paradigms. Today, planning and optimization tools are typically semiautomated and management tasks need to be tightly supervised by human operators. This manual effort by the human operator is time-consuming, expensive, and error-prone and requires a high degree of expertise. SON can be used to reduce operating costs by reducing tasks at hand and to protect proceeds by minimising human error. The subsection below details SON taxonomies.

2.1.1. Self-Configuration. Configuration of base stations (eNBs), relay stations (RS), and femtocells is required during deployment, extension, and upgrade of network terminals. Configurations may also be needed when there is a change in the system, such as the failure of a node or a drop in network performance. In future systems, the conventional process of manual configuration needs to be replaced with self-configuration. It is predictable that nodes in future cellular networks should be able to self-configure all of their initial parameters including IP addresses, neighbour list, and radio access parameters.

2.1.2. Self-Optimization. After the initial self-configuration phase, it is significant to continuously optimize system parameters to ensure efficient performance of the system if all its optimization objectives are to be maintained. Optimization in legacy systems can be done through periodic drive

tests or analysis from log reports generated from network operating centers. Self-optimization includes load balancing, interference control, coverage extension, and capacity optimization.

2.1.3. Self-Healing. Wireless cellular systems are prone to faults and failures because of component malfunctions or natural disasters. In traditional systems, failures are mainly detected by the centralized O&M (Operation and Maintenance) software. Events are recorded and necessary alarms are set off. When alarms cannot be cleared remotely, radio network engineers are usually mobilized and sent to cell sites. This process could take days or even weeks before the system returns to normal operation. In future self-organized cellular systems, this process needs to be improved by consolidating the self-healing functionality. Self-healing is a process that relates the remote detection, diagnosis, and triggering of compensation or recovery actions to blunt the effect of faults in the network's equipment.

2.2. Automatic Neighbour Relation (ANR) [7–9]. The coverage of cells is limited because the cell cannot emit radio frequency with unlimited power, so there are many cells for covering a wide area. If one mobile phone moves from the coverage area of one cell to that of another, it would connect to the new cell and disconnect from the old cell. This procedure is called handover. In LTE radio access network, the cell consists only of eNodeB which communicates with each other directly via the X2 interface. Over this X2 interface, neighbouring eNodeBs communicate with each other to prepare and execute handovers. In order to provide seamless mobility in LTE, it is important to set up the X2 interface without omission because there will be no handover between neighbouring eNodeBs unless the X2 interface is set up and functioning.

The X2 interface is set up by using the neighbour cell lists in the Neighbour Relation Tables (NRT) of each eNodeB, so there is neither X2 interface nor handover between neighbour cells if one eNodeB omits the neighbour cell lists in its own NRT, which is caused by a moving cell or a newly added small cell. In this case, Automatic neighbour Relation (ANR) functionality can detect the new neighbour cell and add its list to the NRT automatically.

2.3. Network Function Virtualization [15]. Service provision within the telecommunications industry has traditionally been based on network operators providing physical proprietary devices and equipment for each function. These dedicated requirements for high quality, stability, and stringent protocol adherence have led to long product cycles, very low service agility, and heavy dependence on specialized hardware.

However, the requirements by users for more diverse and new (short-lived) services with high data rates continue to increase. Therefore, Telecommunication Service Providers (TSPs) must correspondingly and continuously purchase, store, and operate new physical equipment. All these factors lead to high CAPEX and OPEX for TSPs. Moreover, the resulting increase in capital and operational costs cannot result in higher subscription fees, so TSPs have been forced to find ways of building more dynamic and service-aware networks with the objective of reducing product cycles, operating, and capital expenses and improving service agility.

NFV [16, 17] has been proposed as a way to address these challenges by forcing virtualization technology to offer a new way to design, deploy, and manage networking services. The main idea of NFV is the decoupling of physical network equipment from the functions that run on them. The goal of NFV is to transform the way that network operators design networks by evolving virtualization technology to reinforce much of the network equipment onto standard servers, which could be located in data centers, distributed network nodes, and at the end user premises. It involves the implementation of network functions in software—VNFs—that can run on one or more industry standard physical servers and that can be moved to various locations in the networks as required without the need for installation of new equipment.

With this, NFV allows TSPs to get more flexibility to further open up their network capabilities and services to users and other services and the ability to deploy or support new network services faster and cheaper so as to realize better service agility. To achieve these benefits, NFV paves the way for a number of differences in the way network service provisioning is realized in comparison to current practice.

2.4. Virtualized Cellular Infrastructure. A basic architecture of LTE networks without NFV shows that the UE is connected to the Evolved Packet Core (EPC) over the LTE access network (E-UTRAN), in which the eNodeB is the base station for LTE radio. The EPC is made up of the Serving Gateway (S-GW), the Packet Data Network (PDN), the Gateway (P-GW), the Mobility Management Entity (MME), and the Policy and Charging Rules Function (PCRF). All these functions are based on dedicated equipment.

In virtualized cellular infrastructure shown in Figure 1, however, the network entries of EPC get virtualized into a data center. Also, a logic part of eNodeB gets virtualized into data center and only the RF part of eNodeB remains. This division of eNodeB is called C-RAN (cloud radio access network). C-RAN features centralized processing, collaborative radio, real-time cloud computing, and power efficient infrastructure [18]. C-RAN is composed of the BBU (baseband unit), OTN (optical transmission network), and the RRU (remote radio unit). The BBUs implement the base station functionality whereas the RRUs perform radio functions. Also, the BBUs and RRUs are applicable to typical RAN, from macrocell to femtocell. Thus, these BBUs can be centralized and this makes networks more efficient and optimized in terms of cost, resources, and energy through the orchestration management system [18, 19].

2.5. Amarisoft LTE-100 Platform. The Amarisoft LTE-100 platform is a software-based LTE station running on a PC. Like a virtualized cellular infrastructure, the Amarisoft LTE-100 platform provides LTE Enhanced Packet Core (EPC) and base station (eNB) on each PC. The EPC includes Mobility Management Entity (MME) with built-in Packet Gateway (P-GW), Serving Gateway (S-GW), and Home Subscriber

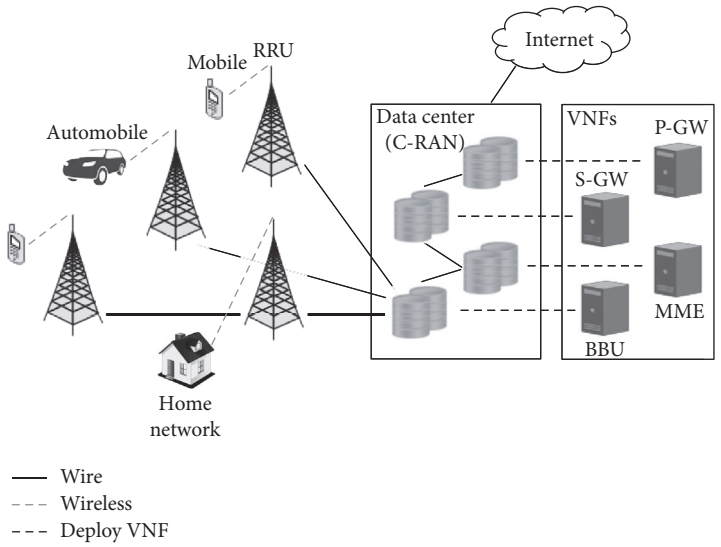


Figure 1: The architecture of virtualized cellular infrastructure.

Servers (HSS). The radio interface in the software-based LTE solution was handled by Ettus Research USRP N210. The antenna configuration used in the experiments was a Single-Input Single-Output (SISO) scheme.

The PHY layer complies with LTE release 13 and supports closed-loop power control, and the protocol layer also complies with LTE release 13 and implements the MAC, RLC, PDCP, and RRC layers. Also, it supports intra-eNodeB, S1, or X2 handovers. For network interface, it supports standard S1AP and GTP-U interfaces to the Core Network and the X2AP interface between eNodeBs.

Like Figure 2, the EPC is connected to the Internet, and the eNodeBs are connected to the EPC, and each eNodeB has an RF unit (USRP N210). It is necessary for each eNodeB to be registered to the EPC and to be connected with the X2 interface between eNodeBs. Then, each eNodeB recognizes other eNodeBs. However, for handover, each eNodeB must have the information of other eNodeBs in the Neighbour Relation Table (NRT). Naturally, it is desirable to have the information of the actual neighbour eNodeBs, but the handover is performed correctly although the NRT has the information of redundant eNodeBs that are not the actual neighbour eNodeBs.

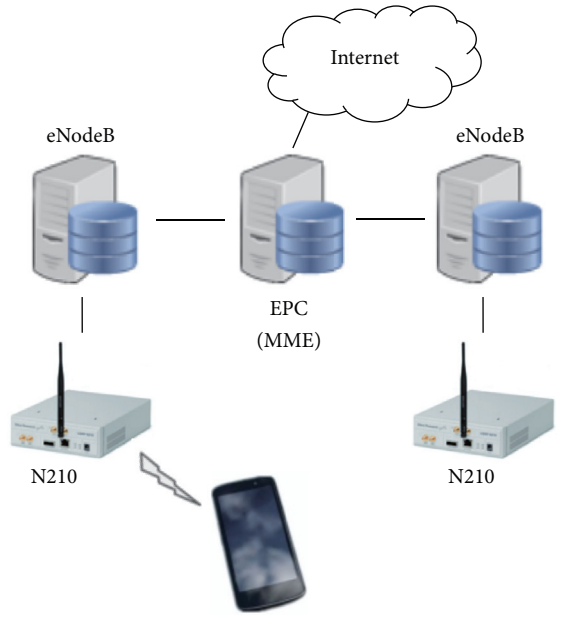


Figure 2: Amari LTE 100 platform configuration.

3. The Experiments for the Strategy of ANR

Since small cells and macrocells are mixed in the transition period to 5G networks and the cell coverage is affected by the surrounding environment, neighbour cell lists will not be fixed in practice and will not be small in number; that is, the number of neighbour cell lists in the NRT of each eNodeB can be increased over the number of physically adjacent cells or can be decreased due to the environmental effects such as shadowing and fading. Furthermore, since the simultaneous handovers will increase due to IoT and massive small cells, it is expected that the handover performance will be more important and has hazard issues. Thus, it is necessary to optimize the NRT (i.e., neighbour cell lists) because redundant neighbour cell lists could be burdensome to perform handover. In order to estimate the performance issues of ANR in 5G, we performed several experiments.

3.1. Experiment Design. In these experiments, the Amarisoft LTE-100 Platform [21], which has one Core Network (EPC) and two eNodeBs, is used, and each component (EPC, eNodeB) is installed in a common computer which has an Intel(R) Core(TM) i7-4790 @ 3.60 GHz CPU and 8 GB memory computer and has Ubuntu 14.04 operating system. This means that network elements like VNF are able to be carried out on a common computer. This platform was

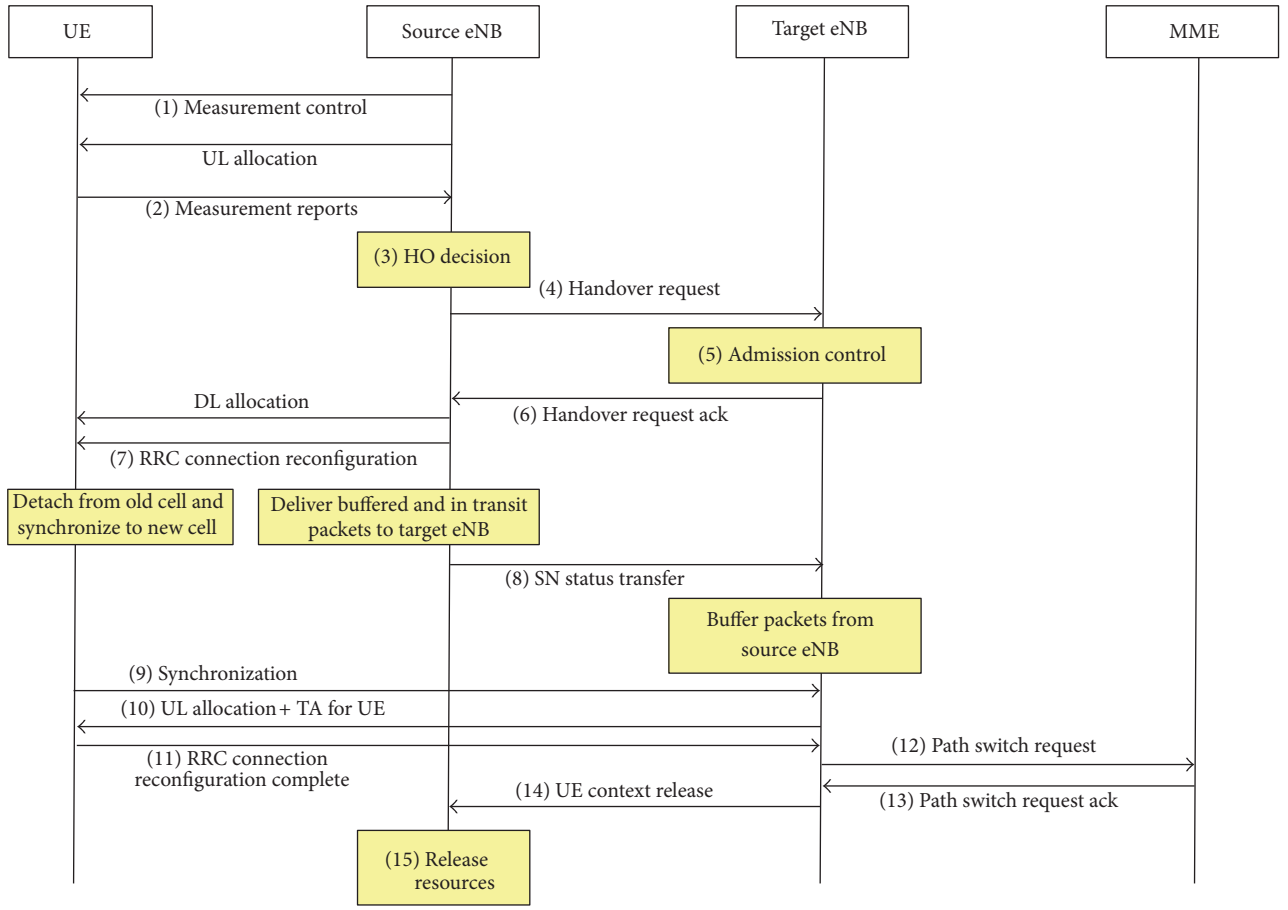


Figure 3: Handover procedure.

used in recent researches [22, 23]. Figure 2 shows the overall network configuration for experiments. There are two eNodeBs and one EPC, and it supports handover. Since there are just two eNodeBs, the optimized NRT of each eNodeB has just one neighbour cell, and the nonoptimized NRT has more neighbour cells. Also, two eNodeBs have the same handover threshold and different frequency, so the handover between two eNodeBs is interfrequency mobility handover. Several devices including a Galaxy S7, iPhone 6s, iPhone 6 plus, and iPhone SE are used as UEs.

For executing handover, the UEs are at a similar distance from two cells and perform handover simultaneously by decreasing the power of the serving cells with command at the same time. In order words, UEs do not move from serving cell to target cell; they are just fixed in the same position with light of sight (LOS). Other power adaptation is not considered. This forces the UEs to have the same handover condition (e.g., the power of the serving cell is less than that of neighbour cell). We increase the number of UEs that perform handover simultaneously (e.g., 2, 3, and 4 UEs), and all these handover experiments are executed over 10 times. During handover, to measure CPU usage, the Linux "top" command is used to measure the CPU usage of the eNodeB program during handover. This "top" command measures CPU usage per second which makes it possible to evaluate the fluctuation of CPU usage for the running the eNodeB program.

For evaluating handover latency, the log information is used. For that, the handover procedure is necessary. Authors in [24] explained the intra-MME/S-GW handover procedure which means that only eNodeB (not MME and S-GW) is changed when handover is executed, and Figure 3 shows this intra-MME/S-GW handover procedure concretely. The handover time (latency) is the period from the Measurement Reports message (2) to the UE Context Release message (14). The handover is triggered by the UE that sends a Measurement Report message (2) to the serving cell, and at the handover completion, the target cell sends the UE Context Release message (14) to the serving cell to inform success of handover via X2AP [25]. Thus, the handover time can be measured by estimating the period from Measurement Report message (2) to UE Context Release message (14). To measure exact handover latency using log information, three computers are synchronized using NTP (network time protocol).

With the handover procedure log, we measure each UE's handover latency, which means the period of each UE's handover when all UEs complete handover simultaneously. We also measure total handover latency, which means the period from the start time of the first UE's handover to the

Table 1: During handover, the average fluctuation of target eNodeB's CPU usage (%) with respect to the number of neighbour cell lists.

Number of neighbour cell lists	Average fluctuation of CPU usage [%]	SEM [%]
1	0.6125	0.280
11	1.1	0.558

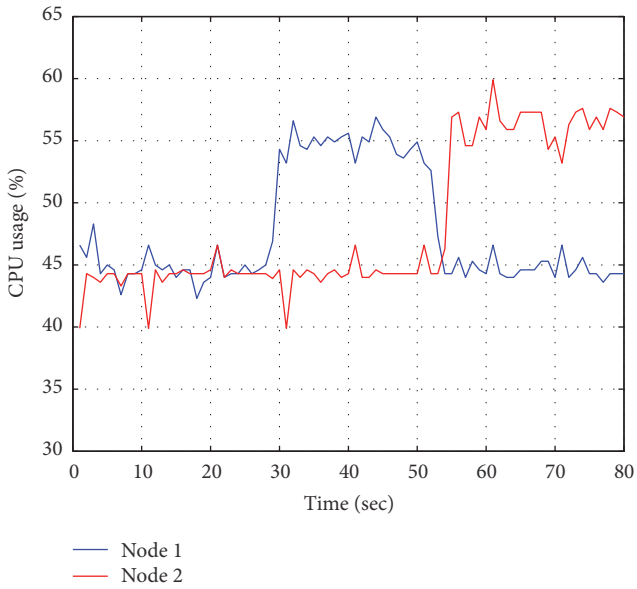


Figure 4: Transition of CPU usage [%] during handover from node 1 to node 2.

end time of the last UE's handover when all UEs complete handover simultaneously.

3.2. Experimental Results. In Table 1, in the case of optimized NRT (i.e., 1 neighbour cell list), the average fluctuation of CPU usage is about 0.6% during handover, and this change is within the average fluctuation at idle. In the case of nonoptimized NRT (i.e., 11 neighbour cell lists), however, the average fluctuation of CPU usage is about 1.1%. It seems that about average 0.5% more of CPU usage is used due to the lookup of neighbour cell lists. In addition, the calculated means from two cases are significantly different at $\alpha = 0.05$ by using ANOVA (the analysis of variance) [26], and we indicate the standard error of the mean (SEM) in all Tables [26]. Even though only the number of neighbour cell lists is increased, the fluctuation of CPU usage is also increased during handover. Thus, it is necessary to optimize neighbour cell lists to stop resource waste.

Meanwhile, if a UE receives packets during handover, the CPU usage for download is transferred from serving cell to target cell. Figure 4 shows this movement of CPU usage from serving cell to target cell when one UE performs handover during download. It seems that resource usage for download is considerable, so proper handover (not ping-pong handover) is important for resource management as well as QoS because correct handover can distribute total resources efficiently.

Table 2: Average handover latency with respect to the number of neighbour cell lists when 1 UE performs handover.

Number of neighbour cell lists	Average latency [msec]	SEM [msec]
1	72.8	2.4
11	73.3	1.9
21	77.2	2.1

Table 3: Each UE's average handover latency with respect to the number of neighbour cell lists when 2 UEs, 3 UEs, or 4 UEs perform handover simultaneously.

	Number of neighbour cell lists	Average latency of handover [msec]	SEM [msec]
2 UEs	11	66.1	1.3
	21	83	3.5
3 UEs	11	68.9	1.2
	21	74.3	1.6
4 UEs	11	79.8	1.5
	21	79.3	1.4

Table 2 shows average handover latency according to the number of neighbour cell lists when one UE performs handover. There are two eNodeBs so that the number of optimal neighbour cell lists is just one. Thus, in Table 2, 21 neighbour cell lists mean that there are an additional 20 neighbour cell lists which do not exist in practice for various reasons. Average handover latency increases along with the number of neighbour cell lists. In particular, the average handover latency rises extremely high when the number of neighbour cell lists is 21. Unfortunately, however, three latencies are statistically equal using ANOVA at $\alpha = 0.05$.

Table 3 indicates the average of each UE's handover latency when 2, 3, and 4 UEs perform handover simultaneously. Although there are slight differences, the average latency of each UE's handover increases according to the increase of UE and neighbour cell lists. In particular, for the case of 2 and 3 UEs, the latency increases statically at $\alpha = 0.05$ when the number of neighbour cells is 21. These latencies, however, do not increase considerably, just 5–10 ms. In addition, it seems that the average of each handover latency does not significantly relate to the number of UEs which simultaneously perform handover. Thus, it is necessary to focus on average total handover latency in each case (i.e., 2, 3, and 4 UEs). To aid the reader's comprehension, Figure 5 describes each UE's handover latency and total handover latency when three UEs perform handover simultaneously.

Table 4 indicates the average of total UEs' handover latency when 2, 3, or 4 UEs perform handover simultaneously. According to the increase of UE and neighbour cell lists, the average of total handover latency is increased. In the case of 21 neighbour cell lists, the average total handover latency of 21 neighbour cell lists is increased statically at $\alpha = 0.05$ compared to the case of 11 neighbour cell lists. Table 5 shows each delay until the 2nd, 3rd, or 4th UE starts handover when 4UEs perform handover simultaneously and the number of

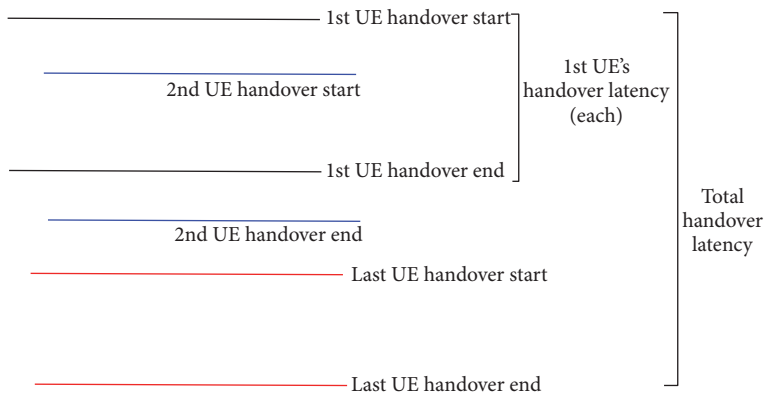


Figure 5: Each UE's handover latency and total handover latency when UEs perform handover simultaneously.

Table 4: Average total handover latency with respect to the number of neighbour cell lists when 2, 3, or 4 UEs perform handover simultaneously.

	Number of neighbour cell lists	Average total latency [msec]	SEM [msec]
2 UEs	11	324.3	79.9
	21	357.2	152.2
3 UEs	11	385.4	49
	21	529.4	32.7
4 UEs	11	513.2	25.9
	21	621.6	22.6

Table 5: Average delay until other UEs start handover since the start of the 1st handover.

	Number of neighbour cell lists	Average delay [msec]	SEM [msec]
2nd UE	11	28.2	6.3
	21	31.7	6.4
3rd UE	11	137.3	10.1
	21	134.2	25.9
4th UE	11	433.6	26.7
	21	506.6	23.5

neighbour cell lists is 11 or 21. When the number of neighbour cell lists is 21, the 4th UE starts handover more later (about 17%) than when the number of neighbour cell lists is 11. This delay leads to the increase of average total handover latency. This shows that as the number of neighbour cell lists and UEs that simultaneously perform handover increase, the average total handover latency increases in a step-like fashion.

As shown in Table 5, the start time (delay) of handover for each UE is different even though all UEs try to perform handover simultaneously. Moreover, this start time is increased according to the increase of UE and neighbour cell lists because the system resources are fixed and more resources will be required. It seems that if there are a lot of neighbour cell lists, the handover decision (3) time in Figure 3 increases because the lookup time of neighbour cell lists also increases, and if the number of UEs which perform handover simultaneously increases, this handover decision (3) time in Figure 3 also increases because more resources are needed.

Therefore, it seems that the average of total handover latency will increase considerably when the number of the neighbour cell lists will be large and these will be not optimized in 5G networks and the number of UEs that perform simultaneous handover will increase dramatically due to IoT and massive small cells. As a result, some UEs will have a delayed handover. This will result in the degradation of QoS.

4. Proposed ANR Model with NFV

As we mentioned, it will be essential to optimize neighbour cell lists in 5G networks. Although recent ANR technology includes the algorithms of optimization of neighbour cell lists and these algorithms are researched continuously, these algorithms are limited (e.g., overreached scenario) and are not sufficient to consider the future network conditions (i.e., an increase of simultaneous handovers and a frequent change of neighbour cell lists due to a natural phenomenon). In this paper, although there is not the algorithm to solve these issues, it is certain that more resources are needed to consider all of them.

In addition, it is also necessary to operate ANR technology flexibly. In some cells, since neighbour cell lists are frequently modified, it is necessary to extend the capacity of the ANR function to quickly optimize neighbour cell lists. Also, it is necessary to diminish the capacity of the ANR function in order to save resources when neighbour cell lists are rarely changed. For this flexible operation of ANR technology, we propose to use NFV for operating ANR. In this case, ANR is an important function for self-optimization as well as self- configuration. Like distributed SON [27], ANR-VNF, which means that network entry is virtualized and has ANR functionality, is deployed in each eNodeB and can be extended and also be diminished. The NFV's properties make it possible. NFV makes it possible to get wanted resources assigned in a common computer and to modify this assigned resource by redeploying VNF.

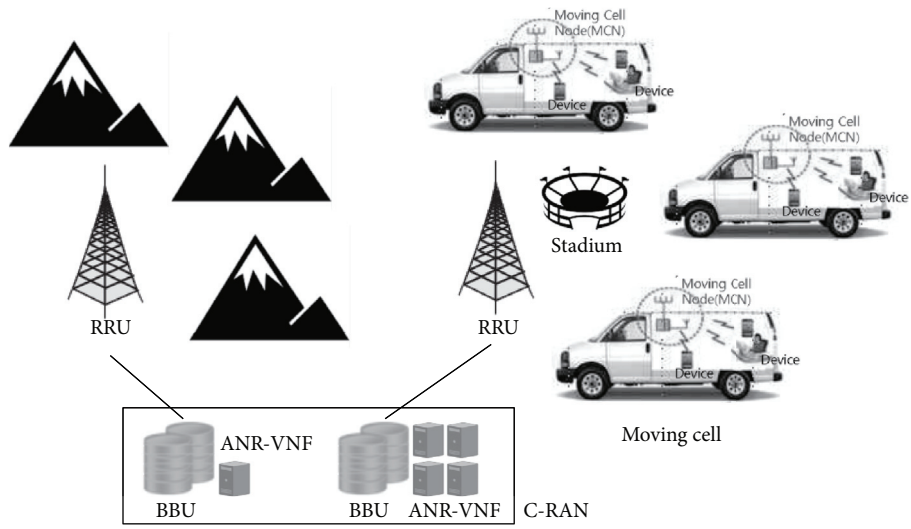


Figure 6: Conceptual diagram of proposed ANR model [20].

Figure 6 depicts a conceptual diagram of the proposed ANR-VNF model. In future, if network entities will be replaced with NFV, eNodeB will run on common computers of data centers like the BBU in this figure. The right side of Figure 6 depicts that ANR-VNF is extended in the BBU because moving cells are located near a stadium for data offload. On the contrary, the left side of Figure 6 shows that, in suburbs and rural areas, ANR-VNF is deployed with minimum resources in the BBU because neighbour cell lists in this eNodeB are rarely changed.

For ANR-VNF, it seems burdensome to deploy typical VMs (virtual machine). Thus, it is necessary to use a Linux container (e.g., Docker) to use ANR-VNF portably [28]. Unlike hypervisor virtualization, containers run in user space on top of an operating system's kernel. Thus, a container allows multiple isolated user space instances to be run on a single host [29]. In other words, if a virtualized eNodeB is deployed on a common computer, ANR-VNF can be run on this host. These containers' functionality makes the proposed strategy possible.

5. Conclusion

In 5G networks, the network environment will change in several cases. Small cells and moving cells will increase due to the frequency property and data offload. These changes will cause neighbour cell lists to be modified more frequently as we mentioned. In addition, since the coverage of cells will become small and many devices will be connected to networks due to IoT, there is no doubt that the number of simultaneous handovers will rise.

These changes will degrade handover performance if neighbour cell lists are not optimized frequently and individually, and this optimization will need more resources. The hazard from these changes is proved through several experiments. Therefore, we propose a new strategy of ANR (i.e., ANR-VNF) by using NFV to overcome this hazard. This strategy can make ANR able to respond to the change of network environments flexibly and efficiently for resource management and handover performance. In future work, we will focus on the algorithm to solve these issues and compare it with any other algorithm. In addition, it is necessary to implement and operate ANR-VNF in each eNodeB.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the "ICT Consilience Creative Program" (IITP-R0346-16-1008), supervised by the IITP (Institute for Information & communications Technology Promotion).

References

[1] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: past, present, and future," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 497–508, 2012.

[2] C. A. Boano, N. Tsiftes, T. Voigt, J. Brown, and U. Roedig, "The impact of temperature on outdoor industrial sensornet applications," *IEEE Transactions on Industrial Informatics*, vol. 6, no. 3, pp. 451–459, 2010.

[3] J. Lee and K. Chung, "An efficient transmission power control scheme for temperature variation in wireless sensor networks," *Sensors*, vol. 11, no. 3, pp. 3078–3093, 2011.

[4] J. Luomala and I. Hakala, "Effects of temperature and humidity on radio signal strength in outdoor wireless sensor networks," in *Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS '15)*, pp. 1247–1255, IEEE, 2015.

[5] C. Ortega-Corral, L. E. Palafox, J. A. García-Macías, J. Sánchez-García, L. Aguilar, and J. I. Nieto-Hipólito, "Parameter optimization of a temperature and relative humidity based transmission power control scheme for wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 2015, article no. 20, Article ID 921319, 2015.

[6] A. J. Fehske, I. Viering, J. Voigt, C. Sartori, S. Redana, and G. P. Fettweis, "Small-cell self-organizing wireless networks," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 334–350, 2014.

[7] C. M. Mueller, H. Bakker, and L. Ewe, "Evaluation of the automatic neighbor relation function in a dense urban scenario," in *Proceedings of the IEEE 73rd Vehicular Technology Conference*, pp. 1–5, Yokohama, Japan, May 2011.

[8] S. Hamalainen, H. Sanneck, and C. Sartori, *LTE Self-Organizing Networks (SON)*, John Wiley & Sons, New York, NY, USA, 2012.

[9] J. T. Penttinen, *The Telecommunications Handbook: Engineering Guidelines for Fixed, Mobile and Satellite Systems*, John Wiley & Sons, Ltd, Chichester, UK, 2015.

[10] Y. Watanabe, Y. Matsunaga, K. Kobayashi, H. Sugahara, and K. Hamabe, "Dynamic neighbor cell list management for handover optimization in LTE," in *Proceedings of the IEEE 73rd Vehicular Technology Conference (VTC '11-Spring)*, Budapest, Hungary, May 2011.

[11] P.-C. Lin, "Minimization of drive tests using measurement reports from user equipment," in *Proceedings of the IEEE 3rd Global Conference on Consumer Electronics (GCCE '14)*, pp. 84–85, IEEE, Tokyo, Japan, October 2014.

[12] L. S. Mohjazi, M. A. Al-Qutayri, H. R. Barada, K. F. Poon, and R. M. Shubair, "Self-optimization of pilot power in enterprise femtocells using multi objective heuristic," *Journal of Computer Networks and Communications*, vol. 2012, Article ID 303465, 14 pages, 2012.

[13] TS ETSI. 136 300 v8. 12.0, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (3GPP TS 36.300 version 8.12.0 Release 8).

[14] D. Ortega-Sicilia, F. Cabrera Almeida, A. Sedeño Noda, and A. Ayala-Alfonso, "Design and evaluation of ANR algorithm for LTE real scenario with high interference," *Electronics Letters*, vol. 51, no. 24, pp. 2057–2058, 2015.

[15] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network function virtualization: state-of-the-art and research challenges," *IEEE Communications Surveys and Tutorials*, vol. 18, no. 1, pp. 236–262, 2015.

[16] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: challenges and opportunities for innovations," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90–97, 2015.

[17] R. Guerzoni, "Network functions virtualisation: an introduction, benefits, enablers, challenges and call for action," in *Proceedings of the SDN and OpenFlow World Congress*, 2012.

[18] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (C-RAN): a primer," *IEEE Network*, vol. 29, no. 1, pp. 35–41, 2015.

[19] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: state of the art, challenges, and implementation in next generation mobile networks (vEPC)," *IEEE Network*, vol. 28, no. 6, pp. 18–26, 2014.

[20] Y. Hwang and J. Shin, "A user centric moving cell management mechanism in LTE-Advanced system," in *Proceedings of the 5th International Conference on Information and Communication Technology Convergence (ICTC '14)*, pp. 243–245, Busan, Korea, October 2014.

[21] "Software LTE base station on PC Amari LTE 100," http://www.amarisoft.com/.

[22] R. Trestian, Q.-T. Vien, P. Shah, and G. Mapp, "Exploring energy consumption issues for multimedia streaming in LTE HetNet small cells," in *Proceedings of the IEEE 40th Conference on Local Computer Networks (LCN '15)*, pp. 498–501, IEEE, Clearwater Beach, Fla, USA, October 2015.

[23] X. Xiong, T. Wu, H. Long, and K. Zheng, "Implementation and performance evaluation of LECIM for 5G M2M applications with SDR," in *Proceedings of the IEEE Globecom Workshops (GC Wkshps '14)*, pp. 612–617, December 2014.

[24] J. Han and B. Wu, "Handover in the 3GPP long term evolution (LTE) systems," in *Proceedings of the Global Mobile Congress (GMC '10)*, IEEE, Shanghai, China, October 2010.

[25] A. Hans, A. Sharma, K. Kumar, and N. Singh, "An overview of handoff procedure in LTE technology," in *Proceedings of the International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom '14)*, pp. 391–394, Greater Noida, India, November 2014.

[26] R. J. Freund, D. Mohr, and W. J. Wilson, *Statistical Methods*, Elsevier, Amsterdam, The Netherlands, 3rd edition, 2010.

[27] S. Feng and E. Seidel, *Self-Organizing Networks (son) in 3gpp Long Term Evolution*, Nomor Research GmbH, Munich, Germany, 2008.

[28] C. Boettiger, "An introduction to docker for reproducible research," *ACM SIGOPS Operating Systems Review*, vol. 49, no. 1, pp. 71–79, 2015.

[29] J. Turnbull, *The Docker Book: Containerization is the new virtualization*, James Turnbull, 2014.

Hindawi
Mobile Information Systems
Volume 2017, Article ID 6740585, 16 pages
http://dx.doi.org/10.1155/2017/6740585

Research Article

A Mobile Network Planning Tool Based on Data Analytics

Jessica Moysen, Lorenza Giupponi, and Josep Mangues-Bafalluy

Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), Av. Carl Friedrich Gauss 7, 08860 Castelldefels, Spain

Correspondence should be addressed to Jessica Moysen; jessica.moysen@cttc.es

Received 3 August 2016; Accepted 14 November 2016; Published 5 February 2017

Academic Editor: Piotr Zwierzykowski

Copyright © 2017 Jessica Moysen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Planning future mobile networks entails multiple challenges due to the high complexity of the network to be managed. Beyond 4G and 5G networks are expected to be characterized by a high densification of nodes and heterogeneity of layers, applications, and Radio Access Technologies (RAT). In this context, a network planning tool capable of dealing with this complexity is highly convenient. The objective is to exploit the information produced by and already available in the network to properly deploy, configure, and optimise network nodes. This work presents such a smart network planning tool that exploits Machine Learning (ML) techniques. The proposed approach is able to predict the Quality of Service (QoS) experienced by the users based on the measurement history of the network. We select Physical Resource Block (PRB) per Megabit (Mb) as our main QoS indicator to optimise, since minimizing this metric allows offering the same service to users by consuming less resources, so, being more cost-effective. Two cases of study are considered in order to evaluate the performance of the proposed scheme, one to smartly plan the small cell deployment in a dense indoor scenario and a second one to timely face a detected fault in a macrocell network.

1. Introduction

Nowadays, we are assisting to the definition of what 5G networks will look like. 3GPP has started multiple work items that will lead to the definition of a novel 5G radio and architecture [1]. The main vendors are publishing numerous white papers presenting their view on 5G networks and architectures. The EU commission has put in place an important 5G Infrastructure Public Private Partnership (5GPPP) program to fund research for 5G networks [2]. What is clear from all these converging visions is that network management in beyond 4G and future 5G networks has to face a whole new set of challenges, due to (1) ultradense deployments, heterogeneous nodes, networks, applications, and Radio Access Networks (RANs) and also heterogeneous spectrum access through novel technologies, such as Long Term Evolution Unlicensed (LTE-U) and License Assisted Access (LAA), all coexisting in the same setting, (2) the need to manage very dynamic networks where part of the nodes is controlled directly by the users (e.g., femtocells), energy saving policies generating a fluctuating number of nodes, active antennas, and so on, (3) the need to support 1000x traffic and 10x users and to improve energy efficiency, (4) the need to improve the experience of the users by enabling Gbps speeds and highly reduced latency, or (5) the need to manage new virtualised architectures.

In this context, it is already widely recognized that the network needs to establish new procedures to become more intelligent, self-aware, and self-adaptive. An initial step in this direction has been introduced in 4G LTE networks, since Release 8, with the introduction of Self-Organising Network (SON). However, this vision needs to be further developed in 5G considering the huge complexity of these networks. As we already observed in [3], a huge amount of data is currently already generated in 4G networks during normal operations by control and management functions, and more data is expected to be gathered in 5G networks due to the densification process [4], heterogeneity in layers and technologies, the additional control and management complexity in Network Functions Virtualisation (NFV) and Software Defined Network (SDN) architectures, the increasing relevance of Machine to Machine (M2M) and Internet of Things (IoT) communications, the increasing variety of applications and services, each with distinct traffic patterns and QoS/Quality of Experience (QoE) requirements, and so on.

The main objective of network management is then to make the network (1) more self-aware, by exploiting and analysing data already generated by the network (this is expected to drive network management from reactive to predictive) and (2) more self-adaptive by exploiting intelligent control decisions tools, offered by ML, based on learning and experience.

In this paper, among many network management problems, we focus on smart network planning, which gives particular emphasis to the QoS offered to the users and the resources used by the operator to offer it. We believe that the use of smart network planning tools is crucial and inevitable for operators running multi-RAT, multivendor, multilayer networks, where an overwhelming number of parameters needs to be configured to optimise the network performance. The state of the art in network planning, in literature, and in the market offers a wide range of platforms systems and applications proposed by the research community and/or oriented to the industry. From an industry perspective, the market of commercial planning tools aims at providing a complete set of solutions to design and analyse networks [5–8]. For instance, in [9], the authors present an open-source network planning tool that includes different planning algorithms to analyse the network under different failures and energy efficiency schemes. However, these works in general focus on several configuration scenarios, RF coverage planning, network recovery test, traffic load analysis, and forecasting traffic, among others, and not directly on QoS offered to end-users and the resources that mobile network operators need to offer. Other works are more targeted to QoS estimation, but not in the area of network planning. The literature already offers different works targeting the problem of QoS prediction and verification, such as [10, 11]. In our preliminary work [12], we focus on complex multilayer heterogeneous networks where we predict QoS independently of the physical location of the UE, that is, based on data learned throughout the whole network. Preliminary results show that by abstracting from the physical position of the measurements we can provide high accuracies in the estimations of QoS in other arbitrary regions. Furthermore, the results presented in [13] show that data analysis achieves better performance in a reduced space rather than in the original one.

This paper presents a smart network planning tool that works in two steps. First, it estimates the QoS at every point of the network based on user-collected measurements, which may have been taken at different time instants and anywhere in the heterogeneous network. We perform this estimation through supervised learning tools. This results in an appropriately tuned QoS prediction model, which is then integrated in the next step. Second, we adjust the network parameters in order to reach certain network objectives by evaluating different combinations and calculating the resulting QoS based on the model of step 1. Network objectives are set in terms of PRB per Mb. We focus on this specific QoS indicator because it combines information that is relevant to the operator (PRBs) and other that is relevant to the end-user (Mb of data) into a single metric. Therefore, the minimization of this indicator allows serving users with an improved spectral efficiency and offered QoS. More specifically, to carry out this optimisation we take advantage of GAs, which are stochastic search algorithms useful to implement learning and optimisation tasks [14], and so are adequate for our purpose.

In order to evaluate the performance of the proposed scheme, we consider two use cases. The first one is in a densified indoor scenario, and the second one is in a more traditional macrocell scenario. In the first use case, we focus on how to plan a dense small cell deployment inspired in a typical 3GPP dual stripe scenario [15], where the parameters to adjust are the number of small cells and their position. In the second use case, we deal with self-healing aspects in macrocellular scenarios. We focus on readjusting the parameters of the surrounding cells to quickly solve an outage problem by automatically adjusting their antenna tilt parameters.

We evaluate the performance of the proposed planning tool through a network simulation campaign carried out over the 3GPP compliant, full protocol stack ns-3 LENA module. We show that the proposed ML-based network planning, differently from other closed optimisation approaches, is extremely flexible in terms of application problem and scenario, and, in this sense, it is generic. Without loss of generality, particular attention has been focused on how to plan a dense 4G small cell deployment and how to quickly solve an outage problem. As can be seen, the approach is equally valid and shows good performance in both studied use cases, characterized by very different optimisation problems and scenarios. As for the RAT, we have focused on 4G and its evolution. The same technique can be applied also to 2G and 3G technologies.

This paper is organised as follows. The general approach is described in Section 2. It describes the ML-based network planning tool, its main design principles, and algorithms we use to build it. In Section 3, we discuss the specific design details and tuning of the QoS prediction model and we put all the pieces (i.e., prediction and optimisation) together. In Section 4, we present the details of the two use cases to which the tool is applied, the simulation platform, and the simulations results. Finally, Section 5 concludes the paper.

2. ML-Based Network Planning Tool Description

We propose designing a network planning tool, which works in two steps. First, we propose to model the QoS through the analysis of data extracted from the networks in the form of measurements. This phase requires to first prepare the data and then to analyse it. And second, we keep on adjusting the parameters and analysing the impact on QoS based on the previous model. In this way, the performance of the network is optimised to meet certain operator targets. Figure 1 presents the different phases required by the proposed network planning.

(1) Data Preparation. This process aims at transforming data into a meaningful format for the estimation at hand. The target is to integrate and prepare large volumes of data

Table 1: Relevant sources of information in mobile networks.

Source	Information	Usage
Control info for short-term network operation	Call/session setup, release, maintenance QoS, RRC, idle and connects mode mobility	Discarded after usage
Control info for SON functions	Info on Radio link failure, intercell interference, UE measurements, MDT measurements, radio resource status, cell load signalling, etc.	Heuristic algorithms typically discard info after
Management information for long-term operation	Fault configuration, accounting, performance and security management (FCAPS), Operations and Management, e.g., in OAM aggregated statistics per eNB, on network performances, # users, successful/failed HO, active bearers, information from active probing	Mainly used for triggering engineer intervention
Customer Relationship Information	Complaints about bad service quality, churn info	Only used by customer service

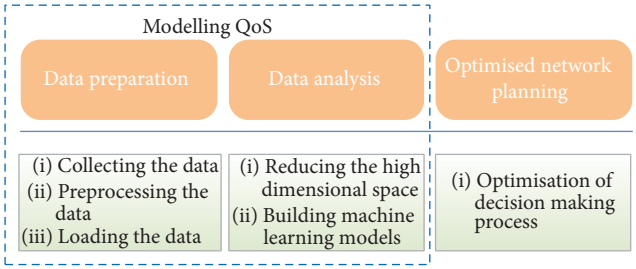


Figure 1: Architecture of the network planning tool.

over the network to provide a unified information base for analysis. To do this, we follow the Extract-Transform-Load (ETL) process, which is responsible for pulling data out of the source and placing it into a database. It involves 3 main steps: data extraction (E), whose objective is to collect the data from different sources; data transformation (T), which prepares the data for the purpose of querying and analysis; data loading (L), which loads the data into the main target, most of the cases into a flat file. This process plays an important role for the design and implementation of planning for future mobile networks. The objective is to create a data structure that is able to provide meaningful insights. Some examples of the kind of sources available in mobile networks are shown in Table 1 [3]. Here the data are classified based on the purpose for which they are generated in the network. The usage that is given nowadays in the network is also suggested in the last column. For the purpose of network planning, we plan to extract data reported by the UEs to the network in the form of UE measurements, in terms of received power, received quality, and offered QoS.

Once the data has been collected, we prepare the data for storing, using the proper structure for the querying.

(2) Data Analysis. The objective of this process is to discover patterns in data that can lead to predictions about the future. This is done by finding this information/correlation among the radio measurements extracted from the network. We do this by applying ML techniques.

(3) Optimised Network Planning. The objective of this process is to find the configuration parameters for the optimised network planning based on the information extracted from the previous data analysis process. In the complex cellular context, we need to deal with several network characteristics that introduce high complexity, for example, the very large number of parameters, the strong cross-tier interference, fast fading, shadowing, and mobility of users. In order to deal with these issues and to guarantee an appropriate network planning, in this work, we propose to use GAs, which allow avoiding some of the problems of typical closed optimisation techniques (e.g., computational intractability) in this complex and dynamic scenario. More specifically, they work with chromosomes (i.e., a given combination of values for the parameters to be tuned in the network). For each of the chromosomes, they calculate its fitness score based on a given objective function (in our case, the QoS predicted by the model) [17]. Then, they select the chromosomes with the best fitness score and generate better child chromosomes by combining the selected ones and they keep on iterating until the objective function of the chromosomes generated reaches the performance target. In this way, GAs perform parallel search from a population of points, which represent the values of the different parameters to be tuned in the scenario and, jointly with other techniques, they have the ability to avoid local minima and use probabilistic search rules.

2.1. Modelling the QoS. We estimate the QoS at every point of the network based on measurements collected in different moments in time and from other regions of the heterogeneous network, that is, based on the measurement history of the network. To do this, we consider the *data preparation* and *data analysis* processes of the network planning tool. As mentioned previously, the objective of these 2 processes is to extract, prepare, and analyse the information already available in the network to provide insightful information from the analysis of it. In fact, in this kind of estimations, ML techniques can be very effective to make predictions based on observations. Therefore, we take advantage of ML techniques to create a model that allows estimating the QoS by learning the relation between PHY layer measurements and QoS measured at the UE. We propose using SL, since among many applications it offers tools for estimation and

prediction of behaviours. In particular, we focus on a regression problem, since we want to analyse the relationship between a continuous variable (PRB per Mb) and the data extracted from the network in the form of UE measurements. Many regression techniques have been developed in the SL literature, and criteria to select the most appropriate method include aspects such as the kind of relation that exists between the input and the output or between the considered features, the complexity, the dimension of the dataset, the ability to separate the information from the noise, the training speed, the prediction speed, the accuracy in the prediction, and so on. We focus on regression models, and we select the most representative approaches. We then use ensemble methods to sub-sample the training samples, prioritizing criteria such as the low complexity and the high accuracy.

We then build a dataset of user measurements, based on the same data contained in the Minimization of Drive Tests (MDT) database. The MDT is a standardized database used for different 3GPP use cases. The dataset contains training samples (rows) and features (columns) and is divided into 2 sets, the training set to train the model and the test set to make sure that the predictions are correct. That training data develop a predictive model and evaluate the accuracy of the prediction, by inferring a function $f(\mathbf{x})$, returning the predicted output $\widehat{y}$. The input space is represented by an n-dimensional input vector $\mathbf{x} = (x^{(1)}, \ldots, x^{(n)})^T \in \mathbf{R}^n$. Each dimension is an input variable. In addition, a training set involves m training samples $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$. Each sample consists of an input vector $\mathbf{x}_i$ and a corresponding output y_i of one data point i. Hence, $x_i^{(j)}$ is the value of the input variable $x^{(j)}$ in training sample i, and the error is usually computed via $|y_i - \widehat{y}_i|$ or with the root mean square error.

In addition to regression analysis, we exploit Unsupervised Learning (UL) techniques for dimensionality reduction to filter the information in the data that is actually of interest (thus reducing the computational complexity) while maintaining the prediction accuracy. We do this by feeding the data into an ensemble method consisting of Bagging/AdaBoost to manipulate the training samples. And after that, the SL techniques under evaluation are then applied. Details for each step are given in the following, and the whole process is depicted in Figure 2.

(1) Collecting the Data. The data we take into account comes from mobile networks, which generate data in the form of network measurements, control, and management information (Table 1). As we mentioned previously, we focus on MDT functionality, which enables operators to collect User Equipment (UE) measurements together with location information, if available, to be used for network management, while reducing operational costs. This feature has been introduced by 3GPP since Release 10; among the targets there are the standardization of solutions for coverage optimisation, mobility, capacity optimisation, parametrization of common channels, and QoS verification. In this context, the literature already offers different solutions for this feature. An example of that can be observed in, [18, 19]. Since operators are also interested in estimating QoS performance, in Release 11, the

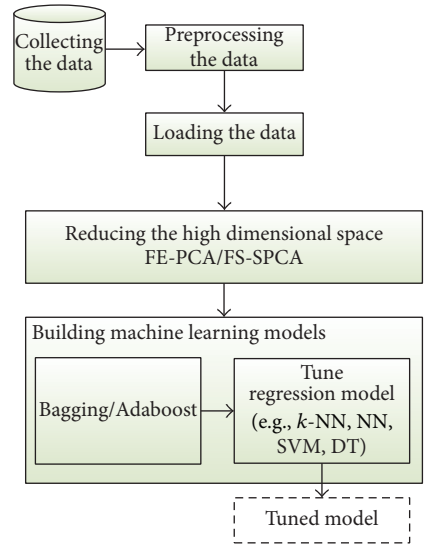


Figure 2: Modelling the QoS.

MDT functionality has been enhanced to properly dimension and plan the network by collecting measurements of throughput and connectivity issues [20]. Therefore, we collect for each UE (1) the Reference Signal Received Power (RSRP) and (2) the Reference Signal Received Quality (RSRQ) coming from the serving and neighbouring eNBs. The size of the input space is $[l \times n]$. The number of rows is the number l of UE in the scenario, and the number of columns corresponds to the number of measurements n. The size of the output space is $[l \times 1]$, which corresponds to the QoS performance in terms of the PRB per transmitted Mb. These measurements gathered at arbitrary points of the network throughout its lifetime are exploited to plan other arbitrary future deployments.

(2) Preprocessing the Data. In order to obtain a good performance during the evaluation, the input variables of the different measurements must be in a similar scale and range. So, a common practice is to normalize every variable between $-1 \leq x^{(j)} \leq 1$ range and replace $x^{(j)}$ with $x^{(j)} - \mu^{(j)}$, over the difference between the maximum and minimum values of the input variables in the dataset, where $\mu^{(j)}$ is the average of the input variable (j) in the dataset. The normalized data is then split into training and test set. We create a random partition from the l sets of input. This partition divides the observations into a training set of m samples and a test set $p = l - m$ samples. We randomly select approximately $p = (1/5) \times l$ observations for the test set.

(3) Loading the Data. This process varies widely. As we mentioned before, depending on the operator requirements, the data can be updated or new data can be added in a historical form at regular intervals. For our propose, in this particular work, we maintain a history of all changes to the data loaded in the network.

(4) Reducing the High Dimensional Space. One of the problems that mobile operators have to face in this kind of

networks is the huge amount of potential features we have as input. In our particular case, features would substantially increase as networks densify. Therefore, to deal with the huge amount of features, we propose applying regression techniques in a reduced space, rather than in the original one. The idea behind that comes from our previous work [12], in which we observed that using a high dimensional space did not result in the best performance. Therefore, we suggest applying the regression analysis in a reduced space. As a result, we take advantage of dimensionality reduction techniques to reduce the number of random variables under consideration. These methods can be divided into Feature Extraction (FE) and Feature Selection (FS) methods. Both methods seek to reduce the number of features in the dataset. FE methods do so by creating new combinations of features (e.g., Principal Component Analysis (PCA)), which project the data onto a lower dimensional subspace by identifying correlated features in the data distribution. They retain the c Principal Components (PCs) with greatest variance and discard all others to preserve maximum information and retain minimal redundancy [21]. Correlation-based FS methods include and exclude features present in the data without changing them. An example is Sparse Principal Component Analysis (SPCA), which extends the classic method of PCA for the reduction of dimensionality of data by adding sparsity constraints on the input features; that is, by adjusting a set of weights over the input features, it induces a matrix in which most of the elements are zero in the solution. In FS-SPCA the sparsity is used to select the f features that give the most useful information; as we increase the weight of SPCA, the number of features is reduced. That is, by adding sparsity constraints on the input features, we promote solutions in which only a small number of input features capture most of the variance. Some preliminary work on these features was presented in [13].

(5) Building the Machine Learning Models. We select some representative regression models, prioritizing criteria such as low complexity and high accuracy: (1) k-Nearest Neighbours (k-NN), (2) Neural Networks (NN), (3) Support Vector Machines (SVMs), and (4) Decision Tree (DT), and we analyse them by performing an empirical comparison of these algorithms, observing the impact on the prediction of the different kinds and amounts of UE measurements.

(1) k-NN can be used for classification and regression [22]. The k-NN method has the advantage of being easy to interpret and fast in training and parameter tuning is minimal.

(2) NN is a statistical learning model inspired by the structure of a human brain where the interconnected nodes represent the neurons to produce appropriate responses. NN support both classification and regression algorithms. NN methods require parameters or distribution models derived from the dataset, and in general they are susceptible to overfitting [23].

(3) SVMs can be used for classification and regression. The estimation accuracy of this method depends on a good setting of the regularization parameter C, ϵ, and the kernel parameters. This method in general shows high accuracy in the prediction, and it can also behave very well with nonlinear problems when using appropriate kernel methods [24].

(4) DT is a flow-chart model, which supports both classification and regression algorithms. Decision trees do not require any prior knowledge of the data, are robust, and work well on noisy data. However, they are dependent on the coverage of the training data, as is the case for many classifiers, and they are also susceptible to overfitting [25, 26].

In order to enhance the performance of each learning algorithm described before, instead of using the same dataset to train, we can use multiple data sets by building an ensemble method. Ensemble methods are learning models that combine the opinions of multiple learners. This technique has been investigated in a huge variety of works [27, 28], where the most useful techniques have been found to be Bagging and AdaBoost [29]. Bagging manipulates the training examples to generate multiple hypothesis. It runs the learning algorithm n_{iter} times, each one with a different subset of training samples. AdaBoost works similarly, but it maintains a set of weights over the original training set and adjusts these weights by increasing the weight of samples that are misclassified and decreasing the weight of examples that are correctly classified [30].

In summary, once the extracted data has been processed and loaded into a file, it is fed into a dimensionality reduction step, and subsequently into an ensemble step, which manipulates the training set by applying Bagging/AdaBoost techniques. The learning algorithm is then applied to produce a regression (see Algorithm 1).

(6) Evaluation of Accuracy. To evaluate the accuracy of the model, the performance of the learned function is measured on the test set. That is, we use a set of samples used to tune the regression algorithm. For each test value, we predict the average QoS and compare it with the actual value in terms of the Root Mean Squared Error (RMSE) of the prediction as follows $\text{RMSE} = \sqrt{\sum_{i=1}^{p}(y_i - \widehat{y}_i)^2/p}$, where p is the length of the test set, $\widehat{y}_i$ indicates the predicted value, and y_i is the testing value of one data point i. In order to compare the RMSE with different scales, the input and output variable values are normalized as follows: $\text{NRMSE} = \text{RMSE}/(y_{\max} - y_{\min})$, where $y_{\max}$ and $y_{\min}$ represent the max and min values in the output space Y_{test} of size $[p \times 1]$, respectively.

2.2. Optimised Network Planning. As we mentioned before, the objective of this process is to close the loop by adjusting the parameters, and so the network performance, through a GA. This results in a GA organised into different phases, as depicted in Figure 3.

(1) Create $\mathbf{S}$ *Feasible Solutions.* We create a set $\mathbf{S} = \{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(p_{\text{size}})}\}$ of feasible solutions (also called chromosomes or individuals), where p_{size} is the starting population size. We denote by $\boldsymbol{\theta}^{(s)} = (\theta_1^{(s)}, \ldots, \theta_M^{(s)})$ the configuration

Input: Input space of size $[m \times n](X_{\text{train}})$,
output space of size $[m \times 1](Y_{\text{train}})$,
Input space of size $[p \times n](X_{\text{test}})$,
output space of size $[p \times 1](Y_{\text{test}})$,
number of iterations (n_{iter})
Output: *model*
// Given the training set ($X_{\text{train}}, Y_{\text{train}}$)
$X_{\text{train}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$
$Y_{\text{train}} = \{y_1, \ldots, y_m\}$
// Apply dimensional reduction step (if it is necessary)
obtain c-dimensional input vector
Apply regression analysis in a reduced space
// Set up the data for Bagging/Adaboost
for $k = 1$ to n_{iter} **do**
  // Call regression algorithm
  model(k) := regression algorithm, namely k-NN, NN, SVM, DT
  // Predict the average QoS
  $\text{QoS}_{\text{predicted}}$:= predict(model(k), X_{test})
  Evaluate performances against the actual value (Y_{test}) by NRMSE
end for
// Result of training base learning algorithm
model := best(model)
return (model)

Algorithm 1: Train regression algorithm.

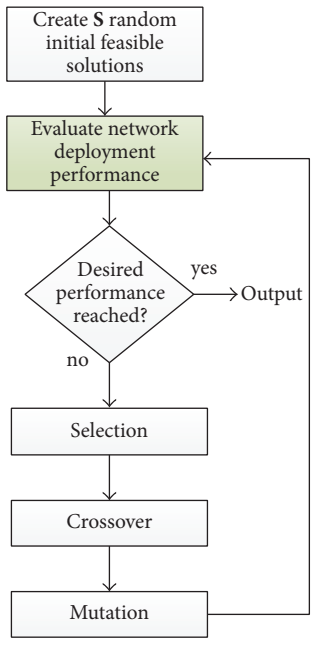


Figure 3: Optimised network planning.

parameters vector of an individual s, with $\theta_j^{(s)}$ denoting the value for the parameter of eNB j, for instance, the transmitted power ($\theta_{j_{\text{txp}}}^{(s)}$), the antenna tilt ($\theta_{j_{\text{tilt}}}^{(s)}$), or the action to switch ON or OFF ($\theta_{j_{\text{sc}}}^{(s)}$) the jth small cell.

(2) Evaluate Network Deployment Performance. This function is responsible for evaluating the network deployment performance. The objective is to calculate the objective function (fitness) of each individual. Given a particular configuration parameter vector $\boldsymbol{\theta}^{(s)}$, this function is responsible for returning the average offered QoS, that is, the average network performance predicted based on the measurement history of the network. This function takes as input **S** feasible solutions and the model produced by the regression algorithm discussed in the previous section. That is, the output of Algorithm 1.

The behaviour of this module is as follows. In each iteration, we collect X_{eval} measurements in some arbitrary q points in the scenario. These measurements are obtained as a consequence of having configured the parameters of the scenario according to each $\boldsymbol{\theta} \in \mathbf{S}$. Based on these measurements, the QoS in the points of interest is predicted by using the model generated in step 1 (see **evalNetPerformance** function in Algorithm 2). Finally, for a given individual, the average of the predicted QoS at the q points is returned as an indicator of the performance of the system with this setup.

As we mentioned before, our metric of interest is the PRB per transmitted Mb, since reducing it allows improving the QoS of the users while also improving the spectral efficiency of the operator. Therefore, the fitness function aims at finding the configuration of parameters for which the total PRB per transmitted Mb is minimized. The operator can target a desired value for the total PRB per transmitted Mb, and, based on this, the network planning tool can decide when the objective has been achieved and interrupt the operation. Therefore, as in any GA, we try to improve the tuning of parameters for each new generation through the processes

Input: Configuration parameters vector ($\boldsymbol{\theta}$),
the tuned model (model)
Output: average QoS
evalNetPerformance := function($\boldsymbol{\theta}$, model){
// Call ns-3 network simulator
evaluate $\boldsymbol{\theta}^{(s)} = (\theta_1^{(s)}, \ldots, \theta_M^{(s)})$
// Collect measurements at q points in the scenario
$X_{\text{eval}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_q\}$
// Predict the average QoS
$\text{QoS}_{\text{predicted}} :=$ predict(model, X_{eval})
average QoS := mean($\text{QoS}_{\text{predicted}}$)
return (average QoS)}

Algorithm 2: Evaluate network deployment performance.

described below (i.e., selection, crossover, and mutation). And our measure of improvement is the average predicted QoS values for the individuals belonging to each generation.

(3) Selection. We select the best fit individuals for reproduction based on their fitness, that is, this function generates a new population of individuals from the current population. This selection is known as elitist selection. Elitism copies the best e fittest candidates into the next generation.

(4) Crossover. This function forms a new individual by combining part of the genetic information from their parents. The idea behind crossover is that the new individual may be better than both parents if he takes the best characteristics from each of them. We use an arithmetic crossover, which creates new individuals (β) that are the weighted arithmetic mean of two parents. If $\boldsymbol{\theta}^{(a)}$ and $\boldsymbol{\theta}^{(b)}$ are the parents, the function returns as follows:

$$\beta = \alpha \times \boldsymbol{\theta}^{(a)} + (1 - \alpha) \times \boldsymbol{\theta}^{(b)}, \quad (1)$$

where α is a random weighting factor chosen before each crossover operation. That is, the arithmetic crossover operator combines two parent chromosome vectors to produce a new individual, where α is a random value between $[0, 1]$.

(5) Mutation. This function randomly selects a parameter based on a uniform random value between a minimum and maximum value. It maintains the diversity in the value of the parameters for subsequent generations. That is, it avoids premature convergence on a local maximum or minimum. For that, we set to δ the probability of mutation in a feasible solution $\boldsymbol{\theta}^{(s)}$. If δ is too high, the convergence is slow or it never happens. Therefore, most of the times δ tends to be small. Finally, we replace the worst fit population with new individuals. The whole genetic algorithm is described in Algorithm 3.

3. Design of the Network Planning Tool

This section presents a discussion on how to tune the model for the ultradense complex scenarios under consideration and then explains how this model is integrated into the global network planning tool.

Table 2: Overall model accuracy.

Approaches	Regression model	Bagging	AdaBoost
(1) FE-PCA	(1.1) k-NN	90.33%	91.98%
	(1.2) NN	93.44%	92.28%
	(1.3) SVM	**94.70%**	94.07%
	(1.4) DT	92.84%	93.60%
(2) FS-SPCA	(2.1) k-NN	89.69%	90.88%
	(2.2) NN	92.41%	91.78%
	(2.3) SVM	93.62%	92.87%
	(2.4) DT	91.22%	92.08%

3.1. Design and Evaluation of the QoS Modelling Component. This section compares the different options for each of the components that define our QoS model, namely, dimensionality reduction, ensemble methods, and regression methods. We consider different options for each of them: (1) FE-PCA and FS-SPCA for dimensionality reduction; (2) Bagging and AdaBoost as ensemble methods; and (3) k-NN, NN, SV, and DT as regression models, as anticipated in Section 2.1.

Table 2 summarizes the performance accuracy of each learning algorithm. Accuracy is measured as $(1 - \text{NRMSE}) \times 100$. Based on this table, we discuss potential design decisions:

(1) FS-SPCA is a very useful approach if we are interested in excluding features to retain minimal information redundancy. This can be observed in Figure 4, which shows the NRMSE as a function of different number of features (f) selected by the SPCA, and where results reveal that for $f = 92$ features we obtain the lowest NRMSE value. As a consequence, we select the 92 features that give us the most useful information. On the other hand, for the FE-PCA approach, we observed in Figures 5 and 6 that we can obtain the 70% of cumulative variance if we consider $c = 10$ PCs. That is, with the first 10 PCs we already capture the main variability of the data. Therefore, on the one hand, FE-PCA would present less inputs to the SL step of the data processing chain at the cost of a prior processing of features. On the other hand, FS-SPCA would simplify the initial feature processing, since selected features are taken as they are, but the cost would be the higher needed storage.

In terms of the overall accuracy of the prediction, by implementing the FS-SPCA approach, we can reduce the dimensionality of the data down to f features and still maintain almost the same accuracy with respect to FE-PCA approach. That is, if we consider the 92 features, we lose only 1% of accuracy with respect to FE-PCA (i.e., 92 inputs versus 10 inputs to the SL step). Therefore, it will depend on the specific network and operator to select whether computing or storage

Input: Initial population (**S**),
size of **S** (p_{size}),
the tuned model (model),
number of generations (g),
rate of elitism e,
rate of mutation δ
Output: solution $\boldsymbol{\theta}^{(*)}$
//Initialization
for $i = 1$ to g **do**
// Return the value of average QoS describing the fitness of each individual $\boldsymbol{\theta} \in \mathbf{S}$
for all $\boldsymbol{\theta} \in \mathbf{S}$ **do**
average QoS := **evalNetPerformance**($\boldsymbol{\theta}$, model)
end for
// Elitism based selection
select the best e solutions
// Crossover
number of crossover $n_c = (p_{\text{size}} - e)/2$
for $j = 1$ to n_c **do**
randomly select two solutions $\boldsymbol{\theta}^{(a)}$ and $\boldsymbol{\theta}^{(b)}$
generate $\boldsymbol{\theta}^{(c)}$ by arithmetic crossover to $\boldsymbol{\theta}^{(a)}$ and $\boldsymbol{\theta}^{(b)}$
end for
// Mutation
for $j = 1$ to n_c **do**
mutate each parameter of $\boldsymbol{\theta}^{(j)}$ under the rate δ and generate a new solution
end for
// The GA keeps on iterating until the new solution reaches the performance target.
end for
return the best solution $\boldsymbol{\theta}^{(*)}$

Algorithm 3: GA scheme.

should be reduced and to decide whether the price paid in terms of accuracy is acceptable.

(2) When we build ensemble methods, SVM and NN regression models perform better when they are bagged than when they are boosted. This was expected, as Bagging combines many weak predictors (i.e., the predictor is only slightly correlated with the true prediction) to produce a strong predictor (i.e., the predictor is well-correlated with the true prediction). This works well for algorithms where by changing the training set the output changes.

The opposite behaviour can be found in k-NN and DT regression models; that is, when these algorithms are boosted the models tend to provide better results than when they are bagged. That is, in order to improve the performance of AdaBoost, we use suboptimal values, for the number of the neighbours (k) for k-NN, and the number of trees (T) for DT; that is, we use values that are not that good, but at least better than random. Therefore, we make weak predictors by tuning the parameters to avoid the cases in which the regressors respond similarly [31]. This is not the case for SVM and NN. Since these learning algorithms do not have an input parameter that we can adjust to obtain a weak predictor without affecting the accuracy of the model, the probability that these algorithms provide better performance when they are boosted than when are bagged is lower. Some initial results about how a SVM can be used as a weak predictor can be found in [32]. Another option could be treating this kind of algorithm as a weak regressor by using fewer samples to train, as stated in [33].

(3) By applying different regression models, and in particular when the SVM regression model is bagged, we improve by 5% the overall accuracy of the prediction with respect to the k-NN model. More specifically, in terms of the NRMSE, k-NN exhibits an error of 10%, while SVM halves this value to 5%.

Results suggest that while all the regression models exhibit high accuracy, the bagged-SVM learning model is the one that better fits our needs and exhibits more accurate predictions. Therefore, in this work we focus on bagged-SVM to build the best model that fits the data.

3.2. Putting It Together: ML-Based Network Planning Tool. In summary, the proposed tool exploits the power of both components working together, namely, ML-based QoS modelling and GA-based network optimisation. The former is capable of extracting the most relevant information out of wealth of operational data measured in complex ultradense networks and make meaningful predictions for the operator and the end-user. This results in powerful network performance

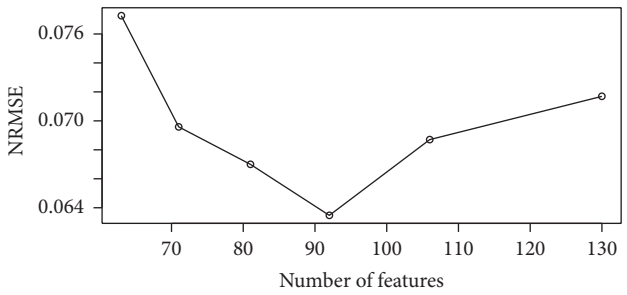


Figure 4: NRMSE, a function of different number of features selected by the SPCA.

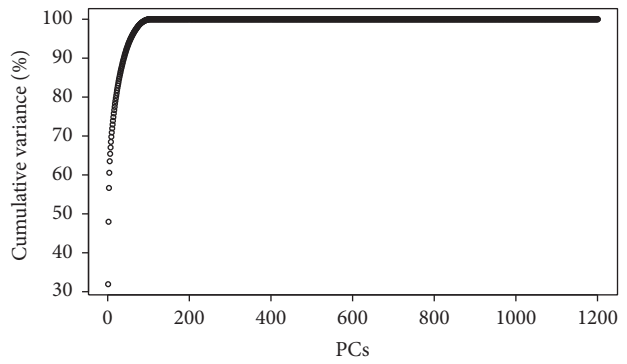


Figure 5: Cumulative contribution of each PC to the original data's variance.

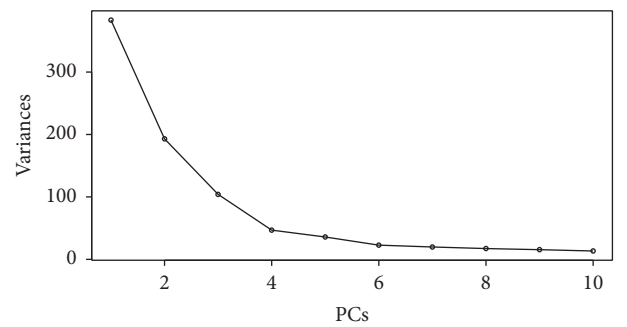


Figure 6: Variability of the data set as a function of the $c = 10$ PCs.

prediction models that are fed into the latter, which explores and finds the best combination of parameters for configuring the network elements, where *best* means the one that gives the best predicted QoS according to the learned model. The whole network planning process is depicted in Figure 7.

4. Performance Evaluation of the Network Planning Tool

In order to evaluate the performance of the proposed scheme, we consider two cases of study.

(A) Case Study #1: Deployment Planning in a Dense Small Cell Scenario. In this use case, we exploit the experience gained throughout the network to properly dimension and deploy (i.e., locate) the small cells in an indoor deployment. The goal is to improve the QoS offered to end-users while increasing the spectral efficiency by reducing the PRBs used per Mb transmitted.

(B) Case Study #2: Self-Healing to Compensate for Faults. Cell Outage Compensation (COC) is applied to alleviate the outage caused by the loss of service from a faulty cell. For this use case, an adequate reaction is vital for the continuity of the service. As a result, vendor specific Cell Outage Detection (COD) schemes have also to be designed [34–36]. In this case of study we assume the outage has already been detected, and we focus on readjusting the network planning (antenna tilt) to solve the outage problem.

The planning tool aims at guaranteeing that the network meets the operator's needs. We proceed to design the appropriate planning tool by defining the following aspects:

(1) Data: we define the radio measurements that we extract and analyse from the network.

(2) Parameter: we define the network parameters that we aim to tune.

(3) Action: we define the possible actions to take in order to optimise network performance.

(4) Objective: we define the system level target.

Table 3 presents this information for each use case under study.

The results of the use cases of application are described in the rest of the section. We first present the simulation scenario and then the simulation results for each use case.

(A) Case of Study #1: Deployment of Indoor Small Cells. We aim at providing a network planning of a small cell indoor deployment to improve the QoS offered to end-users and to increase the resource efficiency (in PRB per Mb) of our planning.

(1) Simulation Scenario. The scenario that we set up consists of 1 Enhanced Node Base station (eNB), with 3 sectors. We need to plan the deployment of the small cell network defined as the standard dual stripe scenario based on 1 block of 2 buildings [37]. The building has 1 floor, with 20 apartments, which results in 40 apartments, as depicted in Figure 8. We consider that 1 small cell is located in each apartment and the planning will decide which one will be switched ON or OFF. The parameters used in the simulations and the learning parameters are given in Table 4.

(2) Simulation Results. The simulation starts with an initial deployment, where each apartment of the building has randomly deployed a small cell. The Signal to Interference and Noise Ratio (SINR) at each point of the scenario, obtained through this initial deployment, is depicted in Figure 9. The idea is to determine the most effective number and location of small cells by evaluating the performance of each individual $\boldsymbol{\theta}_{sc}$ configuration. The configuration $\boldsymbol{\theta}_{sc}$ is represented by a binary string, with dimension N_{sc}. Each element in the binary

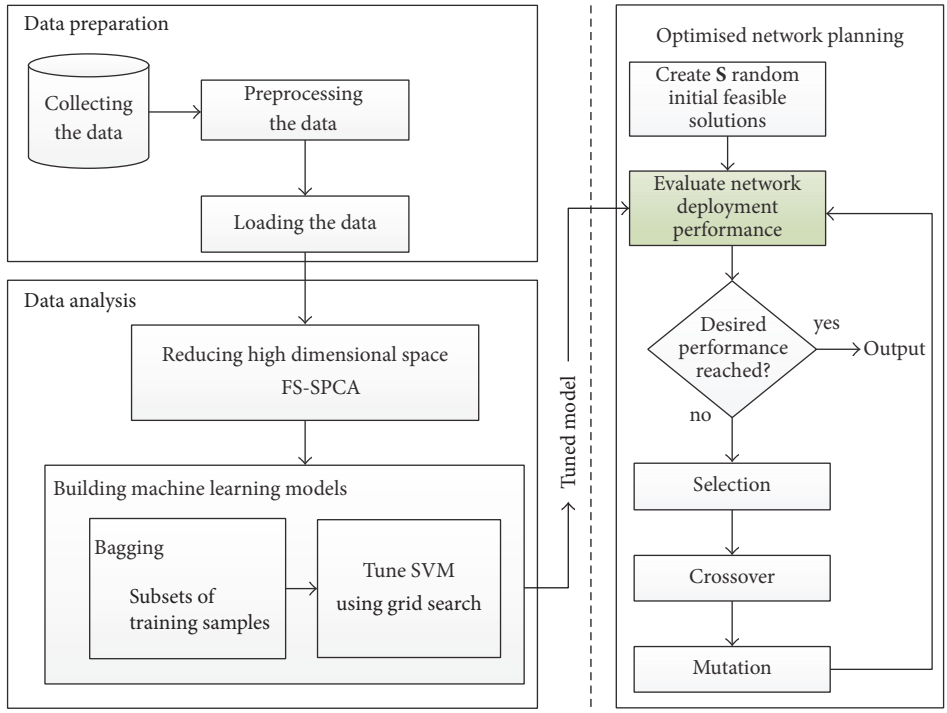


Figure 7: Bagged-SVM network planning tool architecture.

Table 3: Information relevant for the network planning tool.

Information	Case study #1	Case study #2
Data	RSRP and RSRQ coming from the serving and neighbouring cells	RSRP and RSRQ coming from the serving and neighbouring cells
Parameter	$\boldsymbol{\theta}_{sc} = (\theta_{sc_1}, \ldots, \theta_{sc_{N_{sc}}})$, which is a binary vector that denotes if the small cells are switched ON or OFF	$\boldsymbol{\theta}_{tilt} = (\theta_{tilt_1}, \ldots, \theta_{tilt_M})$, which is a vector that denotes the tilt value associated with each of the surrounding cells that are trying to fill the outage gap
Action	Switch ON or OFF each small cell	Adjusting the antenna tilt parameter
Objective	Increasing the resource efficiency of our planning by dimensioning the deployment and locating LTE indoor small cells	Readjusting the network planning to quickly solve an outage problem

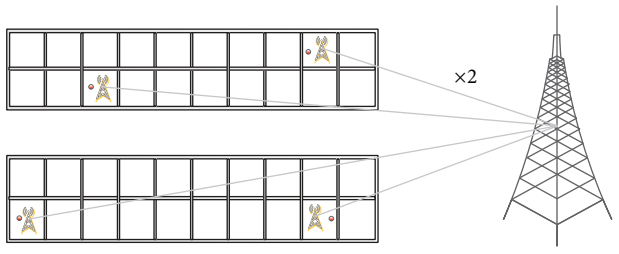


Figure 8: Scenario.

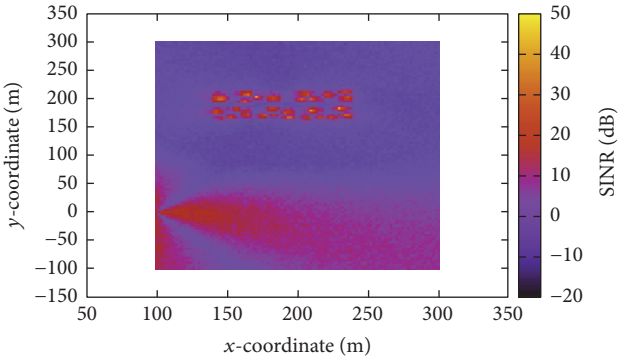


Figure 9: Initial deployment, in which we consider one small cell located in each apartment.

string represents if the jth small cell is ON or OFF. A value of 1 means that the power transmission of the jth small cell is set to 23 dBm, while a 0 means that the jth small cell is switched OFF. At each iteration (i.e., generation of the GA), the $GA_{\boldsymbol{\theta}_{sc}}$ process evaluates different strings, and when the evaluation is done, the $GA_{\boldsymbol{\theta}_{sc}}$ process provides a new configuration of small cells. Therefore, the proposed network planning tool takes advantage of the model generated through the data

Table 4: Case of study #1: simulation parameters.

Parameter	Value
Propagation loss model	HybridBuildings
Shadow fading	Log-normal, std = 8 dB
Scheduler	Proportional Fair (PF)
AMC model	LteAmc::MiErrorModel
Transport protocol	User Datagram Protocol (UDP)
Traffic model	Constant bit rate
Layer link protocol	Radio Link Control (RLC)
Mode	Unacknowledged Mode (UM)
Macro cell scenario	
Number of eNBs	1 site with 3 cells
eNB Tx power	46 dBm
Small cell scenario	
Initial number of small cells (N_{sc})	40
Small cell Tx power	23 dBm
LTE	
Bandwidth	5 MHz
Number of RBs	25
TTI	1 ms
GA	
Type	binary-valued
Size of **S** (*psize*)	100
Elitism e	2
Mutation change δ	0.1
Bagged-SVM	
Number of iteration (γ)	1000
Epsilon ϵ	0.1
Kernel	Radial Basis Function (RBF)
Simulation time	0.25 s

preparation and data analysis processes to estimate the QoS at any random point in the scenario given a certain $\boldsymbol{\theta}_{sc}$. Then, it learns online through the *optimised network planning* tool the $\boldsymbol{\theta}_{sc}^{(*)}$ best configuration to the small cell deployment by improving the configuration with each new generation (see Figure 10). This process is referred hereafter as $GA_{\boldsymbol{\theta}_{sc}}$.

We evaluate the system performance of each deployment, represented by a binary string of dimension N_{sc}, on the ns-3 LENA module. When we get a new configuration from the $GA_{\boldsymbol{\theta}_{sc}}$, we implement it in the system simulator and evaluate it again, until the network deployment reaches the network performance target set by the operator.

Figures 11 and 12 show the fitness of the best individuals found in each generation (Best) and the mean of the fitness values across the entire population (Mean), in terms of PRB/Mb and Average Throughput, respectively. Figure 11 depicts the time evolution of the average PRB per Mb in the scenario. We observe that as the generations proceed and the

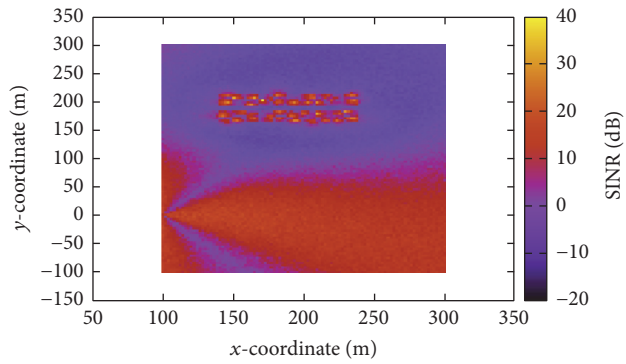


Figure 10: Final deployment, in which the $GA_{\boldsymbol{\theta}_{sc}}$ scheme finds the number of small cells to deploy.

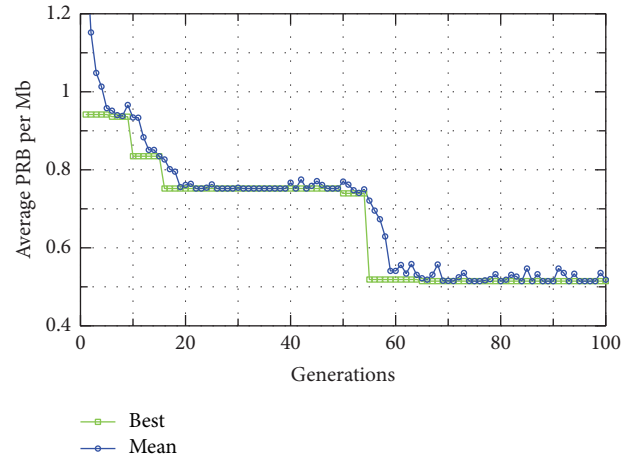


Figure 11: Evolution of the average PRB per Mb.

$GA_{\boldsymbol{\theta}_{sc}}$ evolves, the PRB/Mb decreases as it was expected, and so the efficiency of the planning increases because the average throughout increases and the PRBs consumed decrease. That is, at the 60th generation, the $GA_{\boldsymbol{\theta}_{sc}}$ reaches the best spectral efficiency for the planning. Figure 12 shows the evolution of the $GA_{\boldsymbol{\theta}_{sc}}$ scheme in terms of the average throughput in the whole network. We observe that for each new generation the $GA_{\boldsymbol{\theta}_{sc}}$ scheme finds the number and location of small cells that maximise the throughput. Figure 10 shows the resulting SINR map of the deployment. It can be seen that the SINR in the represented region of the network is also globally improved.

A comparison of different planning schemes is shown in Figure 13. This figure shows the SINR performance for three deployments: (1) one where all the small cells are ON and there is a small cell in each apartment (tagged as *initial deployment*), (2) the proposed $GA_{\boldsymbol{\theta}_{sc}}$ approach (tagged as *final deployment*), which results in 28 deployed small cells, and (3) a benchmark deployment based on a greedy algorithm. The greedy algorithm searches for the best deployment by testing a certain number m_{subset} of string vector configurations, defining the number and position of switched ON nodes. For each configuration, the greedy algorithm computes the QoS and then it selects the configuration that provides the

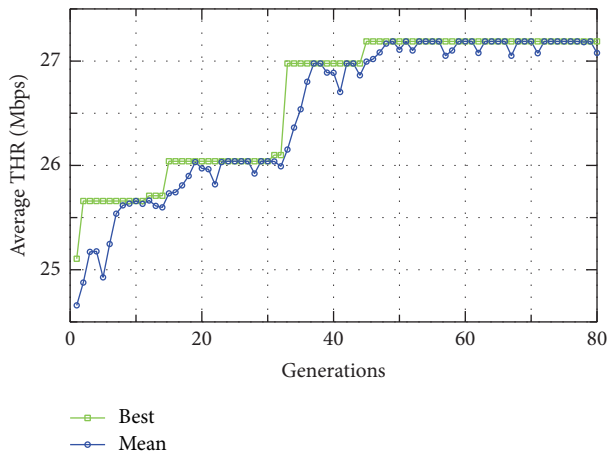


Figure 12: Evolution of the average throughput in the whole network.

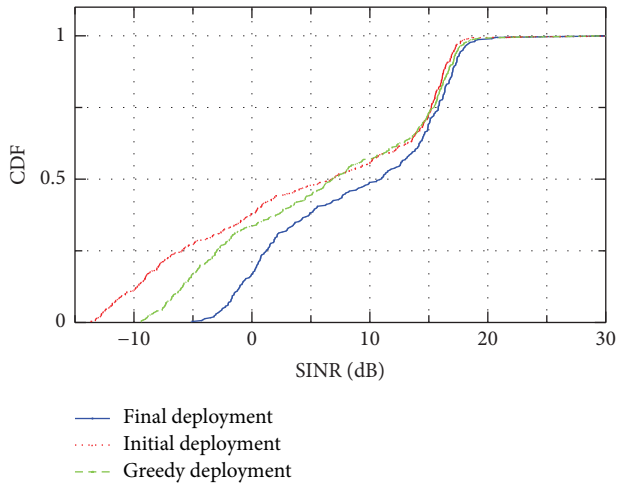


Figure 13: CDF of the SINR (dB) in the building.

best QoS results. The argument m_{subset} is discussed in [38]. It offers a tradeoff between a high number of configurations, which would result in high computational effort, and low number of configurations, which may result in local optima. We have tested different values against performances and we finally selected the $m_{\text{subset}} = 10000$. While the greedy search algorithm rarely outputs optimal solutions, it often provides reasonable suboptimal solutions [39]. This can be observed in Figure 13, where the number of small cells in the initial deployment is 40, in the greedy deployment it is 33, and when applying our $\text{GA}_{\boldsymbol{\theta}_{\text{sc}}}$ scheme it is 28. We observe in this figure that, in general, the $\text{GA}_{\boldsymbol{\theta}_{\text{sc}}}$ tends to work more effectively, since it provides improved QoS while deploying a reduced number of small cells. The reason for this is that the GA approach makes a much deeper estimation of the state of the environment through the regression analysis and counts on a more sophisticated combinatorial search scheme, based on the genetic approach, than the greedy scheme, which performs a limited search and may be trapped at local optima.

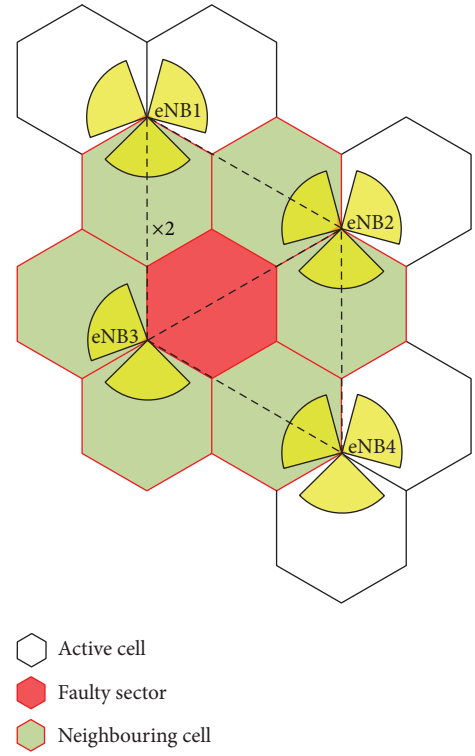


Figure 14: Scenario.

(B) Case of Study #2: Self-Healing to Compensate for Faults. As we already mentioned, in this case of application, we focus on readjusting the network planning to quickly solve an outage problem by adjusting the antenna tilt parameter. To evaluate this, we generate a sector fault in a typical macrocellular deployment. That is, during a certain period of time, a sector is not able to offer service to its users. Therefore, the proposed network planning tool allows setting the antenna tilt for each compensating sector to automatically alleviate the outage caused by the loss of service from a faulty sector [40].

(1) Simulation Scenario. We consider a Long Term Evolution (LTE) cellular network composed of a set of $\mathscr{M}$ eNBs. The $\mathscr{M}$ eNBs form a regular hexagonal network layout with intersite distance D and provide coverage over the entire network. We assume that a sector in the scenario is down (see Figure 14). The parameters to tune are the antenna tilts of the cells neighbouring the affected sector. In particular, the surrounding M cells automatically and continuously adjust their antenna tilt until the coverage gap is filled. The vertical radiation pattern of a cell sector is obtained according to [41], where the gain in the horizontal plane is given by

$$G_h(\phi) = \max\left\{-12\left(\frac{\phi}{\text{HPBW}_h}\right), \text{SLL}_h\right\}. \quad (2)$$

Here ϕ is the horizontal angle relative to the maximum gain direction, HPBW_h is the half power beam-width for the horizontal plane, and SLL_h is the side lobe level for the

horizontal plane. Similarly, the gain in the vertical plane is given by

$$G_v(\rho) = \max\left\{-12\left(\frac{(\rho - \theta_{\text{tilt}})}{\text{HPBW}_v}\right), \text{SLL}_v\right\}. \quad (3)$$

All eNBs in the scenario have the same antenna model where ρ is the vertical angle relative to the maximum gain direction, θ_{tilt} is the tilt angle, HPBW_v is the vertical half power beamwidth, and SLL_h is the vertical side lobe level. Finally, the two gain components are added by

$$G(\phi, \rho) = \max\{G_h(\phi) + G_v(\rho), \text{SLL}_0\} + G_0, \quad (4)$$

where SLL_0 is an overall side lobe floor and G_0 the antenna gain. We consider a cellular network whose system performance has been evaluated on the 3GPP-compliant ns-3 LENA module. The parameters used in the simulations and the learning parameters are given in Table 5. The macrocell scenario that we set up consists of 4 eNB with 3 sectors each, which results in 12 cells, as depicted in Figure 14.

Therefore, each feasible solution corresponds to a vector of $M = 6$ tilt values (i.e., each tilt value is associated with one of the surrounding cells that are trying to fill the outage gap).

(2) Simulation Results. We analyse performance results obtained through the network planning tool described in Section 2. Figures 15–17 show the fitness of the best individual found in each generation (Best) and the mean of the fitness values across the entire population (Mean). We observe that, for each generation, the population (i.e., the possible combinations of antenna tilts) tends to get better as generations proceed. Figure 15 depicts the time evolution of the PRB per offered Mb in the scenario during the evolution of the $\text{GA}_{\boldsymbol{\theta}_{\text{tilt}}}$. We observe that as the $\text{GA}_{\boldsymbol{\theta}_{\text{tilt}}}$ evolves, the efficiency of the planning increases. That is, as the generations proceed, the $\text{GA}_{\boldsymbol{\theta}_{\text{tilt}}}$ finds the configuration parameter vector that minimizes the value of PRBs per transmitted Mb.

In order to analyse the overall impact of the outage and its evolution for each new generation, we depict the average throughput for neighbouring cells only (Figure 16) and for the whole network (Figure 17).

More specifically, Figure 16 depicts the time evolution of the average throughput of the compensating sectors. That is, it shows the performance in terms of the throughput of the 6 neighbouring cells, which adjust their antenna tilt in order to compensate for the faulty sector. Figure 17 describes the average throughput of the whole network. From this figure, we observe that the $\text{GA}_{\boldsymbol{\theta}_{\text{tilt}}}$ scheme achieves at the 30-th generation 23 Mbps of average throughput in the whole network, while, in the neighbouring cells (Figure 16), the achieved throughput is only 15 Mbps, which is reasonable, due to the challenging service conditions in this area.

Finally, the $\text{GA}_{\boldsymbol{\theta}_{\text{tilt}}}$ scheme is compared in Figure 18 with the self-organised Reinforcement Learning- (RL-) based approach for COC proposed in [42], where in order to design the self-healing solution ($\text{AC}_{\theta_{\text{tilt}}}$), the antenna tilt is adjusted. The considered RL approach is an Actor Critic algorithm, which is already proven to outperform different solutions for

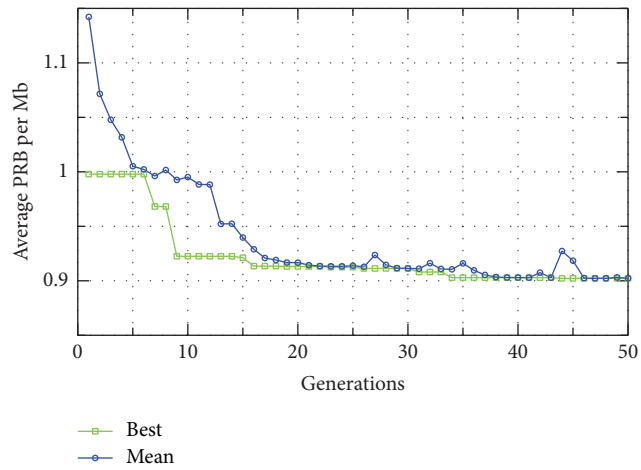


Figure 15: Evolution of the average PRB per Mb.

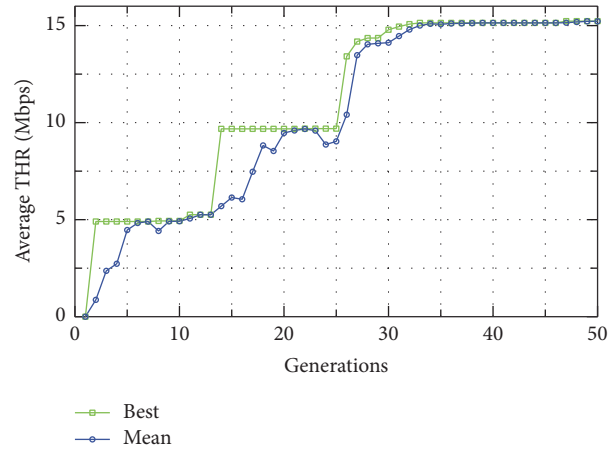


Figure 16: Evolution of the average throughput in the neighbouring cells.

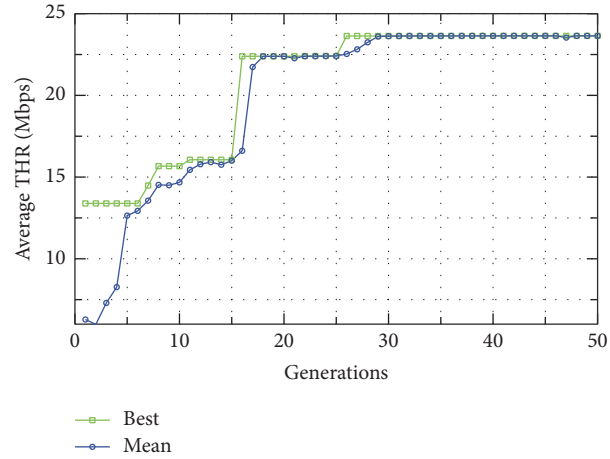


Figure 17: Evolution of the average throughput in the whole network.

Table 5: Case of study #2: Simulation parameters.

Parameter	Value
Propagation loss model	HybridBuildings
Shadow fading	Log-normal, std = 8 dB
Scheduler	PF
AMC model	LteAmc::MiErrorModel
Transport protocol	UDP
Traffic model	Constant bit rate
Layer link protocol-mode	RLC-UM
Macro cell scenario	
$\mathscr{M}$ MacroEnbSites	4 (each site has 3 cells)
Number of cells	12
eNB Tx power	46 dBm
LTE	
Intersite distance D	500 m
Bandwidth	5 MHz
Number of RBs	25
Antenna parameters	
Horizontal angle ϕ	$-180^\circ \leq \phi \leq 180^\circ$
HPBW	Vertical 10° : Horizontal 70°
Antenna gain G_0	18 dBi
Vertical angle θ	$-90^\circ \leq \phi \leq 90^\circ$
SLL	Vertical −18 dB : Horizontal −20 dB
SLL_0	−30 dB
GA	
Type	real-valued
Tilt values (min–max)	0°–15°
Size of $\mathbf{S}$ (*psize*)	100
Elitism e	2
Mutation change δ	0.1
Bagged-SVM	
Number of iteration (γ)	1000
Epsilon ϵ	0.1
Kernel	RBF
Simulation time	0.25 s

COC available in literature in [42]. Therefore, we consider this to be a good benchmark for comparison.

Figure 18 depicts the CDF of the SINR of the network. We assume the user is out of outage when its SINR is above the threshold of −6 dB, as explained in [42]. Therefore, in this figure, we observe that $AC_{\theta_{tilt}}$ is able to recover 95% of UE, while the $GA_{\boldsymbol{\theta}_{tilt}}$ is able to recover all the UE. We observe in this figure that, in general, the $GA_{\boldsymbol{\theta}_{tilt}}$ tends to work effectively, since it maintains the diversity in the values taken by the parameters during the generations and it makes a much deeper estimation of the state of the environment through the regression analysis approach than the $AC_{\theta_{tilt}}$ scheme, which considers only the SINR and CQI feedback from the UE to determine the state of the environment and to estimate the general behaviour of the network.

From these figures, we observe that using the planning tool to adjust the antenna tilt parameter, we are able to alleviate the outage caused by the loss of service from a faulty sector.

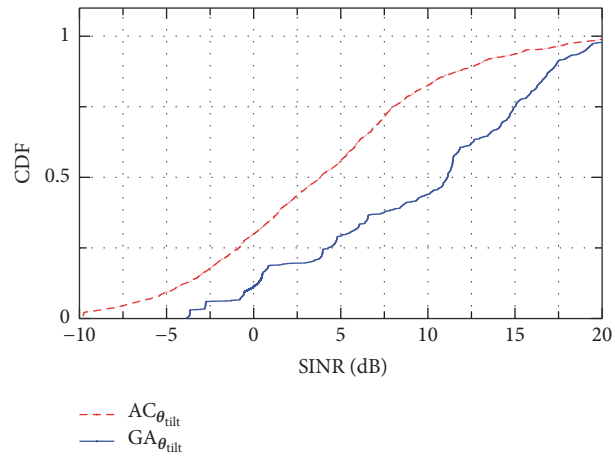


Figure 18: CDF of the average SINR values in the faulty sector by two different COC approaches. The $GA_{\boldsymbol{\theta}_{tilt}}$ based scheme is compared to a RL approach presented in [16].

5. Concluding Remarks

In this paper, we have defined a methodology and built a tool for smart and efficient network planning that exploits and learns from the operational history reflected in the measurements gathered anywhere and at any time throughout the network. To build the QoS prediction model based on this historical data, we collect UE measurements according to 3GPP MDT functionality and we apply regression analysis techniques to estimate the Physical Resource Blocks (PRB) per transmitted Mb. This model is then used as objective function in a genetic algorithm ($GA_{\boldsymbol{\theta}}$). By doing this, one can observe that each new iteration increases the efficiency of resource usage in the network while improving the throughput offered to the user.

To demonstrate the flexibility of the proposed smart ML-based planning tool, in this paper, we have decided to apply it to two very different use cases and scenarios: (1) deployment of indoor small cells and (2) compensation of sector faults in a traditional macrocell deployment. For use case #1, results demonstrate the ability of the proposed scheme to deploy small cells in a network in such a way that the average throughput is increased while the PRBs consumed per Mb transmitted decrease, hence improving the spectral efficiency. Regarding use case #2, results demonstrate the ability of the proposed scheme to compensate 100% of outage users in the scenario and to offer them service. We have compared the performance of our approach in the context of the two proposed cases of study to state-of-the-art solutions based on a greedy algorithm for case 1 and reinforcement learning for case 2. Our scheme outperformed state-of-the-art solutions in both cases. We believe that the same technique can be successfully applied to many other planning problems of interest for operators.

As a future work, we will analyse the planning of scenarios where a huge number of devices have to be provided with service. This is expected to benefit the performance of the estimation, since the data base will be enriched by many measurements, which is the basis of the intelligence of the approach.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The research leading to these results has received funding from the Spanish Ministry of Economy and Competitiveness FPI Research Programme (BES-2011-047309) under the Grant TEC2010-21100. The work of J. Moysen is also funded by 5GNORM Project (TEC2011-29700-C02-01).

References

[1] 3GPP on track to 5G, http://www.3gpp.org/ news-events/3gpp-news/1787-ontrack 5g.

[2] 5GPPP. The 5G Infrastructure Public Private Partnership, https://5g-ppp.eu/.

[3] N. Baldo, L. Giupponi, and J. Mangues-Bafalluy, "Big data empowered self organized networks," in *Proceedings of the 20th European Wireless Conference (EW '14)*, pp. 181–188, Barcelona, Spain, May 2014.

[4] B. Romanous, N. Bitar, A. Imran, and H. Refai, "Network densification: challenges and opportunities in enabling 5G," in *Proceedings of the IEEE 20th International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (CAMAD '15)*, Guildford, United Kingdom, September 2015.

[5] Riverbed OPNET NetOne, https://support.riverbed.com/content/support/software/steelcentral-npm/net-planner.html.

[6] Cariden MATE Design, Adquired by CISCO, http://www.cisco.com/c/dam/en/us/td/docs/net_mgmt/wae/6-1/design/user/guide/MATE_Design_User_Guide.pdf.

[7] RSoft Design Group. Metrowand, https://optics.synopsys.com/rsoft/pdfs/RSoftProductCatalog.pdf.

[8] CelPlan. CelTrace TM Wireless Global Solutions, http://www.celplan.com/products/indoor/celtrace.asp.

[9] Net2Plan, "The open-source network planner," http://www.net2plan.com/publications.php.

[10] F. Chernogorov and T. Nihtilä, "QoS verification for minimization of drive tests in LTE networks," in *Proceedings of the IEEE 75th Vehicular Technology Conference (VTC '12)*, pp. 6–9, IEEE, Yokohama, Japan, May 2012.

[11] F. Chernogorov and J. Puttonen, "User satisfaction classification for minimization of drive tests QoS verification," in *Proceedings of the IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC '13)*, pp. 2165–2169, London, UK, September 2013.

[12] J. Moysen, N. Baldo, L. Giupponi, and J. Mangues-Bafalluy, "Predicting QoS in LTE HetNets based on location-independent UE measurement," in *Proceedings of the 20th IEEE International Workshop on Computer Aided Modelling and Design of Communication Links and Networks*, Guildford, UK, 2015.

[13] J. Moysen, L. Giupponi, and J. Mangues-Bafalluy, "On the potential of ensemble regression techniques for future mobile network planning," in *Proceedings of the IEEE Symposium on Computers and Communication (ISCC '16)*, pp. 477–483, IEEE, Messina, Italy, June 2016.

[14] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Longman, Boston, Mass, USA, 1989.

[15] 3GPP, "Radio performance and protocol aspects (system)—RF parameters and BS conformance," Tech. Rep. TSG RAN WG4 R4-092042, 2009.

[16] J. Moysen and L. Giupponi, "A reinforcement learning based solution for self-healing in LTE networks," in *Proceedings of the 80th IEEE Vehicular Technology Conference (VTC '14-Fall)*, September 2014.

[17] M. Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, Cambridge, Mass, USA, 1996.

[18] J. Turkka, F. Chernogorov, K. Brigatti, T. Ristaniemi, and J. Lempiäinen, "An approach for network outage detection from drive-testing databases," *Journal of Computer Networks and Communications*, vol. 2012, Article ID 163184, 13 pages, 2012.

[19] F. Chernogorov, S. Chernov, K. Brigatti, and T. Ristaniemi, *Data Mining Approach to Detection of Random Access Sleeping Cell Failures in Cellular Mobile Networks*, Computer Science, Networking and Internet Architecture, 2015.

[20] J. Johansson, W. A. Hapsari, S. Kelley, and G. Bodog, "Minimization of drive tests in 3GPP (Release 11)," *IEEE Communications Magazine*, 2012.

[21] S. T. Roweis, *EM Algorithms for PCA and SPCA*, Advances in Neural Information Processing Systems, The MIT Press, 1998.

[22] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[23] M. C. Bishop, *Pattern Recognition and Machine Learning*, Busines Dia, Llc. Springer Science, 2006.

[24] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[25] J. R. Quinlan, *Induction of Decision Trees. Machine Learning*, Kluwer Academic, 1986.

[26] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theory and Applications*, World Scientific, 2008.

[27] D. Opitz and R. Maclin, "Popular ensemble methods: an empirical study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, 1999.

[28] L. Rokach, "Ensemble-based classifiers," in *Artificial Intelligence*, pp. 1–39, 2010.

[29] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization," *Machine Learning*, vol. 40, no. 2, pp. 139–157, 2000.

[30] T. G. Dietterich, *Machine-Learning Research: Four Current Directions*, American Association for Artificial Intelligence, 1997.

[31] X. Li, L. Wang, and E. Sung, "AdaBoost with SVM-based component classifiers," *Engineering Applications of Artificial Intelligence*, vol. 21, no. 5, pp. 785–795, 2008.

[32] E. Mayhua-Lopez, V. Gomez-Verdejo, and A. R. Figueiras-Vidal, *Boosting Ensembles with Subsampled LPSVM Learners*, Universidad Carlos III de Madrid (UC3M), 2013.

[33] E. García and F. Lozano, "Boosting Support Vector Machines," in *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition*, pp. 153–167, Leipzig, Germany, July 2007.

[34] O. Onireti, A. Zoha, J. Moysen et al., "A cell outage management framework for dense heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2097–2113, 2016.

[35] E. J. Khatib, R. Barco, P. Munoz, I. D. La Bandera, and I. Serrano, "Self-healing in mobile networks with big data," *IEEE Communications Magazine*, vol. 54, no. 1, pp. 114–120, 2016.

[36] A. Gómez-Andrades, P. Muñoz, I. Serrano, and R. Barco, "Automatic root cause analysis for LTE networks based on unsupervised techniques," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2369–2386, 2016.

[37] 3GPP, "Technical specification group radio access network; Evolved universal terrestrial radio access (E-UTRA); Further advancements for E-UTRA physical layer aspects," Tech. Rep. TR 36.814, 2010.

[38] P. Romanski and L. Kotthoff, "Package FSelector," https://cran.r-project.org/web/packages/FSelector/FSelector.pdf.

[39] M. Hazewinkel, "Preface," *Discrete Mathematics*, vol. 227-228, pp. 1–4, 2001.

[40] J. Moysen, L. Giupponi, and J. Mangues-Bafalluy, "A machine learning enabled network planning tool," in *Proceedings of the 27th IEEE Personal Indoor and Mobile Radio Communications (PIMRC '16)*, Valencia, Spain, 2016.

[41] F. Gunnarsson, M. N. Johansson, A. Furuskar et al., "Downtilted base station antennas—a simulation model proposal and impact on HSPA and LTE performance," in *Proceedings of the 68th IEEE Vehicular Technology Conference (VTC Fall '08)*, pp. 1–5, Calgary, Canada, September 2008.

[42] J. Moysen and L. Giupponi, "A reinforcement learning based solution for self-healing in LTE networks," in *Proceedings of the 80th IEEE Vehicular Technology Conference (VTC Fall '14)*, pp. 1–6, IEEE, Vancouver, Canada, September 2014.

Hindawi
Mobile Information Systems
Volume 2017, Article ID 5816862, 12 pages
http://dx.doi.org/10.1155/2017/5816862

Research Article

Energy Efficiency and Capacity Tradeoff in Cloud Radio Access Network of High-Speed Railways

Shichao Li,[1,2] Gang Zhu,[1,2] Siyu Lin,[1,2] Qian Gao,[1,2] Lei Xiong,[1] Weiliang Xie,[3] and Xiaoyu Qiao[3]

[1] *State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China*
[2] *School of Electronic Information Engineering, Beijing Jiaotong University, Beijing 100044, China*
[3] *Technology Innovation Center, China Telecom, Beijing 100000, China*

Correspondence should be addressed to Siyu Lin; sylin@bjtu.edu.cn

Received 29 July 2016; Revised 23 November 2016; Accepted 18 December 2016; Published 9 January 2017

Academic Editor: Piotr Zwierzykowski

Copyright © 2017 Shichao Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

To meet the increasing demand of high-data-rate services of high-speed railway (HSR) passengers, cloud radio access network (C-RAN) is proposed. This paper investigates the tradeoff between energy efficiency (EE) performance and capacity in C-RAN of HSR. Considering that the train location can be predicted, we propose a predictable path loss based time domain power allocation method (PPTPA) to improve EE performance of HSR communication system. First, we consider that the communication system of HSR only bears the passenger information services (PISs). The energy-efficient power allocation problem with delay constraint is studied. The formulated problem is nonconvex. To deal with it, an equivalent convex problem is reformulated. Based on PPTPA, we propose an iterative algorithm to improve the EE performance. Second, we consider that the PISs and the train control services (TCSs) are all bore. A capacity optimization problem with joint EE and services transmission delay constraints is formulated. Based on PPTPA, we propose a hybrid power allocation scheme to improve the capacity of the system. Finally, we analyze the effect of small-scale fading on EE performance. The effectiveness of the proposed power allocation algorithm is validated by HSR channel measurement trace based emulation results and extensive simulation results.

1. Introduction

In the past couple of years, high-speed railways (HSR) are expanding rapidly all over the world. More than 16,000 km HSR are deployed in China, which accounts for more than 60 percent of operation length of HSR around the world. The railway communication system plays a key role in HSR to bear the train control services (TCSs) and passenger information services (PISs) [1, 2].

Although current mobile communication technologies can guarantee the safety of HSR as bearing the TCSs, the relatively low transmission data rate (e.g., 2–4 Mbps) cannot provide satisfied quality of experience (QoE) of passengers. Therefore, the mobile communication system with distributed antennas is implemented in HSR scenario to improve the passengers' QoE [3]. Distributed antenna system (DAS) is consisted with radio remote units (RRUs) and base stations (BSs), but the radio resource cannot be shared between the RRUs, which limits the centralized processing gain of DAS. To mitigate this issue, cloud radio access network (C-RAN) is proposed based on DAS structure [4, 5], in which a baseband unit (BBU) pool is instead of the distributed deployed BSs to handle the complex computational tasks. The centralized signal processing structure in the BBU pool has much more advantages than distributed signal processing in separate BSs, including saving the power, reducing the operating expenditure, and improving hardware utilization ratio. C-RAN changes the static relationship between BBU and RRUs, which is by now recognized as a good solution to provide high-speed data services to meet the increasing demand of high-data-rate services of HSR passengers [6, 7]. In addition, C-RAN can reduce the handover frequency of the train. Due to the high movement speed of the trains, the handover happens frequently. C-RAN can provide extensive radio coverage via super cell to limit the handover frequency

and the BBU pool can eliminate interference to improve the coverage quality [3].

As energy consumption for communication technologies has been growing rapidly, a large number of greenhouse gases are discharged [8]. Recent reports suggest that 2% of the world-wide CO_2 is discharged by communication infrastructures and 3% of the world-wide energy is consumed [9]. For the rail transportation system, around 50% of the energy is consumed by trains, and the rest is used by infrastructure facilities (stations, communication equipment, groundwater pumps, tunnel lighting, etc.) to ensure the proper system operation [10, 11]. As we know, HSR is a kind of green transportation system with low energy consumption, in order to save energy, all the infrastructure facilities need to improve EE performance. For the communication equipment of HSR, the BSs are deployed densely along the railway line (2 base stations are deployed each 3 kilometers). For example, there are about 900 BSs in Beijing-Shanghai HSR line. Therefore, how to utilize the limited energy to meet the needs of growing communication traffic is a significant problem [12–15].

Resource allocation in the C-RAN has attracted considerable attention in recent years. To maximize the sum capacity in C-RAN system, Zhou and Yu adapted Wyner-Ziv coding method to enhance the performance of C-RAN [16]. To reduce the energy consumption of C-RAN, a joint power minimization and RRUs selection problem were formulated [17]. Comparing with saving energy, how to improve the capacity with unit power is a more practical problem, so EE is an important performance metric in wireless communication system design. There are several EE performance-improving methods in C-RAN, such as joint distributed compression in the uplink [18] and joint power and subcarriers allocation in the downlink [19].

To pursue the EE maximization, the system capacity may degrade dramatically [20], so the tradeoff between capacity and EE is worth investigating. An effective EE constrained rate optimal power allocation policy for Nakagami-m channels was proposed in [20]. To maximize the EE performance in a distributed antenna system subject to the users' quality of service (QoS), backhaul capacity and antenna transmission power, a joint antenna, subcarrier, and power allocation method were proposed [20]. A joint antenna, subcarrier, and power allocation method was proposed [21]. In order to improve the EE performance of data center networks, a mechanism via the elastic multicontroller software defined networks was proposed [22]. Wu et al. proposed a method to improve the EE performance when considering multiple users harvest energy from a power station and then communicate with an information station in a time-division manner [23]. A convex relaxation and global optimization method was proposed to improve the EE performance in multiuser multicarrier broadband wireless systems [24]. There are also some works about the balance between the capacity and EE performance; for example, Ng et al. proposed an iterative resource allocation policy which considered the tradeoff between network capacity, EE performance, and backhaul capacity in multicell networks [25]. All the previous works are designed for conventional cellular networks.

Since C-RAN can provide various high-speed wireless services for users, it can be deployed along the high-speed lines to improve the QoE of passengers. As HSR is a kind of green transportation system with low energy consumption, the energy consumption of mobile communication system in HSR also needs to be considered [26]. All the existing works about EE optimization of C-RAN are investigated in the conventional cellular networks. Therefore, how to improve the EE performance in HSR considering the HSR characteristic is an practical interesting problem. Generally, resource allocation requires the accurate channel state information (CSI). Comparing with the conventional cellular networks, it is hard to estimate the accurate CSI in HSR scenario. However, due to the line of sight (LoS) scenario and predicted location of train in HSR system, the CSI can be simplified as predicted path loss information, which can be estimated via location information of the train.

There are two major concerns in the EE optimization in HSR. First, the QoE requirements of passengers should be considered. For the C-RAN structure, complex computational tasks are done by Virtual Machines (VMs) in BBU pool, which introduces extra latency to the service transmission. To achieve the delay and transmission rate requirements of passengers, the processing time of VMs should be considered in the EE optimization of C-RAN. Second, the communication system of HSR bears not only the PISs but also the TCSs. To pursue the EE maximization, the system capacity may degrade dramatically [20], which is inconsistent with the high system capacity requirement. Therefore, the tradeoff between EE performance and capacity should be considered.

In this paper, considering the train location can be predicted, we propose a predictable path loss based time domain power allocation method (PPTPA) to improve EE performance of HSR communication system. First, we only consider the communication system of HSR bearing the PISs. We focus on the EE optimization problem with delay constraint of C-RAN in the HSR scenario. The VMs processing latency in the BBU pool and radio transmission delay of services are considered in the proposed power allocation scheme to maximize EE performance in C-RAN. Since the EE maximization formulation is a nonconvex problem, we reformulate the objective function. Based on PPTPA, we propose an iterative algorithm to improve the EE performance. Second, we consider the communication system of HSR bearing the PISs and the TCSs. A capacity optimization problem subject to joint EE requirement and services transmission delay constraints is formulated. Based on PPTPA, we propose a hybrid power allocation scheme to improve the capacity of the system. Finally, the effects of small-scale fading on the proposed power allocation scheme performance are evaluated. Zhengzhou-Xian HSR line channel measurements trace based emulation results and extensive simulation results are provided to validate that the proposed two power allocation policies can meet the QoE requirements of PISs and TCSs.

The rest of the paper is organized as follows. The system model is described in Section 2. In Section 3, we analyze power allocation problem with services transmission delay constraint and utilize PPTPA to maximize EE performance. In Section 4, we investigate the optimal power allocation to

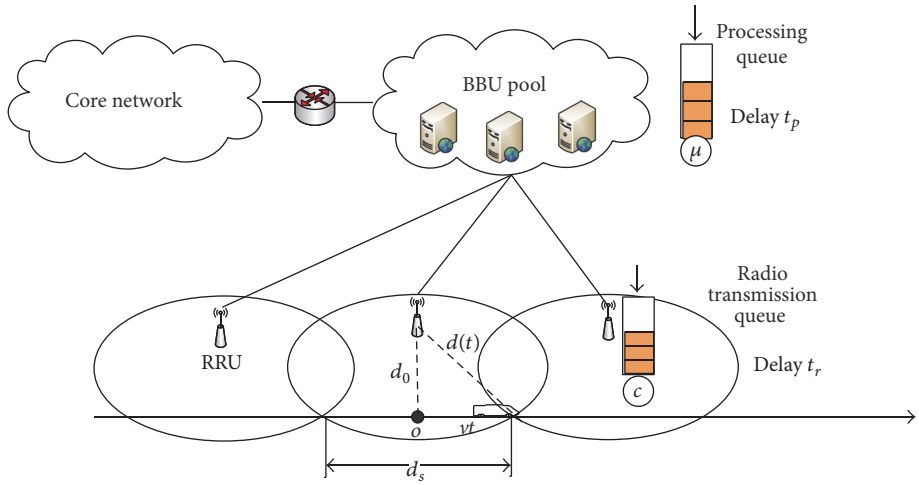


Figure 1: System model.

maximize capacity with joint services transmission delay and EE constraints. In Section 5, emulation results and extensive simulation results are provided to justify our analysis. Conclusion is given in Section 6.

2. System Model

The C-RAN under consideration consists of a set of RRUs and a BBU pool as shown in Figure 1. The RRUs are deployed along the railway line to provide the radio coverage. In this work, we only focus on the downlink of C-RAN, as it is always the bottleneck of the mobile communication system. The downlink traffic transmission can be divided into two phases, one is data preprocessing in the BBU pool and the other is radio transmission by the RRUs. The two phases can be modeled as two queue models, respectively [27]. One is the service data preprocessing queue, in which the data packets are preprocessed (such as encoded) by the VMs of BBU pool. The processing rate of VMs in BBU pool is denoted as μ, which can be considered as the service rate in the service data preprocessing queue. The other queue is the radio transmission queue, in which the service data are transmitted from the RRUs to the passengers packet by packet. The radio transmission rate is denoted as c, which can be considered as service rate in the radio transmission queue. The BBU pool connects RRUs via fiber links, so the transmission delay of these links can be ignored. We assume that the service arrival process from core network is a Poisson process with mean rate λ. The service time in the preprocessing queue follows exponential distribution with mean $1/\mu$ and the service time of the radio transmission queue follows exponential distribution with mean $1/c$.

In HSR scenario, most of the services are sensitive to delay, that is, video on demand and voice services. Therefore, delay is a key metric to measure QoE. The transmission delay of the traffic in the C-RAN should be less than the QoE requirement of passengers. In this system, $t_p = 1/(\mu - \lambda)$ is denoted as the data processing delay and $t_r = 1/(c - \lambda)$ is denoted as the radio transmission delay. t_t is the expected delay in this system, which can be expressed as $t_t = t_p + t_r = 1/(\mu - \lambda) + 1/(c - \lambda)$, for $\mu > \lambda, c > \lambda$. The transmission delay requirement can be described as

$$t_t \leq \tau, \tag{1}$$

where τ is the transmission delay requirement of passengers. If the service is HTTP/web (emails), it can be represented as a new service with different QoE requirements. As the HTTP/web (emails) requires low data rate and loose delay, we can adjust the delay constraint of QoE requirement.

The coverage model of RRUs is shown in Figure 1, the distance between RRU and railway line is d_0, and the coverage diameter of each RRU is d_s. The train speed is v and the period of the train crossing one cell covered by RRU is $T = d_s/v$. RRUs can eliminate the interinterference by BBU pool, so we do not consider the interference between RRUs [28].

3. EE Maximization Problem

Due to the rapid increase of the operating cost of wireless communication system and carbon emission, the EE performance has become a main issue for the design of future wireless communication system [26]. In this section, we formulate the EE maximization with QoE provisioning problem firstly. As the objective function is a nonlinear fractional form, the objective function is reformulated, and then an iterative algorithm based on PPTPA is proposed to solve this problem.

3.1. Problem Formulation. To guarantee the QoE requirement of passengers, the transmission delay should be satisfied. Constant power allocation policy cannot meet the transmission delay requirement due to the varying fading channel [29]. Therefore, we adopt variable transmit power allocation policy. In this section, our objective is to maximize EE performance, and the EE performance can be defined as C/P bits/joule [3].

Let $t = 0$, when the train arrives at the point O as shown in Figure 1, $h(t)$ denotes the channel gain at time t. In HSR

scenario, it is difficult to get the accurate CSI of the train. However, due to the moving track of the train being certain, the predicted path loss information instead of accurate CSI is used for resource allocation. The path loss is described as

$$h(t) = \frac{1}{\left(d_0^2 + (vt)^2\right)^{\alpha/2}} + \Delta, \quad -\frac{d_s}{2v} \le t \le \frac{d_s}{2v}, \tag{2}$$

where α is the path loss exponent and Δ is a constant related to the height of the transmit antenna and the frequency. Because the channel variation is periodical from one cell to another and that is symmetrical in one cell, we only consider half of a period of the channel variation to design power allocation policy.

We denote $\mathscr{P}$ as a power allocation policy, and the corresponding channel capacity can be calculated as

$$C(t) = \log\left(1 + h(t)\,\mathscr{P}(h(t))\right). \tag{3}$$

Energy consumption plays an important role in greenhouse gas emissions, so a high EE performance is a main concern point for the design of future wireless communication. In addition, delay is a key metric to measure QoE. In this paper, our aim is to maximize the EE performance from RRUs to the passengers with the constraints of radio transmission rate, transmission delay requirement, and average power. Inequation (1) is transmission delay requirement. From (1), we can get $c \ge \lambda + 1/\tau + 1/(\tau^2(\mu - \lambda) - \tau)$.

Problem 1 (EE maximization problem). Considering the transmission delay constraint and average power constraint, the EE maximization problem in the downlink C-RAN of HSR is formulated as

$$(\mathbf{P1}) \max_{\mathscr{P}} \quad \frac{\int_0^{T/2} C(t)\,dt}{\int_0^{T/2} \mathscr{P}(h(t))\,dt + P_c} \tag{4a}$$

$$\text{subject to} \quad \frac{2}{T}\int_0^{T/2} C(t)\,dt \ge c, \tag{4b}$$

$$c \ge \lambda + \frac{1}{\tau} + \frac{1}{\tau^2(\mu - \lambda) - \tau}, \tag{4c}$$

$$\frac{2}{T}\int_0^{T/2} \mathscr{P}(h(t))\,dt \ge P_{\text{ave}}, \tag{4d}$$

where P_c is the circuit power consumption and P_{ave} is average transmission power of RRU. Constraint (4b) means that the average channel capacity should be large or equal to the radio transmission rate in half of a period. Constraint (4c) corresponds to the transmission delay requirement. Constraint (4d) means the average power constraint of RRU in half of a period.

Constraint (4b) is a concave function; (4c) and (4d) are linear functions. Therefore, (4b), (4c), and (4d) are all convex constraints. But the objective function is a nonlinear fractional programming, so Problem 1 is a nonconvex optimization problem [30].

3.2. EE Maximization Problem Reformulation. Problem 1 is a nonconvex optimization problem, so it cannot be solved by classical convex optimization methods. In this section, we reformulate the objective function of Problem 1.

We define a nonnegative variable γ^* as the energy-efficient optimal value $\gamma^* = \int_0^{T/2} C^*(t)dt / \int_0^{T/2} \mathscr{P}^*(h(t))dt + P_c$, where $\mathscr{P}^*(h(t))$ is the optimal power allocation policy and $C^*(t)$ is the corresponding channel capacity with the optimal power allocation policy [31].

Lemma 2. *γ^* can be achieved if and only if*

$$\begin{aligned} \max_{\mathscr{P}} \quad & \int_0^{T/2} C(t)\,dt - \gamma^*\left(\int_0^{T/2} \mathscr{P}(h(t))\,dt + P_c\right) \\ &= \int_0^{T/2} C^*(t)\,dt \\ &\quad - \gamma^*\left(\int_0^{T/2} \mathscr{P}^*(h(t))\,dt + P_c\right) = 0. \end{aligned} \tag{5}$$

Proof. The proof of Lemma 2 is in the Appendix. □

From Lemma 2, if we can find the energy-efficient optimal value γ^*, Problem 1 can be solved. However, γ^* cannot be calculated directly; we propose an iterative algorithm (Algorithm 1) to update γ while ensuring the corresponding solution $\mathscr{P}(h(t))$ remains feasible in each iteration. Then, optimal resource allocation policy to solve Problem 3 can be derived.

Problem 3 (reformulated EE maximization problem).

$$\begin{aligned} (\mathbf{P3}) \max_{\mathscr{P}} \quad & \int_0^{T/2} C(t)\,dt \\ & - \gamma^*\left(\int_0^{T/2} \mathscr{P}(h(t))\,dt + P_c\right) \\ \text{subject to} \quad & \frac{2}{T}\int_0^{T/2} C(t)\,dt \ge \lambda + \frac{1}{\tau} + \frac{1}{\tau^2(\mu - \lambda) - \tau}, \\ & \frac{2}{T}\int_0^{T/2} \mathscr{P}(h(t))\,dt \le P_{\text{ave}}, \end{aligned} \tag{6}$$

where γ^* is the energy-efficient optimal value.

3.3. Proposed Iterative Algorithm. For Problem 3, we propose an iterative algorithm to update γ as Algorithm 1. The outer loop can update $\gamma^{(i+1)}$ through $C^{(i)}(t)$ and $\mathscr{P}^{(i)}(h(t))$ of each iteration. The inner loop can calculate the power allocation policy of $\mathscr{P}^{(i)}(h(t))$ and $C^{(i)}(t)$ by using the Lagrangian dual method. The outer loop can calculate $\gamma^{(i)}$.

(1) Initialize $\gamma^{(1)} = 0$, set threshold value ε, maximum number of iteration $I_{\max}$;
(2) i is the number of iteration, set $i = 1$ and begin iteration (Outer Loop);
(3) **for** $1 \le i \le I_{\max}$ **do**
(4) Solve the power allocation with $\gamma^* = \int_0^{T/2} C^*(t)dt / \int_0^{T/2} \mathscr{P}^*(h(t))dt + P_c$; (Inner Loop)
(5) Use the Lagrangian dual method to obtain $C^{(i)}(t)$ and $\mathscr{P}^{(i)}(h(t))$;
(6) **if** $\int_0^{T/2} C^{(i)}(t)dt - \gamma^{(i)}(\int_0^{T/2} \mathscr{P}^{(i)}(h(t))dt + P_c) < \varepsilon$ **then**
(7) Set $\gamma^* = \gamma^{(i)}$;
(8) **break**;
(9) **else**
(10) set $\gamma^{(i+1)} = \int_0^{T/2} C(t)dt / \int_0^{T/2} \mathscr{P}(h(t))dt + P_c$;
(11) **end if**
(12) **end for**

Algorithm 1: Energy-efficient power allocation.

Problem 4 (optimal resource allocation in the inner loop).

$$\begin{aligned}(\mathbf{P4}) \max_{\mathscr{P}} \quad & \int_0^{T/2} C^{(i)}(t)\,dt \\ & - \gamma^{(i)}\left(\int_0^{T/2} \mathscr{P}^{(i)}(h(t))\,dt + P_c\right) \\ \text{subject to} \quad & \frac{2}{T}\int_0^{T/2} C^{(i)}(t)\,dt \\ & \ge \lambda + \frac{1}{\tau} + \frac{1}{\tau^2(\mu-\lambda)-\tau}, \\ & \frac{2}{T}\int_0^{T/2} \mathscr{P}^{(i)}(h(t))\,dt \le P_{\text{ave}},\end{aligned} \tag{7}$$

where $\gamma^{(i)}$ is the i-th iteration energy-efficient value.

3.4. Lagrangian Dual Method. Based on the proposed iterative algorithm, we have obtained the i-th iteration energy-efficient value $\gamma^{(i)}$; then the Problem 1 becomes a convex problem; we can use Lagrangian dual method to solve the optimal power allocation results [30]. The Lagrangian function of objective function can be written as

$$\begin{aligned}L(\mathscr{P}(h(t)), \upsilon, \beta) = & \int_0^{T/2} C(t)\,dt \\ & - \gamma^*\left(\int_0^{T/2} \mathscr{P}(h(t))\,dt + P_c\right) \\ & - \upsilon\int_0^{T/2} (C(t)-c)\,dt \\ & + \beta\int_0^{T/2} \left(\mathscr{P}(h(t)) - P_{\text{ave}}\right)dt,\end{aligned} \tag{8}$$

where υ and β are the Lagrange multipliers.

To maximize the Lagrangian function, we only need to maximize the power allocation pointwise.

$$\begin{aligned}L(\mathscr{P}(h(t)), \upsilon, \beta) = & C(t) - \gamma^*(\mathscr{P}(h(t)) + P_c) \\ & - \upsilon(C(t) - c) \\ & + \beta(\mathscr{P}(h(t)) - P_{\text{ave}}) \\ = & \log(1 + h(t)\mathscr{P}(h(t))) \\ & - \gamma^*(\mathscr{P}(h(t)) + P_c) \\ & - \upsilon(\log(1 + h(t)\mathscr{P}(h(t))) - c) \\ & + \beta(\mathscr{P}(h(t)) - P_{\text{ave}}).\end{aligned} \tag{9}$$

And then, using the Lagrangian dual method differentiating $L(\mathscr{P}(h(t)), \upsilon, \beta)$ with respect to $\mathscr{P}(h(t))$ and setting the result to zero, we can calculate $\mathscr{P}(h(t))$.

$$\begin{aligned}\frac{\partial L(\mathscr{P}(h(t)), \upsilon, \beta)}{\partial \mathscr{P}(h(t))} = & \frac{h(t)}{\ln 2\,(1 + h(t)\mathscr{P}(h(t)))} - \gamma^* \\ & - \upsilon\frac{h(t)}{\ln 2\,(1 + h(t)\mathscr{P}(h(t)))} \\ & + \beta = 0.\end{aligned} \tag{10}$$

$$\mathscr{P}(h(t)) = \frac{1-\upsilon}{\ln 2\,(\gamma^* - \beta)} - \frac{1}{h(t)}. \tag{11}$$

(11) is the water filling form power allocation scheme. The power allocation requires accurate CSI. In HSR scenario, considering the moving track of the train is certain, we utilize the predicted path loss information instead of accurate CSI. Therefore, we call (11) as *predictable path loss based time domain power allocation method* (PPTPA).

Furthermore, since $C(t) \ge c$, we obtain

$$\mathscr{P}(h(t)) = \max\left\{\frac{1-\upsilon}{\ln 2\,(\gamma^* - \beta)} - \frac{1}{h(t)}, \frac{2^c - 1}{h(t)}\right\}. \tag{12}$$

We can calculate the value of υ and β by using the bisection method. (12) is a kind of hybrid power allocation scheme

based on PPTPA. We call it as QoE Constrained EE Power Allocation (QCEPA).

4. Capacity Maximization with EE Constraint Problem

EE performance is a main concern point for the design of wireless communication. However, if we only consider the EE maximization, the system capacity may degrade seriously [20, 32]. In Section 3, we only considered the PISs. In practice, the communication system of HSR may bear not only the PISs but also the TCSs. In this section, we consider the communication system of HSR bears both of them and then design a hybrid power allocation scheme based on PPTPA to support more services subject to EE performance and transmission delay constraints.

4.1. Problem Formulation. From Section 3, we get the optimal EE performance subject to transmission delay and average power constraints. To analyze the EE performance and capacity tradeoff, we formulate the problem to maximize capacity subject to constraints on delay, average power, and EE requirement in this section.

Problem 5 (capacity maximization problem). Considering the transmission delay constraint, average power, and EE requirement, the capacity maximization problem in the downlink C-RAN of HSR can be formulated as

$$(\mathbf{P5}) \max_{\mathscr{P}} \quad \int_0^{T/2} C(t)\,dt \tag{13a}$$

$$\text{subject to} \quad \frac{2}{T}\int_0^{T/2} C(t)\,dt \geq c, \tag{13b}$$

$$c \geq \lambda + \frac{1}{\tau} + \frac{1}{\tau^2(\mu-\lambda)-\tau}, \tag{13c}$$

$$\frac{2}{T}\int_0^{T/2} \mathscr{P}(h(t))\,dt \geq P_{\text{ave}}, \tag{13d}$$

$$\frac{\int_0^{T/2} C(t)\,dt}{\int_0^{T/2} \mathscr{P}(h(t))\,dt + P_c} \geq \eta\gamma^*, \tag{13e}$$

where η is chosen as a weight of the optimal EE. γ^* is the optimal EE value of Problem 1, which is calculated in Section 3. Constraints (13b)–(13d) have the same meaning with constraints (4b)–(4d) in Problem 1. Constraint (13e) means that the EE performance is large or equal to the EE requirement.

Constraint (13e) can be written as

$$\int_0^{T/2} C(t)\,dt - \eta\gamma^*\left(\int_0^{T/2} \mathscr{P}(h(t))\,dt + P_c\right) \geq 0. \tag{14}$$

It becomes a convex constraint. Intuitively, objective function is a convex function, constraint (13b) and constraint (13c) are all convex constraints. Therefore, Problem 5 is a convex optimization problem.

4.2. Lagrangian Dual Method. Problem 5 is a convex optimization problem, so it can be solved by classical convex optimization methods. In this section, we use Lagrangian dual method to solve Problem 5. The Lagrangian function of objective function can be written as

$$\begin{aligned} L(\mathscr{P}(h(t)), \upsilon, \beta, \mu) &= \int_0^{T/2} C(t)\,dt \\ &\quad - \nu\int_0^{T/2} (C(t)-c)\,dt \\ &\quad + \beta\int_0^{T/2} (\mathscr{P}(h(t)) - P_{\text{ave}})\,dt - \mu\left(\int_0^{T/2} C(t)\,dt\right. \\ &\quad \left. - \eta\gamma^*\left(\int_0^{T/2} \mathscr{P}(h(t))\,dt + P_c\right)\right), \end{aligned} \tag{15}$$

where υ, β, and μ are the Lagrange multipliers.

As in Problem 3, our aim is to maximize the Lagrangian function, and we only need to maximize the power allocation pointwise.

$$\begin{aligned} &L(\mathscr{P}(h(t)), \upsilon, \beta, \mu) \\ &\quad = C(t) - \upsilon(C(t)-c) + \beta(\mathscr{P}(h(t)) - P_{\text{ave}}) \\ &\qquad - \mu(C(t) - \eta\gamma^*(\mathscr{P}(h(t)) + P_c)) \\ &\quad = C(t)(1-\upsilon-\mu) + (\mu\eta\gamma^* + \beta)\mathscr{P}(h(t)) + \upsilon c \\ &\qquad - \beta P_{\text{ave}} + \mu\eta\gamma^* P_c. \end{aligned} \tag{16}$$

And then, using the Lagrangian dual method differentiating $L(\mathscr{P}(h(t)), \upsilon, \beta, \mu)$ with respect to $\mathscr{P}(h(t))$ and setting the result to zero, we can calculate $\mathscr{P}(h(t))$.

$$\begin{aligned} &\frac{\partial L(\mathscr{P}(h(t)), \upsilon, \beta, \mu)}{\partial \mathscr{P}(h(t))} \\ &\quad = (1-\upsilon-\mu)\frac{h(t)}{\ln 2\,(1 + h(t)\,\mathscr{P}(h(t)))} + \mu\eta\gamma^* + \beta \\ &\quad = 0. \end{aligned} \tag{17}$$

$$\mathscr{P}(h(t)) = \frac{\upsilon + \mu - 1}{\ln 2\,(\mu\eta\gamma^* + \beta)} - \frac{1}{h(t)}.$$

Furthermore, since $C(t) \geq c$, we get

$$\mathscr{P}(h(t)) = \max\left\{\frac{\upsilon + \mu - 1}{\ln 2\,(\mu\eta\gamma^* + \beta)} - \frac{1}{h(t)}, \frac{2^c - 1}{h(t)}\right\}. \tag{18}$$

(18) is also a kind of hybrid power allocation scheme based on PPTPA. We call it as QoE Constrained Capacity Maximization Power Allocation (QCCMPA).

5. Power Allocation Considering Small-Scale Fading

In the previous sections, we only considered the influence of path loss on the channel. However, the fast time-varying

small-scale fading is the characteristic of wireless channel of HSR; we need to consider the effect of small-scale fading on system performance. As we know, small-scale fading cannot be predicted accurately in HSR scenario. But the statistics information of small-scale fading can be known in the transmitter and it remains unchanged for a long time. Because the small-scale fading can reduce the system capacity, the EE performance can be reduced when we consider the small-scale fading. In this section, we evaluate the effect of small-scale fading on the energy-efficient power allocation policy.

The Nakagami-m model is used to describe the small-scale fading of HSR [29, 33]. m is the fading factor, which increases from 1/2 to ∞. The fading becomes Rayleigh when $m = 1$.

We consider flat fading, and for narrowband signals, the received signal at time t is

$$y(t) = \sqrt{h_s(t)\, h(t)}\, x(t) + z(t), \quad 0 \le t \le \frac{2}{T}, \tag{19}$$

where $\sqrt{h_s(t)}$ is small-scale fading with Nakagami-m distribution. $z(t)$ is complex Gaussian noise with variance σ^2. The instantaneous channel capacity is

$$C(t) = \log\left(1 + h_s(t)\, h(t)\, \mathscr{P}(h(t))\right), \quad 0 \le t \le \frac{2}{T}. \tag{20}$$

Therefore, the channel capacity in a half period is

$$\int_0^{T/2} C(t)\, dt = \int_0^{T/2} \log\left(1 + h_s(t)\, h(t)\, \mathscr{P}(h(t))\right) dt. \tag{21}$$

In HSR scenario, we are interested in small delay constraint and large data arrival rate. The vertical distance d_0 and the coverage distance d_s are small. In this case, the train can receive high SNR even when the train is on the edge of the cell [34]. Therefore, the channel capacity is

$$\begin{aligned}\int_0^{T/2} C(t)\, dt &\approx \int_0^{T/2} \log\left(h_s(t)\, h(t)\, \mathscr{P}(h(t))\right) dt \\ &= \int_0^{T/2} \log\left(h(t)\, \mathscr{P}(h(t))\right) dt \\ &\quad + \int_0^{T/2} \log\left(h_s(t)\right) dt.\end{aligned} \tag{22}$$

From (22), we can see that the channel capacity would be reduced because of the factor $\int_0^{T/2} \log(h_s(t))dt$. Because the period that the train passes through one cell is much larger than the channel coherence time, the small-scale fading is ergodic. Therefore, we get

$$\int_0^{T/2} \log\left(h_s(t)\right) dt = E\left(\log\left(h_s(t)\right)\right) \frac{T}{2}. \tag{23}$$

$\sqrt{h_s(t)}$ is Nakagami-m distribution with parameters Ω and m, where Ω is the average SNR and m is the fading factor. Thus, we obtain

$$E\left(\log\left(h_s(t)\right)\right) = \left[\psi(m) + \ln\frac{\Omega}{m}\right] \log e, \tag{24}$$

where ψ is Digamma function.

Table 1: Parameters in simulation.

Parameter	Description	Value
d_0	Vertical distance	50 m
d_s	Coverage distance	1500 m
v	Train speed	350 km/h
α	Path loss exponent	3
μ	Service processing capacity	6 bits/s
λ	Data arrival rate	5 bits/s/Hz
τ	Delay constraint	400 ms
Δ	Frequency compensation	1 dB
P_{ave}	Average power threshold	40 w

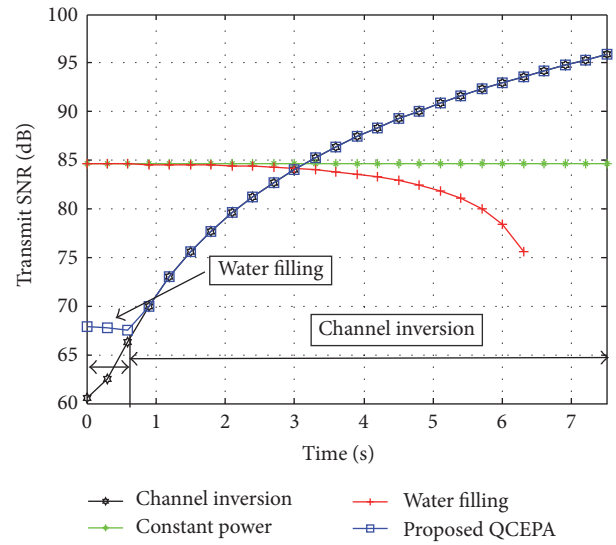


Figure 2: Transmit SNR of proposed QCEPA algorithm compared with other three different power allocation policies.

6. Results and Discussions

In this section, we provide emulation and extensive simulation results to validate previous theoretical analysis results. The simulation parameters are shown in Table 1.

6.1. EE Optimization Power Allocation Policy. We firstly evaluate the EE performance of the proposed QCEPA power allocation solution, and then we analyze the impact of the service processing capacity of VMs μ, channel fading factor m, data arrival rate λ, and different train speeds v on the EE performance. Finally, the effect of small-scale fading on the proposed QCEPA power allocation policy is evaluated.

In Figure 2, the proposed QCEPA algorithm is compared with other three different power allocation policies (water filling, channel inversion, and constant power allocation).

(i) Water filling power allocation: in this scheme, we compare the traditional water filling power allocation scheme. The objective function is to maximize the capacity with average power constraint. Compared to the proposed QCEPA, the traditional water filling

power allocation does not consider the delay constraint, and the objective function is also different.

(ii) Channel inversion power allocation: in this scheme, the power allocation scheme is $\mathcal{P}(h(t)) = (2^c - 1)/h(t)$.

(iii) Constant power allocation: in this scheme, the power is constant in each time slot.

From Figure 2, we can see that the process of QCEPA is divided into two phases; from 0 s to 0.79 s is the first phase, the channel condition is good, and water filling power allocation policy is used; from 0.79 s to 7.5 s is the second phase, the channel condition becomes bad as the train approaches the cell edge, and RRU adopts channel inversion power allocation policy. It also can be seen that the water filling power allocation scheme is no power allocation when the time is 6.5 s. The reason is that the channel condition is bad when the train is at the edge of the cell; the RRU does not allocate power to the passengers. From this figure, it also can be seen that the proposed QCEPA is different from the traditional water filling power allocation. The reason can be explained as follows. Because the channel condition is good from 0 s to 0.79 s, QCEPA adopts the form of water filling power allocation to meet the delay constraint and achieve the maximum EE performance. From 0.79 s to 7.5 s, because the channel condition is bad, QCEPA adopts channel inversion power allocation to meet the delay constraint and achieve the maximum EE performance. However, for the traditional water filling power allocation, because it does not consider the delay constraint and its objective function is to maximize capacity, the allocation results of QCEPA and the traditional water filling power allocation are different. Although these two algorithms are all the form of water filling, the values are different. For example, the solution of QCEPA is $\mathcal{P}(h(t)) = \max\{(1 - \upsilon)/\ln 2(\gamma^* - \beta) - 1/h(t), (2^c - 1)/h(t)\}$, but the solution of traditional water filling is $\mathcal{P}(h(t)) = \max\{1/ - \lambda \ln 2 - 1/h(t), 0\}$.

Figure 3 illustrates the EE performance versus different train speeds under different power allocation policies. From this figure we can see that when the train speed is faster, the EE performance is worse. This is because that the average channel quality of the high-speed train is worse than the low speed train in unit time. It also can be seen that the EE performance of proposed algorithm outperforms the other three power allocation schemes. When the train speed is 350 km/h, the EE performance can reach 0.3408 bps/Hz/W. And the EE performance of channel inversion power allocation policy is the worst. This is because the channel inversion power allocation policy tries to keep a constant transmission rate, no matter whether the channel condition is good or bad. When the channel condition is not very good, the RRU has to increase transmit power, but the channel capacity increases inconspicuously. Therefore, the EE performance becomes bad.

Service preprocessing capability of VMs μ can affect the transmission time and further affect the EE performance. Therefore, μ is an indirect factor affecting the EE performance. Figure 4 shows EE performance versus service processing capability of VMs under different m factors. From this figure, as can be seen, the more serious small-scale fading

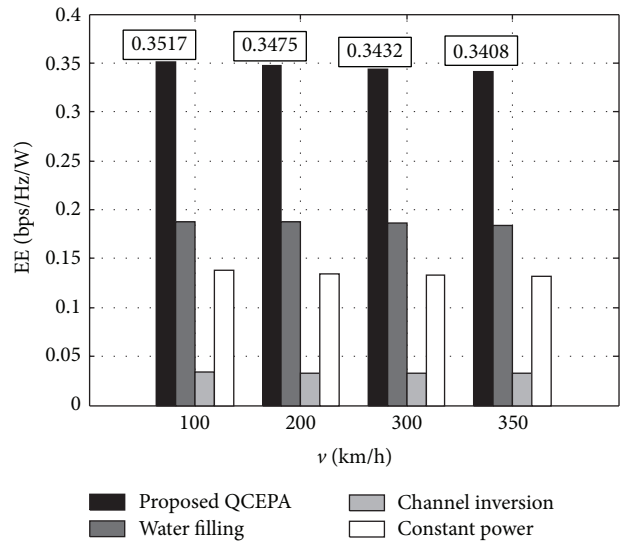


Figure 3: EE performance versus different train speeds under different power allocation policies.

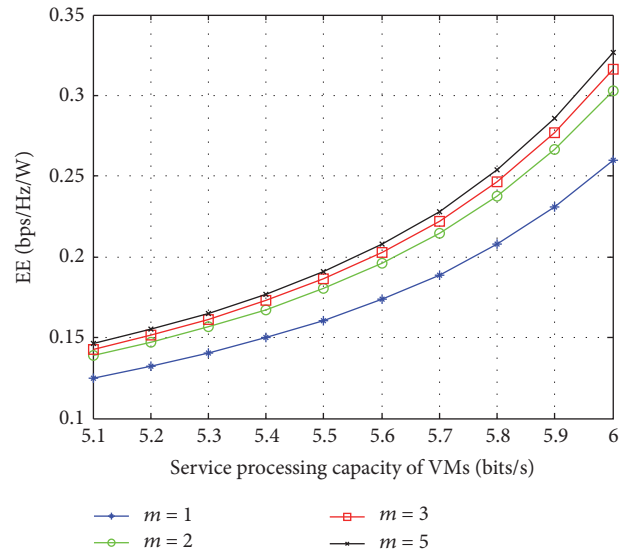


Figure 4: EE performance versus service processing capacity of VMs under different m factors.

leads to worse EE performance. Also we can see that with the increasing service processing capability of VMs, the EE performance also increases. We can explain as follows: when the service processing capability is stronger, the service time is shorter, so the RRU has more time to transmit the data as the channel condition becomes good, which leads to better EE performance.

Another affecting EE performance factor is data arrival rate λ. Figure 5 illustrates EE performance versus data arrival rate with considering small-scale fading or not. In Algorithm 1, we only consider the effect of large-scale fading on power allocation scheme. However, the small-scale fading effect cannot be eliminated. Therefore, we must evaluate the effect of small-scale fading on the power allocation results. In

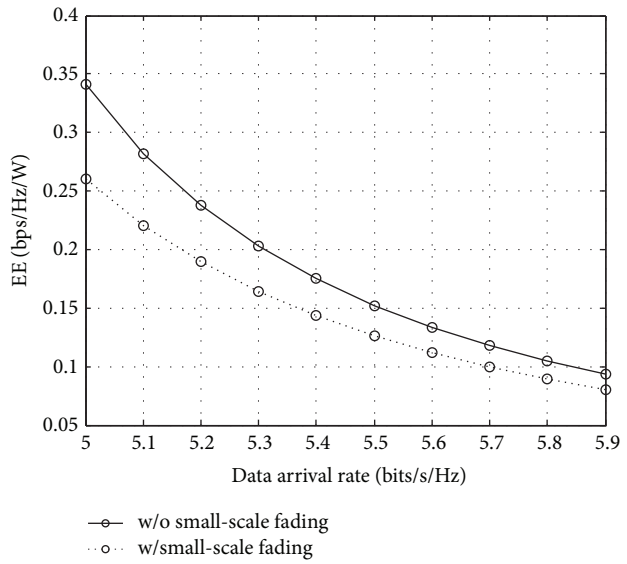


FIGURE 5: EE performance versus data arrival rate with considering small-scale fading when $m = 1$ or not.

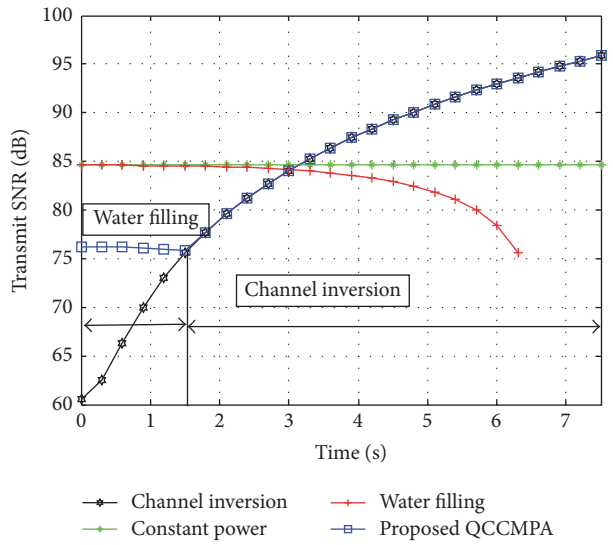


FIGURE 6: Transmit SNR of proposed QCCMPA algorithm compared with other three different power allocation policies when $\eta = 95\%$.

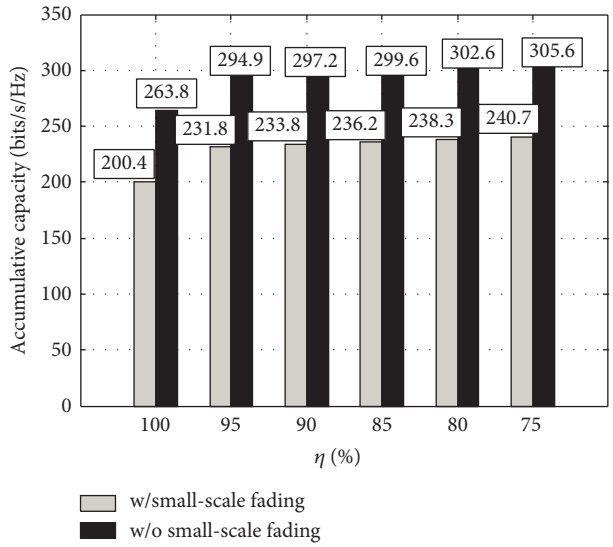


FIGURE 7: Capacity under different EE performance ratio requirements with considering small-scale fading or not.

Section 5, we analyze the effect of small-scale fading on EE performance. For this figure, our goal is to show the effect of small-scale fading on EE performance. It can be seen that with the data arrival rate increases, the EE performance decreases. This is because when the data arrival rate gets higher, the RRU needs more power to increase the transmission rate, so the EE performance becomes bad. Furthermore, the EE performance is obviously decreased when the small-scale fading is considered. It is because that the channel prediction is not accurate, which leads to the unprecise power allocation policy. However, from Figure 5 we can see that the EE performance only decreases by 19.3% when the small-scale fading is considered. Therefore, the power allocation policy is still effective.

6.2. Maximum Capacity Power Allocation Policy. In this section, we evaluate the proposed QCCMPA power allocation scheme and then analyze the impact of small-scale fading on the capacity.

Figure 6 illustrates transmit SNR of proposed QCCMPA algorithm compared with other three different power allocation policies when $\eta = 95\%$. The process of QCCMPA is also divided into two phases; water filling power allocation is used in the first phase, and it takes longer time than that in Figure 2; channel inversion power allocation is used in the second phase. The form of QCCMPA is the same as QCEPA in Figure 2. However, the transmission power and the duration of each phase are different. This is because that the objective function of Problem 5 is to maximize capacity. If we want to get larger capacity, we need take more time to adopt the water filling power allocation scheme. From Figure 6 we can see that the power value in first phase of QCCMPA is larger than that in first phase of QCEPA. This is because the channel condition is good in the first phase, and we increase the transmission power to get larger capacity in this phase. However, as the train moves away from the cell center, the channel condition becomes bad. If we also increase transmission power in the second phase, the channel capacity increases very slightly but the EE performance decreases seriously, which cannot satisfy our requirements. Therefore, to increase system capacity, the best solution is to take more time to adopt the water filling power allocation scheme when the channel condition is good. From this figure, it also can be seen that the proposed QCCMPA is different from the traditional water filling power allocation, which can be explained for the same reason as Figure 2.

Figure 7 illustrates the accumulative capacity under different EE performance ratio requirements with considering

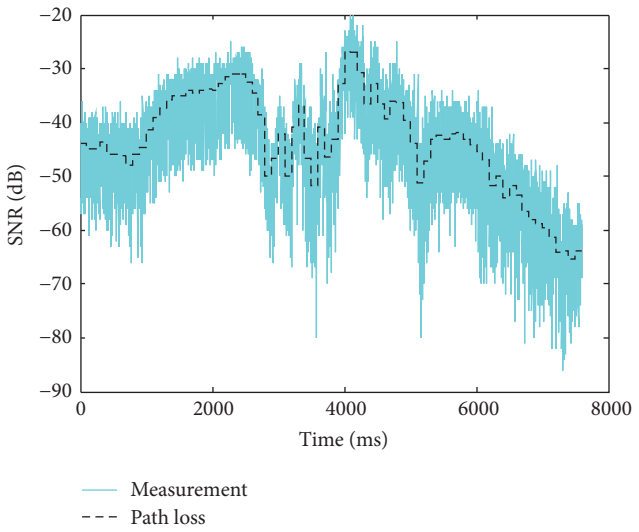


Figure 8: The measured path loss information.

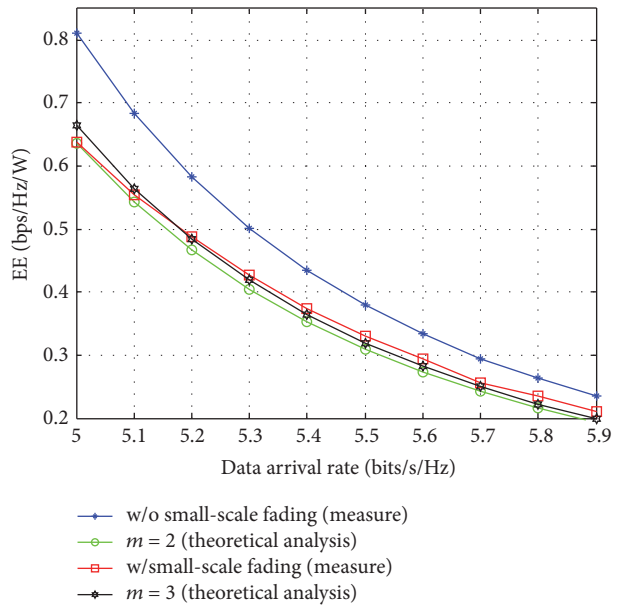


Figure 9: EE performance versus data arrival rate with measured values and theoretical analysis.

small-scale fading or not. As can be seen, when the EE performance reduces from 100% to 95%, the capacity is increased by 11.76% compared to 5% EE loss. However, when the EE performance reduces from 100% to 75%, the capacity only increases by 15.84%. In practical scenarios, we can make a balance between EE and capacity according to our requirements. In Figure 7, we also consider small-scale fading with $m = 1$, where the EE performance reduces from 100% to 95%, and the capacity increases by about 15.66%. When EE performance ratio reduces to 75% of the optimal value, the capacity increases by about 20.12%. Therefore, the power allocation scheme is also effective when considering small-scale fading.

6.3. Performance Emulation. In order to evaluate the effectiveness of the proposed PPTPA in the real HSR scenario, we utilize the channel measurement of Zhengzhou-Xian HSR line [33] to perform it. The channel measurement and path loss information are shown in Figure 8.

We utilize the measured path loss and small-scale fading values to emulate the EE performance versus data arrival rate. And we compare the theoretical analysis with the measured data. Figure 9 has the same trend with Figure 5 and it can be explained for the same reasons. In Figure 5, the EE performance of theoretical analysis decreases by 19.3% when the small-scale fading is considered. However, in Figure 9 the EE performance only decreases by 15.81% when we utilize the measured values. In addition, the EE performance with measured small-scale fading is between $m = 2$ and $m = 3$. Therefore, the proposed power allocation algorithm is effective.

7. Conclusion

In this paper, we investigated two problems in C-RAN of HSR. Because the train location can be predicted, we utilized the path loss information to simplify the CSI of the train and proposed PPTPA to solve two problems. The first problem is the energy-efficient power allocation problem with delay constraint when the communication system of HSR only bears PISs. The second problem is the capacity optimization with joint EE and services transmission delay constraints when the communication system of HSR bears PISs and TCSs. The effect of small-scale fading on the proposed power allocation schemes was evaluated. Emulation results and extensive simulation results based on Zhengzhou-Xian HSR line channel measurements trace have demonstrated that proposed two power allocation policies can achieve EE maximization and capacity maximization. For the first problem, compared with water filling scheme, the EE performance of proposed algorithm is improved by 85%. For the capacity maximization with EE constraint problem, the capacity in the proposed scheme can be increased by 11.76%, when the EE performance reduces from 100% to 95%.

Appendix

Proof of Lemma 2. We prove Lemma 2 with two steps as follows [25].

We prove the sufficient condition of Lemma 2 firstly. We define γ^* as the optimal energy-efficient value of Problem 1, $\gamma^* = \int_0^{T/2} C^*(t)dt / \int_0^{T/2} \mathcal{P}^*(h(t))dt + P_c$, where $\mathcal{P}^*(h(t))$ is the optimal power allocation policy and $C^*(t)$ is the corresponding channel capacity. Obviously, γ^* should satisfy

$$\begin{aligned} \gamma^* &= \frac{\int_0^{T/2} C^*(t)\,dt}{\int_0^{T/2} \mathcal{P}^*(h(t))\,dt + P_c} \\ &\geq \frac{\int_0^{T/2} C(t)\,dt}{\int_0^{T/2} \mathcal{P}(h(t))\,dt + P_c}. \end{aligned} \tag{A.1}$$

From (A.1), we can get

$$\int_0^{T/2} C(t)\,dt - \gamma^* \left(\int_0^{T/2} \mathscr{P}(h(t))\,dt + P_c \right) \le 0,$$
$$\int_0^{T/2} C^*(t)\,dt - \gamma^* \left(\int_0^{T/2} \mathscr{P}^*(h(t))\,dt + P_c \right) = 0. \tag{A.2}$$

Therefore, we can conclude $\max_{\mathscr{P}} \int_0^{T/2} C(t)dt - \gamma^*(\int_0^{T/2} \mathscr{P}(h(t))dt + P_c) = 0$. The sufficient condition is proved.

Secondly, we prove the necessary condition. Suppose that $\widehat{\mathscr{P}}^*(h(t))$ is the optimal power allocation policy of the reformulated objective function and $\widehat{C}^*(t)$ is the corresponding channel capacity. Therefore, $\int_0^{T/2} \widehat{C}^*(t)dt - \gamma^*(\int_0^{T/2} \widehat{\mathscr{P}}^*(h(t))dt + P_c) = 0$. For a feasible power allocation policy $\mathscr{P}(h(t))$ and corresponding channel capacity $C(t)$, they can be written as

$$\begin{aligned} &\int_0^{T/2} C(t)\,dt - \gamma^* \left(\int_0^{T/2} \mathscr{P}(h(t))\,dt + P_c \right) \\ &\le \int_0^{T/2} \widehat{C}^*(t)\,dt - \gamma^* \left(\int_0^{T/2} \widehat{\mathscr{P}}^*(h(t))\,dt + P_c \right) \\ &= 0. \end{aligned} \tag{A.3}$$

From above inequality, we can get

$$\frac{\int_0^{T/2} C(t)\,dt}{\int_0^{T/2} \mathscr{P}(h(t))\,dt + P_c} \le \gamma^*,$$
$$\frac{\int_0^{T/2} \widehat{C}(t)\,dt}{\int_0^{T/2} \widehat{\mathscr{P}}(h(t))\,dt + P_c} = \gamma^*. \tag{A.4}$$

Therefore, the optimal power allocation policy $\widehat{\mathscr{P}}^*(h(t))$ for the reformulated objective function is also the optimal power allocation policy for the original function. The necessary condition is proved. □

Disclosure

This paper has been presented in part at the IEEE 83rd Vehicular Technology Conference (VTC spring) [35], Nanjing, China, 15–18 May 2016.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was partly supported by the Fundamental Research Funds for the Central Universities (no. 2015RC032), the State Key Laboratory of Rail Traffic Control and Safety (nos. RCS2015K011 and RCS2015ZT001), the Key Laboratory of Wireless Sensor Network and Communication, Chinese Academy of Sciences (no. 2013005), the National Natural Science Foundation of China (nos. 61501023, U1334202, and U1534201), and the Project of China Railway Corporation under Grant no. 2016X003-O.

References

[1] B. Ai, X. Cheng, T. Kurner et al., "Challenges toward wireless communications for high-speed railway," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 2143–2158, 2014.

[2] S. Xu, G. Zhu, C. Shen, Y. Lei, and Z. Zhong, "Analysis and optimization of resource control in high-speed railway wireless networks," *Mathematical Problems in Engineering*, vol. 2014, Article ID 781654, 13 pages, 2014.

[3] J. Wang, H. Zhu, and N. J. Gomes, "Distributed antenna systems for mobile communications in high speed trains," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 4, pp. 675–683, 2012.

[4] *C-RAN: The Road Green RAN*, China Mobile Research Institute, 2011.

[5] M. Behjati, M. H. Alsharif, R. Nordin, and M. Ismail, "Energy efficient and high capacity tradeoff in distributed antenna system for a green cellular network," *Journal of Computer Networks and Communications*, vol. 2015, Article ID 170854, 9 pages, 2015.

[6] I. Chih-Lin, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward green and soft: a 5G perspective," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 66–73, 2014.

[7] D. Feng, C. Jiang, G. Lim, L. J. Cimini Jr., G. Feng, and G. Y. Li, "A survey of energy-efficient wireless communications," *IEEE Communications Surveys and Tutorials*, vol. 15, no. 1, pp. 167–178, 2013.

[8] G. Fettweis and E. Zimmermann, "ICT energy consumption—trends and challenges," in *Proceedings of the 11th International Symposium on Wireless Personal Multimedia Communications (WPMC '08)*, pp. 2006–2009, 2008.

[9] M. De Sanctis, E. Cianca, and V. Joshi, "Energy efficient wireless networks towards green communications," *Wireless Personal Communications*, vol. 59, no. 3, pp. 537–552, 2011.

[10] A. González-Gil, R. Palacin, P. Batty, and J. P. Powell, "A systems approach to reduce urban rail energy consumption," *Energy Conversion and Management*, vol. 80, pp. 509–524, 2014.

[11] A. González-Gil, R. Palacin, and P. Batty, "Optimal energy management of urban rail systems: key performance indicators," *Energy Conversion and Management*, vol. 90, pp. 282–291, 2015.

[12] A. Dahane, A. Loukil, B. Kechar, and N.-E. Berrached, "Energy efficient and safe weighted clustering algorithm for mobile wireless sensor networks," *Mobile Information Systems*, vol. 2015, Article ID 475030, 18 pages, 2015.

[13] L. Kong, X.-Y. Liu, M. Tao et al., "Resource-efficient data gathering in sensor networks for environment reconstruction," *Computer Journal*, vol. 58, no. 6, pp. 1330–1343, 2014.

[14] L. Kong and X. Liu, "MZig: enabling multi-packet reception in ZigBee," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking (MobiCom '15)*, pp. 552–565, Paris, France, September 2015.

[15] G. P. Fettweis, K.-G. Chen, and R. Tafazoli, "Green radio: energy efficiency in wireless networks," *Journal of Communications and Networks*, vol. 12, no. 2, pp. 99–102, 2010.

[16] Y. Zhou and W. Yu, "Optimized backhaul compression for uplink cloud radio access network," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1295–1307, 2014.

[17] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green cloud-RAN," *IEEE Transactions on Wireless Communications*, vol. 13, no. 5, pp. 2809–2823, 2014.

[18] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Robust and efficient distributed compression for cloud radio access networks," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 2, pp. 692–703, 2013.

[19] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 11, pp. 5275–5287, 2015.

[20] L. Musavian and Q. Ni, "Effective capacity maximization with statistical delay and effective energy efficiency requirements," *IEEE Transactions on Wireless Communications*, vol. 14, no. 7, pp. 3824–3835, 2015.

[21] X. Li, X. Ge, X. Wang, J. Cheng, and V. C. M. Leung, "Energy efficiency optimization: joint antenna-subcarrier-power allocation in OFDM-DASs," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7470–7483, 2016.

[22] K. Xie, X. Huang, S. Hao, M. Ma, P. Zhang, and D. Hu, "E^3 MC: improving energy efficiency via elastic multi-controller sdn in data center networks," *IEEE Access*, vol. 4, pp. 6780–6791, 2016.

[23] Q. Wu, W. Chen, D. W. Kwan Ng, J. Li, and R. Schober, "User-centric energy efficiency maximization for wireless powered communications," *IEEE Transactions on Wireless Communications*, vol. 15, no. 10, pp. 6898–6912, 2016.

[24] C. C. Zarakovitis and Q. Ni, "Maximizing energy efficiency in multiuser multicarrier broadband wireless systems: convex relaxation and global optimization techniques," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 7, pp. 5275–5286, 2016.

[25] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 11, no. 9, pp. 3292–3304, 2012.

[26] X. Yang, X. Li, B. Ning, and T. Tang, "A survey on energy-efficient train operation for urban rail transit," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 2–13, 2016.

[27] J. Tang, W. P. Tay, and T. Q. S. Quek, "Cross-layer resource allocation with elastic service scaling in cloud radio access network," *IEEE Transactions on Wireless Communications*, vol. 14, no. 9, pp. 5068–5081, 2015.

[28] P. Demestichas, A. Georgakopoulos, D. Karvounas et al., "5G on the Horizon: key challenges for the radio-access network," *IEEE Vehicular Technology Magazine*, vol. 8, no. 3, pp. 47–53, 2013.

[29] C. Zhang, P. Fan, K. Xiong, and P. Fan, "Optimal power allocation with delay constraint for signal transmission from a moving train to base stations in high-speed railway scenarios," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 12, pp. 5775–5788, 2015.

[30] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[31] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Communications*, vol. 21, no. 6, pp. 126–135, 2014.

[32] L. Musavian and Q. Ni, "Delay-QoS-driven spectrum and energy efficiency tradeoff," in *Proceedings of the 1st IEEE International Conference on Communications (ICC '14)*, pp. 4981–4986, Sydney, Australia, June 2014.

[33] S. Lin, L. Kong, L. He et al., "Finite-state Markov modeling for high-speed railway fading channels," *IEEE Antennas and Wireless Propagation Letters*, vol. 14, pp. 954–957, 2015.

[34] S. Lin, Z. Zhong, L. Cai, and Y. Luo, "Finite state Markov modelling for high speed railway wireless communication channel," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM '12)*, pp. 5421–5426, IEEE, Anaheim, Calif, USA, December 2012.

[35] S. Li, G. Zhu, S. Lin, Q. Gao, S. Xu, and L. Xiong, "Energy-efficient power allocation in cloud radio access network of high-speed railway," in *Proceedings of the 83rd IEEE Vehicular Technology Conference (VTC Spring '16)*, Nanjing, China, May 2016.

Hindawi
Mobile Information Systems
Volume 2017, Article ID 7410546, 13 pages
http://dx.doi.org/10.1155/2017/7410546

Research Article

Macrocell Protection Interference Alignment in Two-Tier Downlink Heterogeneous Networks

Jongpil Seo, Hyeonsu Kim, Jongmin Ahn, and Jaehak Chung

Department of Electronic Engineering, Inha University, Incheon, Republic of Korea

Correspondence should be addressed to Jaehak Chung; jchung@inha.ac.kr

Received 1 August 2016; Revised 4 November 2016; Accepted 1 December 2016; Published 3 January 2017

Academic Editor: Mariusz Głąbowski

Copyright © 2017 Jongpil Seo et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Conventional interference alignment (IA) has been developed to mitigate interference problems for the coexistence of picocells and macrocells. This paper proposes a macrocell protection interference alignment (MCP-IA) in two-tier MIMO downlink heterogeneous networks. The proposed method aligns the interference of the macro user equipment (UE) and mitigates the interference of the pico-UEs with a minimum mean squared error interference rejection combining (MMSE-IRC) receiver. Compared to the conventional IA, the proposed MCP-IA provides an additional array gain obtained by the precoder design of the macro BS and a diversity gain achieved by signal space selections. The degrees of freedom (DoF) of the proposed MCP-IA are equal to or greater than that of the conventional IA and are derived theoretically. Link level simulations show the link capacity and the DoF of the macro UE, and also exhibit the proposed MCP-IA attaining additional array gain and diversity gain. The system level simulation illustrates that the proposed method prevents the interference of the macro UE completely and preserves the throughput of the pico-UE irrespective of the number of picocells. For 4×2 antenna configuration, the system level simulation demonstrates that the proposed MCP-IA throughput of the macro UE is not affected by the number of picocells and that the proposed MCP-IA throughput of the picocells approaches that of single-user MIMO (SU-MIMO) with a 3% loss.

1. Introduction

Heterogeneous networks have been researched for improving the system capacity [1–3]. Pico-, femto-, or relay base stations (BSs) are adopted to cover the shadowing regions and increase the system capacity with a lower power compared to that of the traditional macro BSs. In heterogeneous networks, all the networks utilize the same carrier frequency to increase the spectral efficiency and to avoid bandwidth segmentation. Severe intercell interference at the boundaries of the small cells, however, reduces the system capacity [3].

To overcome intercell interference, enhanced intercell interference coordination (eICIC) techniques were developed [4, 5]. eICIC scheme partitions the coordinated resources between the macro- and the picocells and allocates the interference-free resources of the almost blank subframes (ABSs) to the user equipment devices (UEs). The semistatic coordination technique of eICIC provides an advantage in intercell interference management but such a preserved resource management results in inefficient resource utilizations [6, 7].

For high-efficiency resource utilization, interference alignment (IA) was researched [8]. The concept of IA is to align the interference with the other transmitters into a reduced dimensional subspace at each receiver [9, 10]. Then, IA enables an interference-free communication through the remaining signal subspaces for all the receivers and achieves half the multiplexing gain or the degrees of freedom (DoF).

Centralized IA [8] utilizes symbol extensions in any available domain. If the number of users of IA increases, the dimensions of the signal spaces induced by the symbol extensions increase and cause complex matrix operations. To avoid symbol extensions in the time or frequency domains, decentralized IA such as interference leakage minimization algorithm (ILMA) [11] was developed, enabling the alignment of the interference in a finite spatial domain. The deployment of the decentralized IA is restricted by the number of antennas and requires an additional complexity of at least tens of

iterations for convergence [12]. Conventional IA schemes for heterogeneous networks, however, are restricted by the limited number of users and antenna configurations. For a large two-tier network, the number of participant users also cannot be increased. To overcome this problem, some researches proposed user selection schemes requiring additional system complexity that arranges the adequate number of the BS-UE pairs [13–16].

In practical heterogeneous networks, interference channels have not been clearly developed and some of the weak interference channels can be ignored. For example, picocells are deployed in a macrocell to cover the shadowing region and the interference from a macro BS to pico-UE is weak. However, if the location of the macro UE moves into the pico-BS, the macro UE is exposed to interference from the pico-BSs. In this case, protection against the interference from a pico-BS to the macro UE is more important than that from a macro BS to the pico-UEs. Thus, the requirements of the precoding matrix designs for all the BSs in the conventional IA scheme can be relaxed.

Therefore, we propose a macrocell protection IA (MCP-IA) method that utilizes finite dimensional orthogonal subspaces in the spatial domains without iterative computations. To align the interference of the macro UE from the pico-BSs, signal subspaces of the macro UE with closed-forms are utilized and the interference rejection method is applied to the pico-UEs for reducing the interference from the macro BSs and the other pico-BSs. The proposed MCP-IA works well because the interference of the pico-UEs is small in practice. In addition, the proposed MCP-IA achieves diversity gain by selecting the subspace and the array gain by beamforming for the macro UE because the proposed MCP-IA constructs orthogonal subspaces of the macro UE between the signal and the interference subspaces. These two advantages are not obtained in the conventional IA. The DoF of the proposed MCP-IA is equal to or greater than that of the conventional IA. The proposed scheme is also effective to 5G systems since the carrier frequencies in 5G are considered as millimeter wave (mmWave), its indoor propagation and penetration losses are large, and it does not affect large interference to other indoor picocells [17–19].

To demonstrate the advantages of the proposed method, link level simulations are performed to compare the DoF and the capacity of the proposed MCP-IA with ILMA as the conventional IA, eICIC, and single-user multiple-input multiple-output (SU-MIMO) schemes; system level simulations are executed to provide the throughput of the pico- and the macrocells, for each scheme.

The rest of this paper is organized as follows. In Section 2, the system model is discussed. In Section 3, the design methods for precoding and the receiver matrices of the proposed scheme are described, and the achievable DoF and the optimality of the proposed MCP-IA are analyzed, and the complexity of the proposed MCP-IA is calculated. The numerical results are executed for the link level and the system level simulations in Section 4, and the conclusion follows in Section 5.

The notations used in this paper are defined as follows. Vectors and matrices are written in boldface with the matrices in capitals. All the vectors are column vectors. $\mathbf{A}^H$ denotes the conjugate transpose of $\mathbf{A}$. $\mathscr{C}(\mathbf{A})$ denotes the column space of $\mathbf{A}$. $\mathscr{R}(\mathbf{A})$ denotes the range space of $\mathbf{A}$. $\mathbf{A}^\perp$ and $\|\mathbf{A}\|_F$ denote the orthogonal complement and the Frobenius norm of $\mathbf{A}$, respectively. $\mathbf{I}_n$ denotes an identity matrix of size n. The random vector, $\mathbf{x} \sim \mathscr{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, denotes that $\mathbf{x}$ is drawn from a complex Gaussian distribution with a mean vector, $\boldsymbol{\mu}$, and a covariance matrix, $\boldsymbol{\Sigma}$. $v^d_{\min}(\mathbf{A})$ denotes a matrix whose columns are the eigenvectors of $\mathbf{A}$ corresponding to the d smallest eigenvalues of $\mathbf{A}$. $v^d_{\max}(\mathbf{A})$ denotes a matrix whose columns are the eigenvectors of $\mathbf{A}$ corresponding to the d largest eigenvalues of $\mathbf{A}$. $\mathbb{E}[\cdot]$ denotes the expected value of a random variable.

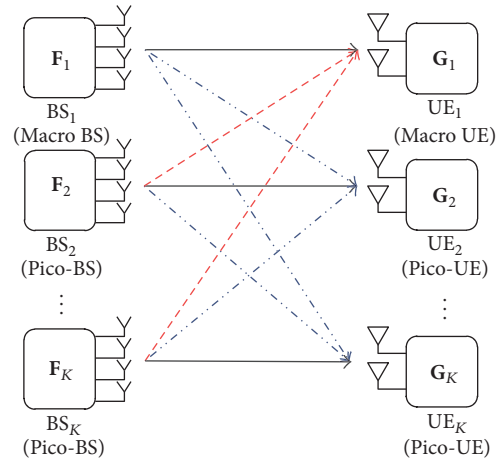


Figure 1: System model.

2. System Model

For the heterogeneous networks, a single macrocell and $K-1$ picocells are considered, as in Figure 1. Each BS serves one UE per cell and the BSs are connected by backhaul links for exchanging a small amount of channel information. Let BS_1 be a macrocell BS and BS_k, where $k \neq 1$, be the picocell BSs. Each BS and UE are equipped with M transmit antennas and N receive antennas, respectively. A BS can transmit d spatial streams that are bounded by $d \leq \min(M, N)$. Assume that all the MIMO channels experience slow fading and that precoding and data transmission are executed within coherence time. Each subcarrier of OFDM system has narrow band channel. The precoding matrix at each BS needs to be calculated for every channel realization.

The kth BS, BS_k, transmits d symbols of the spatial symbol vector, $\mathbf{x}_k \in \mathbb{C}^d$, through M transmit antennas with a linear precoder, $\mathbf{F}_k \in \mathbb{C}^{M \times d}$, where $\mathbf{x}_k$ satisfies the power constraint of $\mathbb{E}[\mathrm{tr}(\mathbf{x}_k \mathbf{x}_k^H)] = 1$ and $\mathbf{F}_k$satisfies $\mathrm{tr}(\mathbf{F}_k \mathbf{F}_k^H) = d$. Then, the received signal, $\mathbf{y}_k$, of the kth user, UE_k, is given by

$$\mathbf{y}_k = \sqrt{P_{kk}}\mathbf{H}_{kk}\mathbf{F}_k\mathbf{x}_k + \sum_{j \neq k} \sqrt{P_{kj}}\mathbf{H}_{kj}\mathbf{F}_j\mathbf{x}_j + \mathbf{n}_k, \quad (1)$$

where $\mathbf{x}_k$ and $\mathbf{x}_j$ in (1) denote the desired signal and interference, respectively. P_{kk} denotes the desired received signal

power of UE_k. $\mathbf{H}_{kj}$ denotes a channel matrix from BS_j to UE_k with $\mathbf{H}_{kj} \in \mathbb{C}^{N\times M}$. P_{kj} denotes the received interference power from BS_j to UE_k, including the path loss and the shadowing. $\mathbf{n}_k$ denotes the additive complex white Gaussian noise at UE_k with a zero mean and a σ_n^2 variance, $\mathcal{CN}(0, \sigma_n^2\mathbf{I}_N)$.

3. Macrocell Protection Interference Alignment

In this section, we propose a macrocell protection interference alignment in downlink two-tier heterogeneous networks. In two-tier heterogeneous networks, picocells are utilized to provide coverage extensions, shadow regions coverage, and offloading in a cell. If the picocells are randomly deployed, interference channel models for the heterogeneous networks may not be obtained and the conventional IA cannot be directly utilized for practical systems. If macro UEs are located near the picocells, the macro UEs are interfered by the picocell BSs, but pico-UEs are less interfered from the macro BSs because of their deployed locations. Thus, macro UEs need to be protected from the interference of the picocell BSs, while the pico-UEs may not require protection. The proposed MCP-IA considers practical cell environments and adopts IA concepts for the scenarios. The proposed algorithm is described in the following subsection.

3.1. Proposed MCP-IA. In order to suppress interference in practical heterogeneous networks, since the interference of the pico-UEs are small, full channel knowledge may not be required to align all the interference and IA can be developed for specific users who suffer from interference. In this paper, we develop the macrocell protection interference alignment wherein the cross-tier interference of the macro UE is suppressed by IA and the cross-tier interference of the pico-UE is suppressed by receiver processing, based on the result of the IA. To mitigate the interference from the pico-BSs to the macro UE, the interference at the macro UE should be aligned. In addition, to achieve the DoF, all the interference to the macro UE should fall in the interference spaces of the macro UE. Then, the signal and interference spaces of the macro UE should be preserved and nulled, respectively.

For the first interference mitigation requirement, a perfect alignment of the cross-tier interference at the macro UE should be satisfied and is given by

$$\mathscr{C}\left(\mathbf{H}_{12}\mathbf{F}_2\right) = \mathscr{C}\left(\mathbf{H}_{13}\mathbf{F}_3\right) = \cdots = \mathscr{C}\left(\mathbf{H}_{1K}\mathbf{F}_K\right). \quad (2)$$

Equation (2) forms subspace equivalence based on the orthogonal complement and can be rewritten as the orthogonality of the desired signal and the interference subspaces of the macro UE, given by

$$\mathbf{G}_1^H\mathbf{H}_{1k}\mathbf{F}_k = \mathbf{0}, \quad \forall k \neq 1, \quad (3)$$

$$\operatorname{rank}\left(\mathbf{G}_1^H\mathbf{H}_{11}\mathbf{F}_1\right) = d, \quad (4)$$

where $\mathbf{G}_1$ denotes the interference suppression matrix of the macro UE. Since (3) is a system of bilinear equations, the system may have infinite solutions by initial values. If $\mathbf{G}_1$ is chosen as an appropriate value, however, the system can be easily solved without an iterative approach.

Let $\mathbf{u}_i$ be an arbitrary orthonormal basis to construct the orthogonal subspaces as follows:

$$\begin{aligned}\mathcal{S}_1 &= \mathscr{C}\left(\left[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d\right]\right),\\ \mathcal{I}_1 &= \mathscr{C}\left(\left[\mathbf{u}_{d+1}, \mathbf{u}_{d+2}, \ldots, \mathbf{u}_N\right]\right),\end{aligned} \quad (5)$$

where $\mathcal{S}_1$ and $\mathcal{I}_1$ denote the desired signal and the interference subspaces of the macro UE, respectively, and $\mathcal{S}_1, \mathcal{I}_1 \neq \emptyset$. The orthonormal basis, $\mathbf{u}_i$, can be obtained by several well-known orthogonalization techniques such as Gram-Schmidt process, QR decomposition (QRD), or singular value decomposition (SVD).

3.2. Feasibility of the Proposed MCP-IA. The proposed MCP-IA constructs the signal and interference subspaces arbitrarily, different from the conventional IA. Therefore, a feasibility condition for the proposed MCP-IA should be newly determined. The number of available spatial symbols between the macro BS and the pico-BS without cross-tier interference at the macro UE is investigated, and the following lemma is derived.

Lemma 1. *For given $\mathbf{G}_1$, the number of spatial symbols, d, is bounded by*

$$d \leq \min\left(\frac{M}{2}, N\right). \quad (6)$$

Proof. Let d_m and d_p be the number of spatial streams transmitted from a macro BS and a pico-BS, respectively. For given $\mathbf{G}_1$, the number of equations in (3) is $d_m d_p$. Using the uniqueness of the subspace in [9], the number of variables in (3) is given by $d_p(M - d_p)$. Therefore, d_m and d_p must satisfy the following condition:

$$d_p\left(M - d_p\right) \leq d_m d_p. \quad (7)$$

Substituting $d_m = d_p = d$ in (7), d can be obtained by

$$d \leq \frac{M}{2}. \quad (8)$$

As $d \leq \min(M, N)$, by the property of MIMO in Section 2, the number of spatial streams is bounded by

$$d \leq \min\left(\frac{M}{2}, N\right). \quad (9)$$

□

Lemma 1 shows that the feasibility of the proposed MCP-IA is not determined by the number of picocells but by the MIMO antenna configurations. Different from the conventional IA, if the number of transmit antennas is greater than that of the receive antennas, that is, $M \geq 2N$, the proposed MCP-IA can achieve a greater DoF compared to the conventional IA. For example, in 4×2 MIMO systems, the conventional IA achieves DoF = 1 only, even if it is feasible. The proposed IA, however, always attains DoF = 2, indicating that the proposed MCP-IA obtains larger spatial multiplexing gain without cross-tier interference.

3.3. Precoding and Receiver Matrix Design of the Proposed MCP-IA. In this subsection, the precoding matrix of the transmitters and receivers for the proposed MCP-IA is derived based on the requirements of the previous subsection. The interference suppression matrix, $\mathbf{G}_1$, of the macro UE is directly derived when the orthogonal subspaces of the macro UE are chosen. Since $\mathbf{G}_1$ maps the macro BS's symbols into $\mathcal{S}_1$, that is, $\mathcal{S}_1 = \mathcal{C}(\mathbf{G}_1)$, $\mathbf{G}_1$ can be set as

$$\mathbf{G}_1 = \left[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d\right]. \tag{10}$$

For the derivation of the precoding matrix, $\mathbf{F}_1$, $\mathbf{F}_1$ needs to be selected for the signal spaces given by (4). $\mathcal{S}_1$ should be equivalent to a range space of $\mathbf{H}_{11}\mathbf{F}_1$; that is, $\mathcal{S}_1 \subseteq \mathcal{R}(\mathbf{H}_{11}\mathbf{F}_1)$. Then, $\mathbf{F}_1$ is set as

$$\mathbf{F}_1 = \left[\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d\right], \tag{11}$$

where $\mathbf{v}_i$ denotes the right singular vector of $\mathbf{G}_1^H\mathbf{H}_{11}$, corresponding to the ith largest singular value.

If the DoF of the macro UE is greater than two, that is, $d \geq 2$, array processing and the transmit power allocation methods can be applied at the macro BS to improve the capacity performance. In the proposed MCP-IA scheme, the macro UE obtains an array gain and provides power allocations because $\mathbf{v}_i$ in $\mathbf{F}_1$ is calculated from $\mathbf{H}_{11}$ for the desired signal. Thus, the proposed MCP-IA can provide a greater signal-to-interference ratio (SIR) for the macro UE that is located in the edge of the macrocell and suffers from a low received signal power and cross-tier interference. Note that the conventional IA is not designed for maximizing the capacity and the array gain of the desired signal subspace.

The precoding matrix of the picocell BSs, $\mathbf{F}_k$, $\forall k \neq 1$, can be calculated by (3) and is a homogeneous system of linear equations. If $\mathbf{F}_k = \mathbf{0}$, the solution set of the system is trivial and the picocell BSs do not transmit. For a nonnontrivial solution of (3), the nonzero null space of $\mathbf{G}_1^H\mathbf{H}_{1k}$ is the solution. For example, the nonzero null space of $\mathbf{G}_1^H\mathbf{H}_{1k}$ can be calculated by the selection of the right singular vector of the SVD of $\mathbf{G}_1^H\mathbf{H}_{1k}$ with vanishing singular values. As $\mathbf{G}_1^H\mathbf{H}_{1k} \in \mathbb{C}^{d\times M}$ and $d \leq \min(M/2, N)$, $\mathbf{G}_1^H\mathbf{H}_{1k}$ has $M - d$ zero singular values. Then, the precoding matrix, $\mathbf{F}_k$, of picocell BSs, $\forall k \neq 1$, is obtained as

$$\mathbf{F}_k = \left[\overline{\mathbf{v}}_{M-d+1}^k, \overline{\mathbf{v}}_{M-d+2}^k, \ldots, \overline{\mathbf{v}}_M^k\right], \quad \forall k \neq 1, \tag{12}$$

where $\overline{\mathbf{v}}_i^k$ denotes the right singular vector of $\mathbf{G}_1^H\mathbf{H}_{1k}$ corresponding to the ith largest singular value. Thus, $\mathbf{G}_1$ and all $\mathbf{F}_k$ satisfy the requirements of (3) and the proposed MCP-IA provides the DoF to the macro UE and suppresses the cross-tier interference of the macro UE from the other picocell BSs.

Since one of the main goals of the proposed MCP-IA is to protect the macro UE from the cross-tier interference, the achievable DoF of the proposed MCP-IA for the macro UE is important. Thus, the DoF of the proposed MCP-IA is derived by the following lemma.

Lemma 2. *The achievable DoF of the macro UE are given as follows:*

$$\lim_{\rho\to\infty} \frac{C_M(\rho)}{\log_2(\rho)} = d, \tag{13}$$

where $C_M(\rho)$ and ρ denote the capacities of the macro UE and signal-to-noise ratio (SNR), respectively.

Proof. See Appendix A. □

Let us consider the interference suppression matrix of the pico-UE. The interference of the pico-UEs from the macro BSs is not significant and the interference among the picocells is very small. In this scenario, the conventional strict IA need not be applied for designing $\mathbf{G}_k$. In the proposed MCP-IA, the precoding matrix, $\mathbf{F}_k$, $\forall k \neq 1$, of the picocell BS is aligned to the null spaces of the macro UE but not for other pico-UEs. The interference of the pico-UEs exist and are accumulated from the other picocell BSs and the macrocell BSs. Therefore, an interference mitigation method for the pico-UEs needs to be developed. The distribution of the accumulated interference can be modeled as Gaussian by the central limit theorem. In this paper, an MMSE-IRC (minimum mean squared error interference rejection combining) receiver is adopted to mitigate the interference and is given by

$$\mathbf{G}_k = \frac{1}{\beta_k}\left(\mathbf{Q}_k + \sigma_n^2\mathbf{I}_N\right)^{-1}\mathbf{H}_{kk}\mathbf{F}_k, \quad \forall k \neq 1, \tag{14}$$

where β_k denotes the normalization factor satisfying $\|\mathbf{G}_k\|_F^2 = d$. $\mathbf{Q}_k$ denotes the interference covariance matrix of the kth pico-UE and is given by

$$\mathbf{Q}_k = \sum_{j\in\mathcal{B}_k} \frac{P_{kj}}{d}\mathbf{H}_{kj}\mathbf{F}_j\mathbf{F}_j^H\mathbf{H}_{kj}^H, \tag{15}$$

where $\mathcal{B}_k$ denotes the interference of the kth pico-UE from other BSs. $\mathcal{B}_k$ depends on the cross-correlation among the picocells. If the channels between the pico-UEs and the other pico-BSs are correlated or the channel gains are very small, the rank of $\mathbf{Q}_k$ is less than or equal to $(N - d)$ and the kth pico-UE has a d-dimensional interference-free subspace; that is, the pico-UE preserves the DoF without a complex interference alignment. This scenario is common in practical heterogeneous networks [20]. If many picocells are colocated and the interference from the other pico-BSs is strong and uncorrelated, that is, $N - d < \operatorname{rank}(\mathbf{Q}_k) \leq N$, the pico-UEs of the proposed method may not receive any signals from their BSs. In practice, however, this does not happen frequently. In the following Numerical Results section, the system level simulation results show that the proposed MCP-IA demonstrates a greater system throughput compared to the conventional IA.

For the link capacity maximization of the macro UE, the conventional IA schemes do not maximize the capacity since the precoding matrices of them are designed for achieving the DoF. The macro BS of the proposed MCP-IA, however, obtains the maximum link capacity with the optimal precoding matrix, which is important at the low SNR regime

of the cellular environment. The optimal precoder design of pico-BSs that operates in the large SNR regime may not be important because their capacities are close to the optimal capacity. The optimality of the macro BS of the proposed MCP-IA is shown in Lemma 3.

Lemma 3. *The conditions finding the optimal precoding matrix* $\mathbf{F}_1$ *are given as follows:*

(1) *For* $M < 2N$, $\mathbf{F}_1$ *is optimal if the columns of* $\mathbf{G}_1$ *are the left singular vectors of* $\mathbf{H}_{11}$.

(2) *For* $M \geq 2N$, $\mathbf{F}_1$ *is always optimal regardless of* $\mathbf{G}_1$.

Proof. See Appendix B. □

3.4. Complexity of MCP-IA. One of the benefits of the proposed MCP-IA is the low computational complexity compared with that of the conventional IA (ILMA). The comparison is performed by the total number of floating point operations (FLOPs).

Table 1 exhibits the number of FLOPs for the proposed MCP-IA and ILMA. In Table 1, Σ and V and U denote singular values and right and left singular matrices, respectively. $|\mathscr{B}_k|$ denotes the number of interferers of the kth pico-UE from other BSs. $\overrightarrow{\mathbf{Q}_k}$ denotes an interference covariance matrix of reverse link in ILMA. L denotes the number of QR iterations for eigenvalue decomposition. L is set as 5 so that the mean square error of the eigenvector is lower than 10^{-4}. The conventional ILMA requires more than 70 iterations to converge [21]. As seen in Table 1, the significant computational saving of the proposed MCP-IA for 4×2 antenna configuration is obtained compared with the conventional ILMA.

In 5G mmWave small cell environments, interference among picocells is reduced and $|\mathscr{B}_k|$ decreases, too. Thus, the proposed MCP-IA is implemented with lower complexity.

4. Numerical Results

We evaluate the link and the system level performance of the proposed MCP-IA compared to that of the conventional IA, for which interference leakage minimization algorithm (ILMA) is used, and other interference suppression methods. First, in link level simulations, an isolated heterogeneous network with fixed SIRs for the cross-tier and cotier interference is assumed. The capacity of the UEs versus the SNR is evaluated and the achievability of the DoF at the macro UE is verified. The effects of the channel estimation errors are tested and the optimality of the proposed MCP-IA is evaluated. Next, in the system level simulation, the throughputs versus the number of picocells of the proposed MCP-IA under multiple macrocells are executed. The system level simulation parameters are summarized in Table 2.

4.1. Link Level Performance. The link level simulation shows the capacities of the UEs versus the SNR and exhibits the DoF of the proposed MCP-IA, ILMA, and SU-MIMO for heterogeneous networks. The number of picocells varies from one to four for one macrocell. 2×2 antenna configuration and 4×2 antenna configuration are considered. A macrocell and a picocell contain one UE per cell. The received signal power of all the UEs varies and the noise variance, σ_n^2, at the UEs is set to be one. For the interference, the average SIR for the cross-tier is set to be 3 dB and the average SIR for the cotier interference is set to be 20 dB. SU-MIMO is compared to observe interference effects from other BSs. SU-MIMO is not designed to suppress the interference but treats the interference as additive noise.

For link capacity comparison, the capacity equation including the other cell interference, precoding and interference suppression matrices, is defined as follows:

$$C_k = \log_2 \det\left(\mathbf{I}_N + \frac{(P_{kk}/d)\,\mathbf{G}_k^{\mathrm{H}}\mathbf{H}_{kk}\mathbf{F}_k\mathbf{F}_k^{\mathrm{H}}\mathbf{H}_{kk}^{\mathrm{H}}\mathbf{G}_k}{\mathbf{G}_k^{\mathrm{H}}\left(\mathbf{Q}_k + \sigma_n^2\mathbf{I}_N\right)\mathbf{G}_k}\right). \quad (16)$$

In Figures 2, 3, and 4, $C_{\mathrm{MU}} = C_1$ and $C_{\mathrm{PU}} = C_k$, $\forall k \neq 1$, denote the capacity of the macro UE and the pico-UE, respectively.

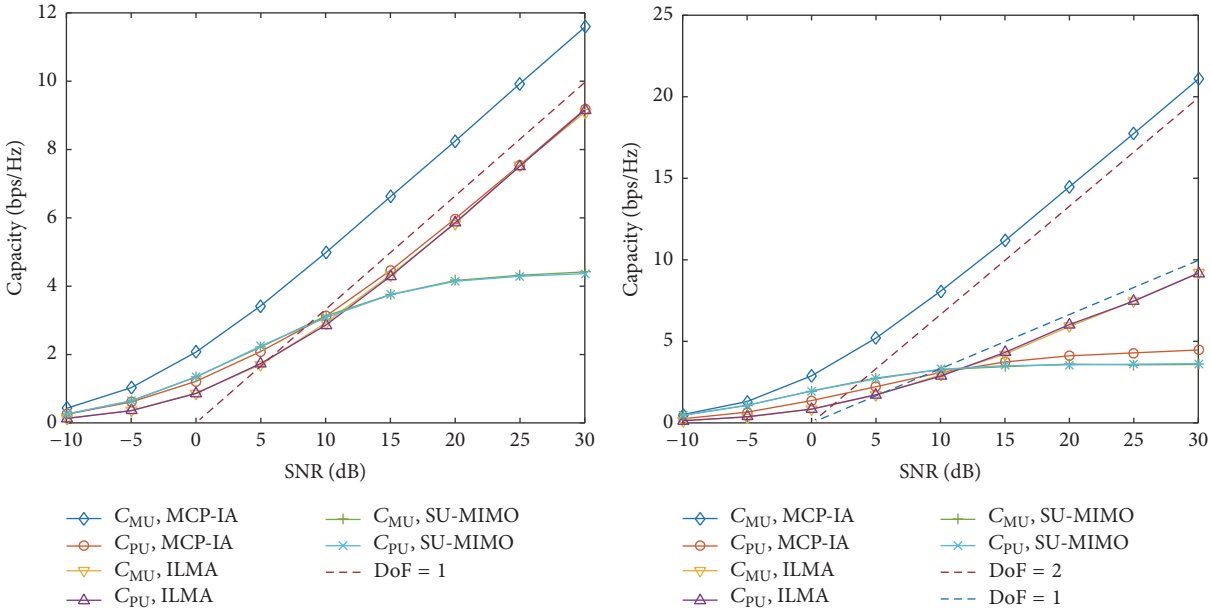
For a 2×2 MIMO system, SU-MIMO utilizes dual-layer spatial multiplexing, while the proposed MCP-IA and ILMA do not. For a 4×2 case, as the DoF of the proposed MCP-IA is two by Lemma 1 and the DoF of ILMA is one by [8], the proposed MCP-IA and SU-MIMO utilize dual-layer spatial multiplexing, while ILMA does not.

The capacities versus the SNR of the UEs for one picocell, two picocells, and four picocells are displayed in Figures 2, 3, and 4, respectively. The diamond-solid line denotes the capacity of the macro UE of the proposed MCP-IA, the circle-solid line denotes that of the pico-UE of the proposed MCP-IA, the inverse triangle-solid line denotes that of the macro UE of ILMA, the triangle-solid line denotes that of the pico-UE of ILMA, the plus-solid line denotes that of the macro UE of SU-MIMO, the x-solid line denotes that of the pico-UE of SU-MIMO, and dashed line denotes the DoF of the IA.

As shown in Figures 2(a) and 2(b), the capacity of ILMA and the proposed scheme increase linearly. The capacity of SU-MIMO does not increase linearly because SU-MIMO is not designed to suppress the interference between the picocell and the macrocell. The capacity of the macro UE of the proposed MCP-IA is greater than that of the other methods. This is because the proposed MCP-IA has diversity and array gains compared to ILMA. The diversity gain is obtained by the selection of the signal space among the orthogonal spaces of macro UE. The array gain is attained by the fact that the precoder is designed for a link between macro BS and macro UE. Recall that the precoders of ILMA are not designed to increase the received signal power, but only to align the interferences of all BSs. In Figure 2(a), for a 2×2 MIMO antenna case, pico-UE of the proposed MCP-IA and ILMA attains the same DoF, but in Figure 2(b), for a 4×2 MIMO antenna case, the proposed MCP-IA achieves a double DoF and a more than a double capacity compared to ILMA. This is because Lemma 1 tells that the proposed MCP-IA with $M \geq 2N$, that is, $M = 4$ and $N = 2$, always achieves $d = N$ DoF for the macro UE, while ILMA has $d \leq N/2$ DoF by (1) in [11] even if it is feasible. In practice, the number of antennas at the BS is greater than that of the UEs, and the proposed MCP-IA is a more effective scheme than ILMA for obtaining a greater capacity.

Table 1: Computational complexity of MCP-IA and ILMA.

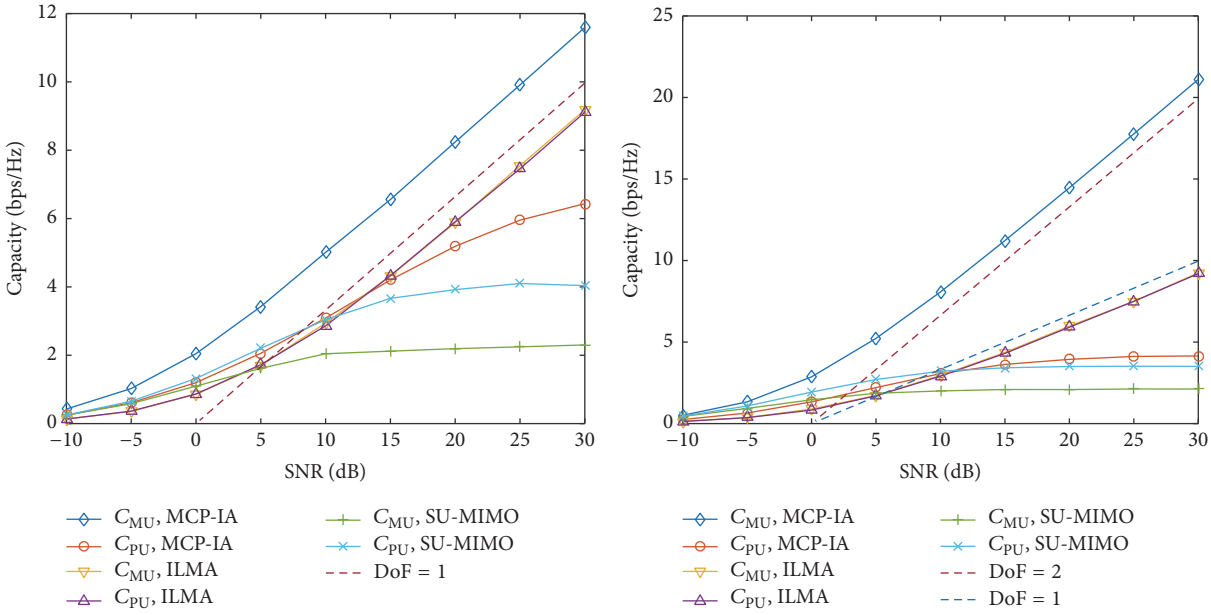
Scheme	Step	Operation	FLOPs	4×2 MIMO
MCP-IA	Compute $\mathbf{F}_1$	SVD with only $\mathbf{\Sigma}$ and $\mathbf{V}$ [26, 27]	$32\left(MN^2 + 2N^3\right)$	1024
	Compute $\mathbf{G}_1$	SVD with $\mathbf{\Sigma}$, $\mathbf{V}$, and $\mathbf{U}$ [26, 27]	$8\left(4M^2N + 8MN^2 + 9N^3\right)$	2624
	Compute $\mathbf{F}_k$	Matrix multiplication [26]	$8dNM - 2dM$	112
		SVD with only $\mathbf{\Sigma}$ and $\mathbf{V}$	$32\left(Md^2 + 2d^3\right)$	1024
	Compute $\mathbf{G}_k$	Matrix multiplication	$4\left\lvert\mathscr{B}_k\right\rvert N^2\left(2d-1\right)$	96
		Matrix inversion [27, 28]	$6\left(2N^3 - N^2 + N\right)$	84
Iteration			*1*	*1*
Total FLOPs				*4964*
ILMA	Compute $\mathbf{Q}_k$	Matrix multiplication	$4N^2\left(K-1\right)\left(2d-1\right)$	32
	Compute $\mathbf{G}_k$	Eigenvalue decomposition [26, 27]	$\frac{16}{3}N^3L$	213
	Compute $\overrightarrow{\mathbf{Q}_k}$	Matrix multiplication	$4M^2\left(K-1\right)\left(2d-1\right)$	128
	Compute $\mathbf{F}_k$	Eigenvalue decomposition	$\frac{16}{3}M^3L$	1706
Iteration			*Convergence is required*	*70 (average)*
Total FLOPs				*145530*



(a) Link capacity of 2 × 2 MIMO antenna configuration

(b) Link capacity of 4 × 2 MIMO antenna configuration

Figure 2: Link capacity of UE with one picocell.



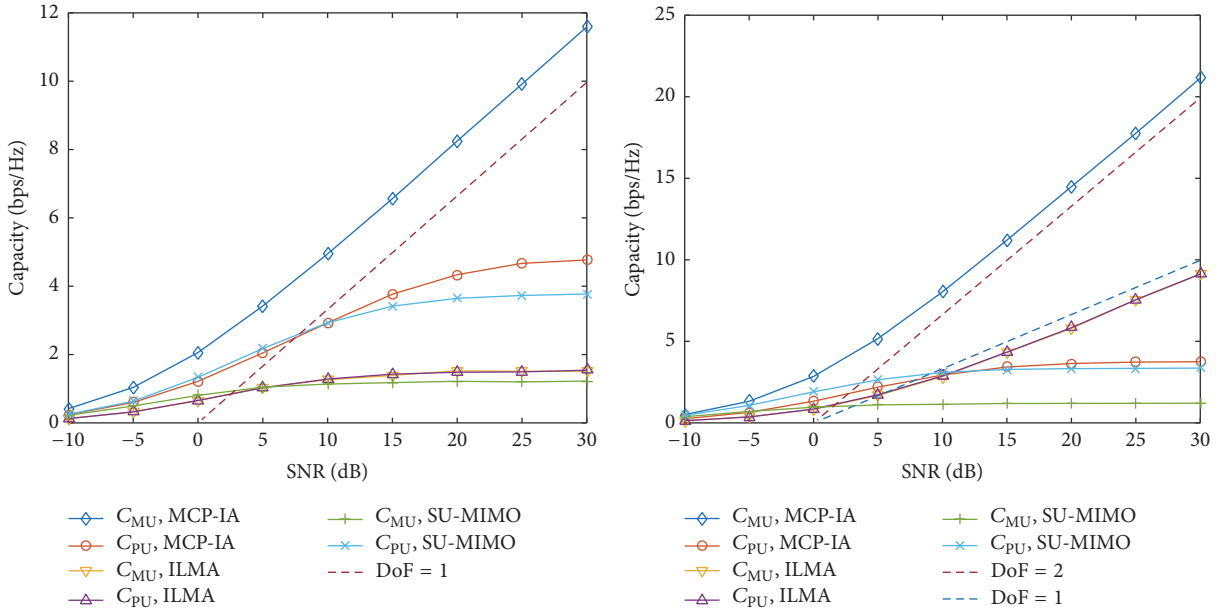
(a) 2 × 2 MIMO antenna configuration

(b) 4 × 2 MIMO antenna configuration

Figure 3: Link capacity of UE with two picocells.

For a two-picocell scenario in Figures 3(a) and 3(b), ILMA still has a feasible solution and the capacities of the macro UE are similar to the one-picocell scenarios. The proposed MCP-IA shows a capacity loss in the pico-UE for an SNR region larger than 15 dB due to the interferences from the other pico-BSs. In practical environments, as the pico-UE is located closer to the pico-BS than the other pico-BSs, the pico-UE always preserves a greater capacity and the capacity loss is negligible. For a 4 × 2 case, the proposed MCP-IA obtains a multiplexing gain that results in a larger capacity than ILMA. This is also seen in the system level simulation.

In Figures 4(a) and 4(b), the capacities versus the SNR of the UEs are displayed for four pico-BSs. In Figure 4(a), the precoder solution of the ILMA for 2 × 2 antenna configuration is not available. The macro UE of ILMA with 2 × 2 MIMO systems shows a capacity loss in an SNR range



(a) 2 × 2 MIMO antenna configuration

(b) 4 × 2 MIMO antenna configuration

Figure 4: Link capacity of UE with four picocells.

Table 2: Simulation parameters.

Parameter	Assumption
Macrocell layout	19 hexagonal cells, 3 sectors per cell
Inter-MBS distance	500 m
Picocell coverage	20 m
Carrier frequency/bandwidth	2 GHz/20 MHz
MBS transmission power	46 dBm
PBS transmission power	23 dBm
Path loss from the MBS to the UE	$128.1 + 37.6 \log_{10} R$ [dB], R in km
Path loss from the PBS to the UE	$140.7 + 36.7 \log_{10} R$ [dB], R in km
MBS antenna pattern	$A(\theta) = -\min\left[12\left(\theta/\theta_{3\,\mathrm{dB}}\right)^2, A_m\right]$
PBS antenna pattern	Omnidirectional
Channel model	ITU-R M.1225 Ped. A
Shadowing standard deviation	8 dB
Penetration loss	20 dB
Noise figure	9 dB
Noise power spectral density	−174 dBm/Hz
Traffic model	Full buffer

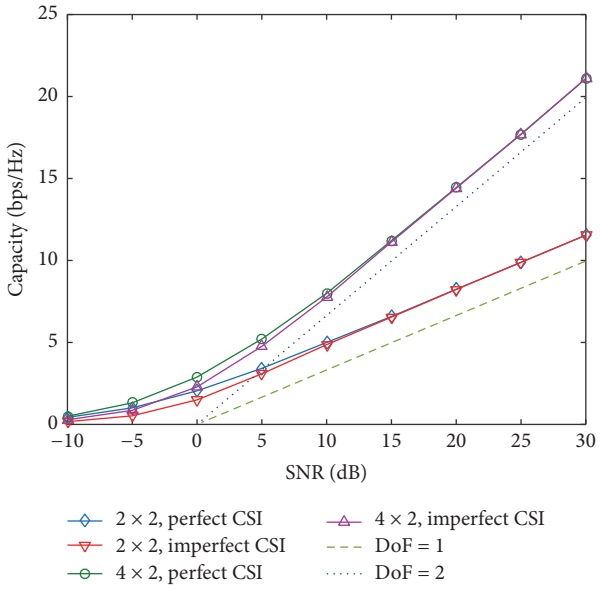


Figure 5: Link capacity of the macro UE with/without channel estimation error.

greater than 10 dB, as does the pico-UE. If the number of picocells increases, ILMA is not feasible and cannot achieve the DoF [9]. The macro UE capacity of the proposed MCP-IA, however, increases linearly even though the number of picocells increases. The pico-UE capacities of the proposed MCP-IA and ILMA do not linearly increase. This is because the proposed MCP-IA does not align all the interferences of the UEs but suppresses the interference of the macro UE from the pico-BSs, as in Figures 2 and 3. Recall that in general heterogeneous networks most picocells are located to cover the shadowing region of the macrocell and are not considerably interfered by the macro BSs. In SU-MIMO case, the capacity is saturated by the cross-tier interferences between the macrocell and the picocells, and it shows the worst capacity result among all the cases. In Figure 4(b) with 4 × 2 antenna configuration, ILMA is feasible and achieves

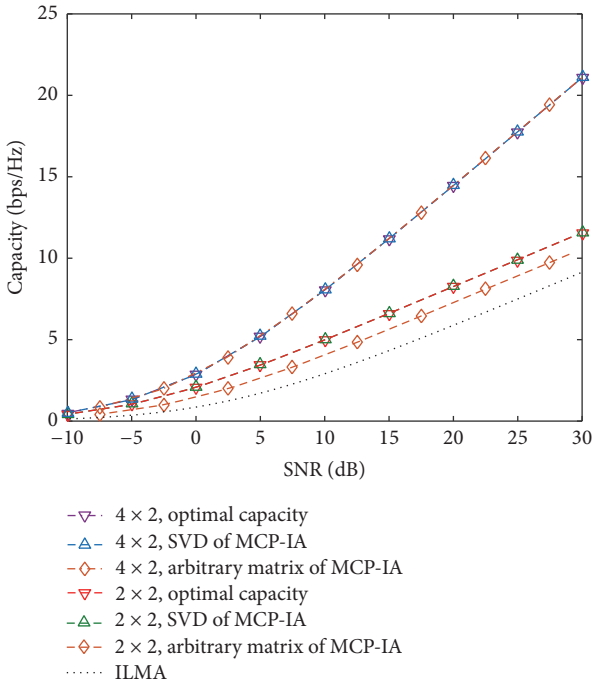


Figure 6: Optimality of capacity comparison for the macro UE.

the DoF for the macro UE and pico-UE. However, as with Figure 3, the DoF and the capacity of the proposed MCP-IA are double those of ILMA and the proposed MCP-IA has additional array and diversity gains.

Figure 5 illustrates the capacity performance of the macro UE of the proposed MCP-IA with the channel estimation error. Least square (LS) is utilized for the channel estimation. In Figure 5, "perfect CSI" and "imperfect CSI" denote the capacity performances without and with the channel estimation error, respectively. The capacity loss by the channel estimation error is calculated in [22, 23]. Even though the channel estimation error occurs, the proposed MCP-IA always achieves the DoF and exhibits the small channel capacity loss which is within the errors derived in [22, 23]. Thus, the proposed MCP-IA is robust to the channel estimation errors.

Figure 6 exhibits capacities by the optimum capacity and SVD with the proposed MCP-IA and arbitrary matrix with the proposed MCP-IA and ILMA. In Figure 6, dashed line and dash-dotted line denote the capacities of 2×2 and 4×2 MIMO antenna configuration, respectively, and dotted line denotes the capacity of ILMA. Let the DoF d be $\min(M/2, N)$. The SVD of MCP-IA and the arbitrary matrix of MCP-IA are obtained in (5). In case of 2×2 MIMO, as seen in Lemma 3, the SVD of MCP-IA demonstrates the same capacity performance as the optimum capacity, and the arbitrary matrix of MCP-IA exhibits smaller capacity. For 4×2 MIMO configuration, all of the proposed MCP-IAs attain the optimal capacity. ILMA, however, has lower capacity than the proposed MCP-IAs. This is because ILMA is not designed for maximizing the capacity but for aligning the interference to achieve the DoF. Therefore, the proposed MCP-IA demonstrates better capacity performance than ILMA.

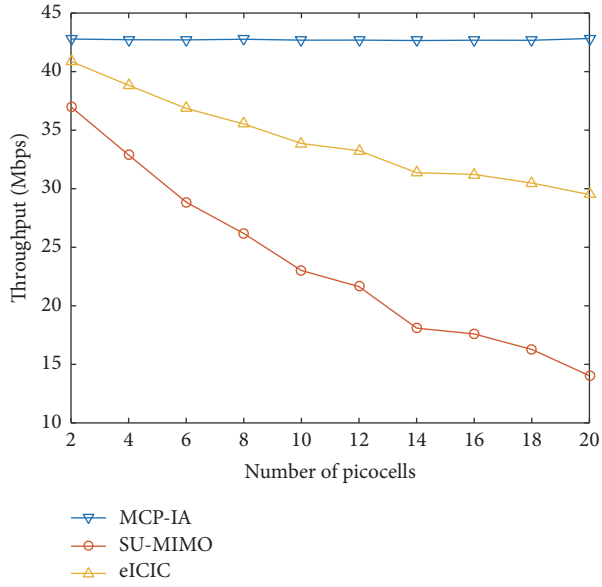
The proposed MCP-IA is designed to protect the macro UEs from the pico-BSs and maximize the capacity. The signal subspace of the macro UE is selected to maximize the capacity and the interference subspace is chosen to align the cross-tier interference. Then, the macro UE achieves the DoF and the optimal capacity. As seen in Figures 2(b), 3(b), and 4(b), for asymmetric antenna configuration such as $M \geq 2N$, the macro UE of the proposed MCP-IA always guarantees $d = N$ DoF by Lemma 1, while ILMA does $d \leq N/2$ DoF even if ILMA is feasible. All the interference suppression schemes have the DoF loss compared to the available DoF of $\min(M, N)$ of the MIMO multiplexing scheme because their DoF are utilized to cancel the interference. The proposed MCP-IA achieves $\min(M/2, N)$ DoF that is greater than $\min(M, N)/2$ DoF of ILMA when the number of antennas of the BS is greater than that of the UEs. The $M \geq 2N$ antenna configuration is common in practical systems and the proposed MCP-IA is more applicable than ILMA.

In the next subsection, the system level simulation is performed to verify the throughputs in randomly deployed picocells and in multiple macrocells.

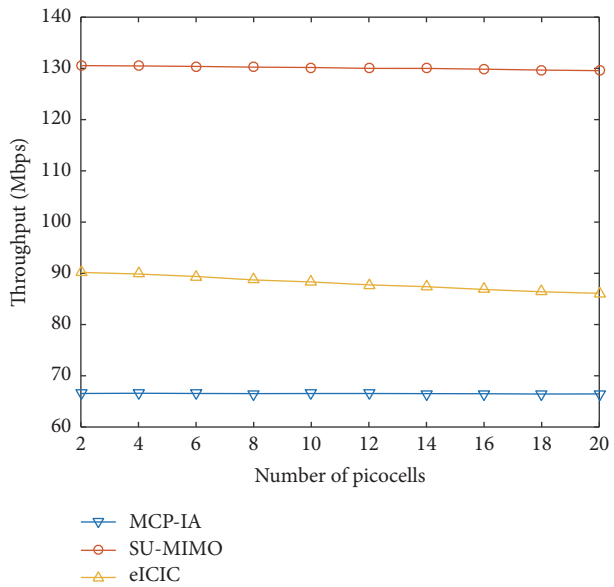
4.2. System Level Performance. In the system level simulation, we measure the throughputs of multiple macrocells with the other cell interference. Since the solutions of the conventional IA schemes such as ILMA with a number of picocells do not exist, eICIC is utilized for comparison. Thus, eICIC, SU-MIMO, and the proposed MCP-IA schemes are evaluated.

We consider that the macro UEs and the pico-UEs move in their cells randomly and that the interference among them occurs by their moving locations. 19 hexagonal macrocells are deployed, each macrocell consists of three sectors, and the number of picocells in a macrocell varies from 2 to 20. The number of macro UEs in a macrocell and pico-UEs in a picocell is set to 20 and 5, respectively. For simplicity, we measure the throughput of the macro UEs and pico-UEs in a center cell. Then, the macro UE and pico-UEs are interfered from the other 18 macrocells. The locations of the macro UE are sometimes close to the picocells resulting in strong interference scenarios. 2×2 antenna configuration and 4×2 antenna configuration are tested. SU-MIMO and the proposed MCP-IA utilize the same spatial symbol dimensions as in the previous link level simulations, and eICIC utilizes dual-layer spatial multiplexing. The other communication parameters for the system level simulations are listed in Table 2. The system level simulations in this paper conform to the evaluation methodology of 3GPP [24].

Figures 7(a) and 7(b) show the throughput of macro UE versus the number of picocells from 2 to 20 for the proposed MCP-IA, eICIC, and SU-MIMO schemes under 2×2 antenna configuration. The throughput denotes an average value of 20 macro UEs for a macrocell. Inverse triangles, triangles, and circles denote the throughputs of the proposed MCP-IA, eICIC, and SU-MIMO schemes, respectively. In Figure 7(a), the proposed MCP-IA demonstrates the maximum throughput performance among the three schemes. Even though the number of picocells increases, the proposed MCP-IA



(a) Macrocell throughput



(b) Per-picocell throughput

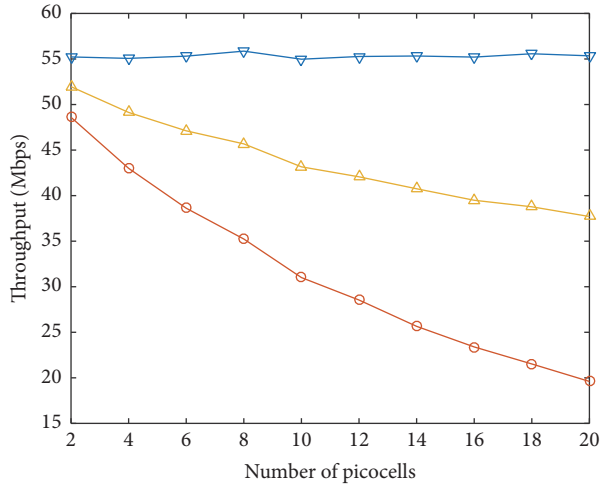
Figure 7: System level performance of the cell throughput corresponding to the number of picocells with 2 × 2 antenna configuration.

maintains the DoF and the maximum throughput of the macro UE, while the throughputs of the other methods decrease. This indicates that the proposed MCP-IA protects against interferences from the pico-BSs completely, even though the number of picocells increases.

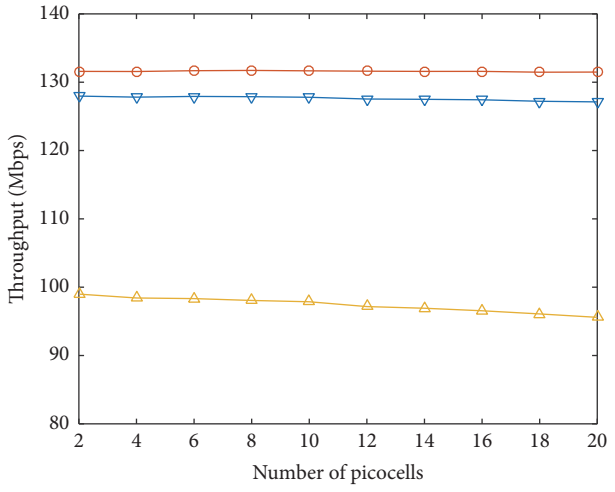
Figure 7(b) displays the throughputs of the picocells versus the number of picocells of the proposed MCP-IA, eICIC, and SU-MIMO. The throughput of the pico-UEs denotes the average value of five pico-UEs in a picocell. The distances between the pico-BS and the pico-UEs are relatively closer than those between the macro BS and the pico-UEs. The cross-tier interference of the pico-UEs is relatively small because of path loss and shadowing, and the cotier interference among the picocells is also small because the density of the picocells in a macrocell is sparse. Therefore, the SNR range of the picocell is greater than that of the macrocell and the average throughput of the picocell is also greater than that of the macrocell. In Figure 7(b), a constant throughput of the proposed MCP-IA demonstrates that the proposed MCP-IA is not affected by the number of picocells and suppresses the interference successfully. The throughput of eICIC and SU-MIMO decreases as the number of picocells increases due to interferences from the other picocells. This is because SU-MIMO and eICIC do not suppress interference and the proposed MCP-IA does. Then, SU-MIMO and eICIC can utilize two spatial streams, and the proposed MCP-IA has one spatial stream with half of the throughput of SU-MIMO. In practice, however, the number of antennas of the BSs is larger than that of UEs, for example, 4 × 2. In this case, the proposed MCP-IA also has two spatial streams and similar throughput of SU-MIMO. This result is shown in Figure 8.

Figure 8 depicts the throughputs versus the number of picocells for 4 × 2 antenna configuration. The throughput of the macrocell in Figure 8(a) shows similar results of the 2 × 2 simulations in Figure 7(a) but increases owing to the large number of transmit antennas. The macrocell throughput of the proposed MCP-IA also demonstrates the largest value among the three schemes and is not affected by the number of picocells. Although the proposed MCP-IA utilizes two spatial streams, the throughput of the macro BS may not be doubled because the SNR regime of the macro UE is low. This result can be seen in Figure 6. At the low SNR region in Figure 6, the capacity of macro UE with the 4 × 2 configuration is not doubled compared to that of 2 × 2. In Figure 8(b), the picocell throughput of the proposed MCP-IA outperforms that of eICIC and approaches that of SU-MIMO. This is because the proposed MCP-IA utilizes two spatial streams and eICIC inefficiently manages its resources due to the ABS subframes. The proposed MCP-IA precoders of the pico-BSs are not designed to improve the capacity of pico-UEs but to align cross-tier interference. The throughput of the proposed MCP-IA is little bit lower than SU-MIMO. However, since the SNR region of the pico-UEs is large, the throughput of the pico-UEs approaches that of SU-MIMO. This throughput gap between the proposed MCP-IA and SU-MIMO is about 3%. Therefore, the proposed MCP-IA provides a better total throughput than eICIC and SU-MIMO in heterogeneous networks.

In 5G systems, larger carrier frequency such as mmWave will be utilized, and many picocells are deployed in indoor environments, and all equipment has MIMO to increase capacity. In the scenarios, cotier interference is relatively small and cross-tier interference is large. This provides relaxed design rules for IA and the proposed MCP-IA is developed. Thus, the proposed MCP-IA is applicable to reduce cross-tier interference of heterogeneous networks in the 5G system.



(a) Macrocell throughput



(b) Per-picocell throughput

FIGURE 8: System level performance of the cell throughput corresponding to the number of picocells for 4×2 antenna configuration.

5. Conclusion

This paper proposes a MCP-IA method for two-tier MIMO downlink heterogeneous networks. The proposed MCP-IA utilizes the concept of IA for the macro UE for protection against cross-tier interference and achieves the same DoF as that of the conventional IA for the macro UE. The pico-UE suppresses the interference from the macro BS using MMSE-IRC. The proposed MCP-IA provides additional array and diversity gains for the macro UE compared to the conventional IA. The array gain of the proposed method is achieved by beamforming precoder for the macro UE and the diversity gain is attained by selecting good signal spaces among the available spaces. The proposed MCP-IA calculates the precoding matrices of the BSs with a closed form. The DoF of the proposed MCP-IA is equal to or greater than the conventional IA and is derived theoretically. The link level simulation results show that the proposed MCP-IA achieves the DoF and the additional array and diversity gains. The system level simulation demonstrates that the proposed MCP-IA suppresses the interference of the macro UE completely and that the pico-UEs maintain the throughput even under the interferences of a large number of picocells. For 4×2 antenna configuration, the system level simulation demonstrates that the proposed MCP-IA obtains the additional multiplexing gain.

Appendix

A. Proof of Lemma 2

The interference-aware ergodic capacity of the macro UE can be written as [25]

$$C_M(\rho) = \mathbb{E}\left[\log_2 \det\left\{\mathbf{I}_d + \frac{\rho}{d}\frac{\mathbf{G}_1^H \mathbf{H}_{11}\mathbf{F}_1\mathbf{F}_1^H\mathbf{H}_{11}^H\mathbf{G}_1}{\mathbf{G}_1^H\left(\mathbf{Q}_1 + \mathbf{I}_N\right)^{-1}\mathbf{G}_1}\right\}\right], \tag{A.1}$$

where

$$\mathbf{Q}_1 = \sum_{j\in\mathscr{B}_1}\frac{P_{1k}}{d}\mathbf{H}_{1k}\mathbf{F}_k\mathbf{F}_k^H\mathbf{H}_{1k}^H \tag{A.2}$$

denotes the interference covariance matrix of the macro UE. In the denominator of the right-hand side of (A.1), $\mathbf{G}_1^H\mathbf{Q}_1\mathbf{G}_1$ becomes a vanishing matrix because

$$\begin{aligned}\mathbf{G}_1^H\mathbf{Q}_1\mathbf{G}_1 &= \mathbf{G}_1^H\left(\sum_{k\in\mathscr{B}_1}\frac{P_{1k}}{\rho d}\mathbf{H}_{1k}\mathbf{F}_k\mathbf{F}_k^H\mathbf{H}_{1k}^H\right)\mathbf{G}_1 \\ &= \sum_{k\in\mathscr{B}_1}\frac{P_{1k}}{\rho d}\left(\mathbf{G}_1^k\mathbf{H}_{1k}\mathbf{F}_k\mathbf{F}_k^H\mathbf{H}_{1k}^H\mathbf{G}_1\right) = \mathbf{0}\end{aligned} \tag{A.3}$$

by (3). Then, $C_M(\rho)$ can be rewritten as

$$C_M(\rho) = \mathbb{E}\left[\log_2 \det\left\{\mathbf{I}_d + \frac{\rho}{d}\mathbf{G}_1^H\mathbf{H}_{11}\mathbf{F}_1\mathbf{F}_1^H\mathbf{H}_{11}^H\mathbf{G}_1\right\}\right]. \tag{A.4}$$

Substituting $\overline{\mathbf{H}} = \mathbf{G}_1^H\mathbf{H}_{11}\mathbf{F}_1$ into $C_M(\rho)$,

$$\begin{aligned}C_M(\rho) &= \mathbb{E}\left[\log_2 \det\left\{\mathbf{I}_d + \frac{\rho}{d}\overline{\mathbf{H}}\,\overline{\mathbf{H}}^H\right\}\right] \\ &= E\left[\sum_{i=1}^{d}\log_2\left(1 + \frac{\rho\lambda_i}{d}\right)\right],\end{aligned} \tag{A.5}$$

where λ_i denotes the ith largest eigenvalue of $\overline{\mathbf{H}}$. By the definition of the DoF,

$$\begin{aligned}\lim_{\rho\to\infty}\frac{C_M(\rho)}{\log_2(\rho)} &= \lim_{\rho\to\infty}\mathbb{E}\left[\sum_{i=1}^{d}\frac{\log_2(1+\rho\lambda_i/d)}{\log_2(\rho)}\right]\\ &= \lim_{\rho\to\infty}\mathbb{E}\left[\sum_{i=1}^{d}\frac{\log_2(\rho\lambda_i/d)+\mathcal{O}(1)}{\log_2(\rho)}\right]\\ &= \lim_{\rho\to\infty}\sum_{i=1}^{d}\frac{\mathbb{E}[\log_2(\rho\lambda_i/d)+\mathcal{O}(1)]}{\log_2(\rho)} \qquad \text{(A.6)}\\ &\approx \lim_{\rho\to\infty}\sum_{i=1}^{d}\frac{\log_2(\rho)+\mathcal{O}(1)}{\log_2(\rho)}\\ &\approx \lim_{\rho\to\infty}\frac{d\log_2(\rho)+\mathcal{O}(1)}{\log_2(\rho)} = d.\end{aligned}$$

B. Proof of Lemma 3

Let us consider an arbitrary complex matrix $\mathbf{A}\in\mathbb{C}^{m\times n}$ and a unitary matrix $\mathbf{R}\in\mathbb{C}^{m\times p}$. If $\mathbf{R}$ is square, that is, $p=m$, $\mathbf{R}^H\mathbf{A}$ can be regarded as the rotation of $\mathbf{A}$. Then, the singular values of $\mathbf{R}^H\mathbf{A}$ are the same as $\mathbf{A}$. If $p<m$, however, $\text{rank}(\mathbf{R}^H\mathbf{A})<\text{rank}(\mathbf{A})$ and those singular values may be varied. In this case, to obtain the same singular values of $\mathbf{A}$, the columns of $\mathbf{R}^H\mathbf{A}$ should be the left singular vectors of $\mathbf{A}$.

For the case that $M<2N$, the achievable DoF is $d=\min(M/2,N)<M$ and N, and $\mathbf{G}_1$ in (10) is a nonsquare $N\times d$ matrix. Therefore, for optimal $\mathbf{F}_1$, the columns of $\mathbf{G}_1$ should be the left singular vectors of $\mathbf{H}_{11}$. For $M\geq 2N$, the achievable DoF is $d=N$, and $\mathbf{G}_1$ is a square $N\times N$ matrix. In this case, $\mathbf{F}_1$ is always optimal for any unitary square matrix $\mathbf{G}_1$ because rank and singular values of $\mathbf{G}_1^H\mathbf{H}_{11}$ are invariant to those of $\mathbf{H}_{11}$. Therefore, the proposed MCP-IA is always optimal for the macro UE for any antenna configuration with $M\geq 2N$.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was a part of the project titled "Development of Distributed Underwater Monitoring & Control Networks," funded by the Ministry of Oceans and Fisheries, Republic of Korea.

References

[1] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: a survey," *IEEE Communications Magazine*, vol. 46, no. 9, pp. 59–67, 2008.

[2] C. Patel, M. Yavuz, and S. Nanda, "Femtocells [industry perspectives]," *IEEE Wireless Communications*, vol. 17, no. 5, pp. 6–7, 2010.

[3] A. Damnjanovic, J. Montojo, Y. Wei et al., "A survey on 3GPP heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 10–21, 2011.

[4] B. Soret, H. Wang, K. I. Pedersen, and C. Rosa, "Multicell cooperation for LTE-advanced heterogeneous network scenarios," *IEEE Wireless Communications*, vol. 20, no. 1, pp. 27–34, 2013.

[5] D. López-Pérez, I. Güvenç, G. De La Roche, M. Kountouris, T. Q. S. Quek, and J. Zhang, "Enhanced intercell interference coordination challenges in heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 22–30, 2011.

[6] I. Hwang, C.-B. Chae, J. Lee, and R. W. Heath Jr., "Multicell cooperative systems with multiple receive antennas," *IEEE Wireless Communications*, vol. 20, no. 1, pp. 50–58, 2013.

[7] S. Sun, Q. Gao, Y. Peng, Y. Wang, and L. Song, "Interference management through CoMP in 3GPP LTE-advanced networks," *IEEE Wireless Communications*, vol. 20, no. 1, pp. 59–66, 2013.

[8] V. R. Cadambe and S. A. Jafar, "Interference alignment and degrees of freedom of the K-user interference channel," *IEEE Transactions on Information Theory*, vol. 54, no. 8, pp. 3425–3441, 2008.

[9] C. M. Yetis, T. Gou, S. A. Jafar, and A. H. Kayran, "On feasibility of interference alignment in MIMO interference networks," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4771–4782, 2010.

[10] L. Ruan, V. K. N. Lau, and M. Z. Win, "The feasibility conditions for interference alignment in MIMO networks," *IEEE Transactions on Signal Processing*, vol. 61, no. 8, pp. 2066–2077, 2013.

[11] K. Gomadam, V. R. Cadambe, and S. A. Jafar, "A distributed numerical approach to interference alignment and applications to wireless interference networks," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3309–3322, 2011.

[12] D. C. Moreira, Y. C. B. Silva, K. Ardah, W. C. Freitas, and F. R. P. Cavalcanti, "Convergence analysis of iterative interference alignment algorithms," in *Proceedings of the International Telecommunications Symposium (ITS '14)*, pp. 1–5, São Paulo, Brazil, August 2014.

[13] S. Chen and R. S. Cheng, "Clustering for interference alignment in multiuser interference network," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 6, pp. 2613–2624, 2014.

[14] B. Guler and A. Yener, "Selective interference alignment for MIMO cognitive femtocell networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 3, pp. 439–450, 2014.

[15] D. Castanheira, A. Silva, and A. Gameiro, "Set optimization for efficient interference alignment in heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 10, pp. 5648–5660, 2014.

[16] G. Gupta and A. K. Chaturvedi, "User selection in MIMO interfering broadcast channels," *IEEE Transactions on Communications*, vol. 62, no. 5, pp. 1568–1576, 2014.

[17] H. Okamoto, K. Kitao, and S. Ichitsubo, "Outdoor-to-indoor propagation loss prediction in 800-MHz to 8-GHz band for an urban area," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 3, pp. 1059–1067, 2009.

[18] S. Y. Lim, Z. Yun, and M. F. Iskander, "Propagation measurement and modeling for indoor stairwells at 2.4 and 5.8 GHz," *IEEE Transactions on Antennas and Propagation*, vol. 62, no. 9, pp. 4754–4761, 2014.

[19] G. R. MacCartney, T. S. Rappaport, S. Sun, and S. Deng, "Indoor office wideband millimeter-wave propagation measurements and channel models at 28 and 73 GHz for ultra-dense 5G wireless networks," *IEEE Access*, vol. 3, pp. 2388–2424, 2015.

[20] V. Chandrasekhar, M. Kountouris, and J. G. Andrews, "Coverage in multi-antenna two-tier networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 10, pp. 5314–5327, 2009.

[21] D. C. Moreira, Y. C. B. Silva, K. Ardah, W. C. Freitas Jr., and F. R. P. Cavalcanti, "Convergence analysis of iterative interference alignment algorithms," in *Proceedings of the International Telecommunications Symposium (ITS '14)*, São Paulo, Brazil, August 2014.

[22] D. Samardzija and N. Mandayam, "Pilot-assisted estimation of MIMO fading channel response and achievable data rates," *IEEE Transactions on Signal Processing*, vol. 51, no. 11, pp. 2882–2890, 2003.

[23] S. J. Lee, "Effect of least square channel estimation errors on achievable rate in MIMO fading channels," *IEEE Communications Letters*, vol. 11, no. 11, pp. 862–864, 2007.

[24] 3GPP TR 36.814 ver. 9.0.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects," March 2010.

[25] R. S. Blum, "MIMO capacity with interference," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 793–801, 2003.

[26] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins Studies in the Mathematical Sciences, The Johns Hopkins University Press, Baltimore, Md, USA, 3rd edition, 1996.

[27] K. Zu and R. C. de Lamare, "Low-complexity lattice reduction-aided regularized block diagonalization for MU-MIMO systems," *IEEE Communications Letters*, vol. 16, no. 6, pp. 925–928, 2012.

[28] K. Zu, R. C. de Lamare, and M. Haardt, "Generalized design of low-complexity block diagonalization type precoding algorithms for multiuser MIMO systems," *IEEE Transactions on Communications*, vol. 61, no. 10, pp. 4232–4242, 2013.

Hindawi Publishing Corporation
Mobile Information Systems
Volume 2016, Article ID 3798523, 17 pages
http://dx.doi.org/10.1155/2016/3798523

Research Article

Playing Radio Resource Management Games in Dense Wireless 5G Networks

Paweł Sroka and Adrian Kliks

The Chair of Wireless Communications, Poznan University of Technology, Poznan, Poland

Correspondence should be addressed to Paweł Sroka; pawel.sroka@put.poznan.pl

Received 1 August 2016; Revised 3 October 2016; Accepted 1 November 2016

Academic Editor: Ioannis Moscholios

Copyright © 2016 P. Sroka and A. Kliks. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper considers the problem of an efficient and flexible tool for interference mitigation in ultradense heterogeneous cellular 5G networks. Several game-theory based approaches are studied, focusing on noncooperative games, where each base station in the end tries to maximize its payoff. An analysis of backhaul requirements of investigated approaches is carried out, with a proposal of a mechanism for backhaul requirements reduction. Moreover, improvements in terms of energy use optimization are proposed to further increase the system gains. The presented simulation results of a detailed ultradense 5G wireless system show that the discussed game-theoretic approaches are very promising solutions for interference mitigation outperforming the algorithm proposed for LTE-Advanced in terms of the achieved spectral efficiency. Finally, it is proved that the introduction of energy-efficient and backhaul-optimized operation does not significantly degrade the performance achieved with the considered approaches.

1. Introduction

As increase in signal coverage, throughput guarantees of efficient service delivery, and finally highly effective spectrum utilization were said to be the main goals of the second, third, and fourth generations of cellular systems [1, 2], it is the integration of various networks with energy efficiency that is foreseen as the key factor for further development of wireless networks. Higher internetwork integration results in higher steering/control data exchange, which influences traffic growth in backhaul and core networks. At the same time, the topic of energy efficiency in future wireless networks became a significant research area in recent years. It is of high interest for mobile network operators (MNOs), since by the application of sophisticated solutions, overall energy consumption can be improved, leading to a further reduction of operational expenditures [2–5].

1.1. Related Work. In general, the minimization of energy consumption has been considered in various scenarios, including cellular networks (e.g., where optimization among all access network [6, 7], backhauling, and core networks [8, 9] can be considered), other noncellular wireless networks (i.e., noncellular networks [10–12]), and wired networks (including optical networks [13–15]). Various solutions have been proposed, targeting different aspects of communications and networking, such as advanced radio resource and interference management enabling throughput enhancement or transmit power reduction (e.g., [16–18]), turning selected network nodes into sleep mode (e.g., base stations, optical modules [19–21]) depending on the traffic requirements, or energy-efficient routing (e.g., [22, 23]). One can observe that the holistic view of all network elements is necessary in order to assess real gains that can be achieved by the application of selected energy-aware solutions. Indeed, it is probable that all benefits observed in one place as a result of the application of a selected algorithm can be lost by the increase of energy consumption in another place. In other words, it is important to analyze as wide spectrum of aspects as possible during the development of any new energy-efficient solution.

1.2. Scope and Novelty. In this work, we concentrate on the application of advanced radio resource and interference management algorithms for dense urban wireless networks (access network) which maximize the total cell throughput and optimize the energy usage by the base stations. The detailed and accurate simulation scenario proposed in the EU METIS2020 project for such an environment has been

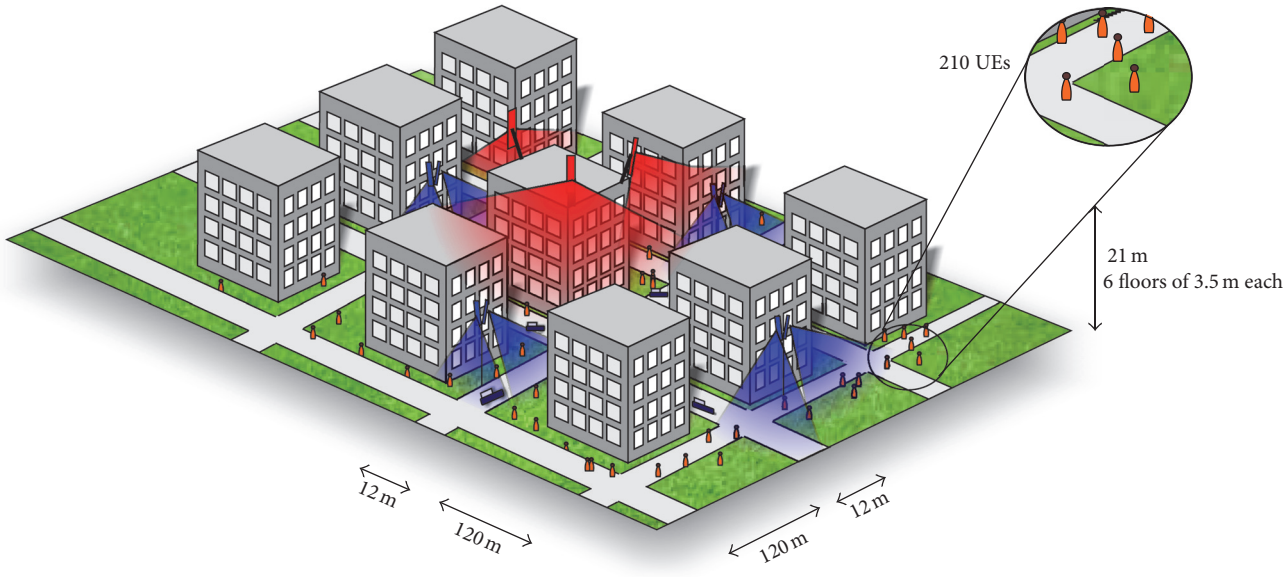


Figure 1: Considered scenario: Manhattan-like dense urban network.

selected as the enabler for making simulations close to reality [24]. The algorithms for interference coordination proposed for 4G networks (known as enhanced intercell interference coordination (eICIC) [25–27]) have been compared with our solutions based on the application of game-theoretic tools. Let us note that the game theory has already been considered many times as a valuable tool in the context of cellular networks. In our work, we tried to analyze the effectiveness of the proposed game-theoretic tools from the point of view of their practical implementation. Thus, we not only discuss the convergence time (which results in a delay in the system) but also analyze how the traffic observed in the backhaul and core networks will be affected by the application of our algorithms. The novel aspects covered by this paper are the following:

(a) Provision of new solutions for radio resource and interference management in dense urban 5G networks that maximize cell throughput with the minimization of overall energy consumption, taking into account the traffic increase in the backhaul network

(b) Summary and definitions of various game-theoretic tools that can be used for achieving this goal and their comparison with the eICIC technique (please note that scenarios with and without network node coordination have been considered)

(c) Analysis of the influence of applying the proposed tool on backhaul traffic and the potential increase of the energy consumed in that part of the network

Let us stress that the main aim of this paper is to compare different slow-time-scale interference management mechanisms that can be considered as candidates for future 5G networks, assuming an identical simulation setup. However, we do not intend to use identical game and auction models for every case, as different definitions can be assumed depending on the game type and optimization criteria.

The paper is organized as follows. First, the system model considered for experimentation is discussed in Section 2. Once it is presented, the proposal for energy-efficient gaming in such a scenario is provided. Then, Section 3 presents detailed game definitions together with suggested ways of achieving the equilibrium expected for each game. The analysis of backhaul traffic, being often a heavy burden of new sophisticated algorithms, such as the coordinated multipoint (CoMP), is provided in Section 4, while the achieved computer simulation results are shown and described in Section 5. Finally, the paper is concluded.

2. System Model of a Dense Network

In order to analyze the efficiency of the proposed algorithms in reducing energy consumption in the context of next-generation networks, a dense urban scenario has been selected for consideration, where multiple outdoor user terminals communicate with macro or micro base stations deployed on buildings and managed by MNO.

2.1. Detailed Model Characterization. In this work, the use-case developed in the METIS 2020 project [24] has been considered, where a Manhattan-like dense urban wireless network is modeled. It is assumed that mobile users utilize the orthogonal frequency division multiple access (OFDMA) technology with the frequency reuse factor equal to 1 to communicate with one of the sectoral antennas deployed by MNO on surrounding buildings, as shown in Figure 1. One can observe that $N = 21$ antennas are installed in the considered area, with three main 120° sectoral macro antennas mounted on the central-building and 9 pairs of 120° sectoral micro antennas deployed close to the neighboring buildings. As the macro base stations are mounted on rooftops (5-meter high masts have been used), the micro base

stations' antennas are located at the heights of 10 m and 3 m separated from the building wall. Each building contains six floors (3.5 m high each) and has a square base of the size of 120 m × 120 m. The streets' width including both sidewalks and lanes is set to 12 m. The number of full-buffer, static, and uniformly distributed user equipment (UE) pieces in the considered area was set to $J = 520$. It means that the average number of users served by one base station $\widehat{J}$ is equal to $\widehat{J} = 520/21 = 24.8$.

The bandwidth available for each of the base stations (BSs) is divided into time-frequency blocks, with the BSs transmitting in each of the blocks using one of the selected power per subcarrier levels. These levels are selected out of the set $P = \{p_{\text{low}}, p_{\text{high}}\}$, as shown in Figure 2.

Let us now describe the scenario using mathematical formalism. The set of all users is represented hereafter as J, with J_i denoting the set of users served by BS i. At each time interval, each BS divides the available resources among up to 10 UE pieces according to the proportional fairness (PF) rule. Let $|h_{i,j}^{(s)}|^2$ denote the channel gain between the ith BS and jth UE on subcarrier s ($h_{i,j}^{(s)} \in \mathbb{C}$), and let σ_j^2 be the noise variance at receiver j (we assume that each mobile user possesses the knowledge on channel attenuation based on the observed pilot signals from all base stations; this information can be then delivered to the serving base station). The signal-to-interference plus noise ratio (SINR) for UE j served by BS i on subcarrier s is given as follows:

$$\gamma_{i,j}^{(s)} = \frac{\left|h_{i,j}^{(s)}\right|^2 p_i^{(s)}}{\sigma_j^2 + \sum_{l \in \{M \cup K\} \setminus i} \left|h_{l,j}^{(s)}\right|^2 p_l^{(s)}}, \tag{1}$$

where $p_i^{(s)}$ denotes the transmit power of BS i on subcarrier s and $\{M \cup K\}$ represents the set of base stations (players).

In our work, we assume that all BSs are interested in achieving at least the minimum throughput $T_{\min}$ and minimizing the operating costs expressed by the total consumed energy. The throughput achieved by the jth UE served by BS i can be then calculated as

$$T_{i,j} = \sum_t \sum_s R_{i,j}^s(t) \cdot \zeta_{i,j}^s(t). \tag{2}$$

Here $\zeta_{i,j}^s(t)$ represents the allocation of subcarrier s at time t to UE j by BS i, with $\zeta_{i,j}^s(t) \in \{0, 1\}$, and the rate of UE j served by BS i on subcarrier s, $R_{i,j}^s$, is calculated depending on the transmitted transport block size determined as specified in the 3GPP Long Term Evolution (LTE) specification [28].

2.2. Backhaul Connection. All BSs can exchange control and signaling information using a dedicated interface. For further backhaul traffic analysis (see Section 4) it is arbitrarily assumed that each micro base station is fiber-connected directly to the macro base station, as illustrated in Figure 3. It is, for example, the fiber-to-the-antenna together with radio-over-fiber technology that can act as the technical enabler for the realization of such a backhaul network. Please note, however, that other realizations of backhaul connections are possible (wired but nonfiber, wireless, etc.), but such a detailed analysis is out of the scope of this paper. Let us note that backhaul load optimization is not directly included in any of the compared interference management solutions, with backhauling considered only as a contribution to overall energy consumption.

2.3. Energy-Efficient Gaming in Dense Networks. Spectrum and energy efficiency are the key figures of merit in the context of next-generation networks, especially in the case of dense user (UE) deployment. Referring to the former requirement, in the considered scenario, all base stations share the same frequency spectrum; that is, a full frequency reuse case is implemented together with advanced interference management solutions among base stations and served users. In the context of contemporary cellular networks, such techniques as intercell interference coordination (ICIC) in LTE or eICIC with almost-blank subframes (ABS) in Long Term Evolution-Advanced (LTE-A) have been proposed [25–27]. We would like to concentrate on a solution that will achieve at least the same efficiency level as the nowadays solutions. Thus, among various proposals for such radio resource management (RRM), we select game-theoretic tools where each wireless network entity is treated as a player in a specified game. These tools have already been proved to be effective in RRM in various scenarios, for example, [29–34].

In this work, we have decided to utilize game-theoretic tools in a practical scenario but focusing on achieving both spectrum and energy efficiency. In particular, the following assumptions have been made: first, the game is played among base stations (mobile users do not participate actively in that game) in a noncooperative way; second, the same set of game strategies is defined for each player; third, the game and players' payoffs are defined in a way that promotes energy efficiency while achieving throughput/rate better than or at least comparable to eICIC solutions. Following the approach known from LTE-A RRM algorithms, we assume that energy efficiency can be achieved by the adaptive assignment of power levels and radio resource blocks among users, depending on their position and effective signal-to-noise ratio. Taking into account the assumptions, 16 transmission strategies have been identified. They define the transmit power levels on certain frequency subbands that might be selected by BSs, as illustrated in Figure 4.

First, the selection of the given strategy will be made for the period of time of one frame, which consists of 10 subframes (also known as transmission time interval (TTI)) of 1 ms each; thus the potential change of the power allocation scheme in the time-frequency plane will be made every 10 ms. In our proposal, the considered frequency band of $N_{\text{RB}} = 100$ resource blocks (RBs) (equivalent to 20 MHz) is further divided into three subbands of 33, 33, and 34 RBs, respectively. In other words, the bandwidth of these subbands equals $33 \cdot 200$ kHz = 6.6 MHz and $34 \cdot 200$ kHz = 6.8 MHz. Each player (macro or micro base station) can transmit with either base or reduced transmit power, denoted as $P_{\text{TX},b}$ and $P_{\text{TX},r}$. Sixteen power allocation schemes among 10 TTIs have been selected, as represented in Figure 4. It has been arbitrarily assumed that the value of the reduced transmit power is 10 dBm lower

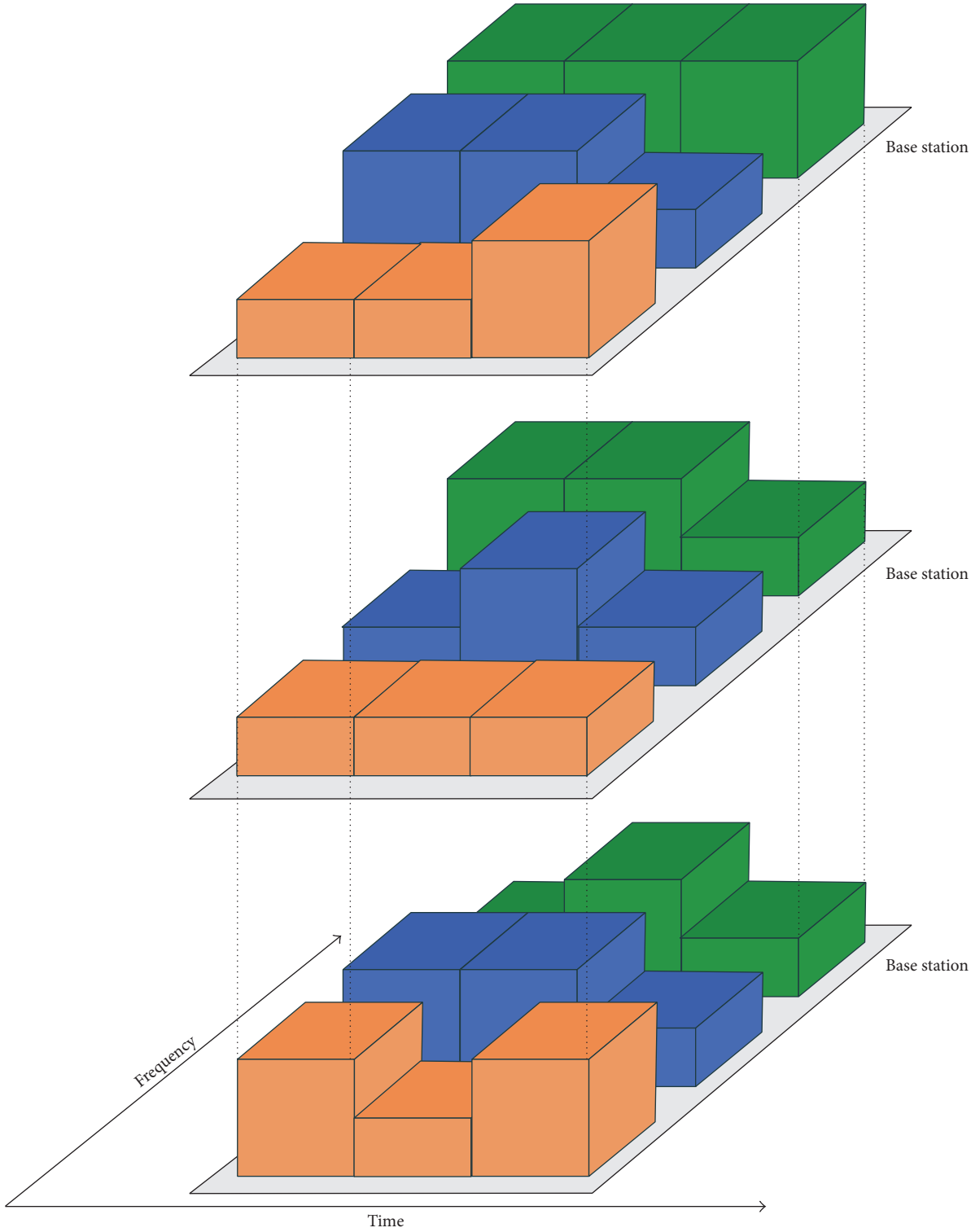


Figure 2: Resource allocation among three base stations.

than the base transmit power in the entire 10 MHz width frequency band. Furthermore, it has been decided that the total transmit power for a macro base station is set to 46 dBm (or equivalently 26 dBm per one RB) and for a micro base station 33 dBm (or equivalently 13 dBm per one RB).

One can observe that in such an approach the energy efficiency will be achieved by the advanced assignment of selected strategies among playing base stations. However, as it will be discussed later, the application of such an approach results in a high need of distributing detailed and accurate

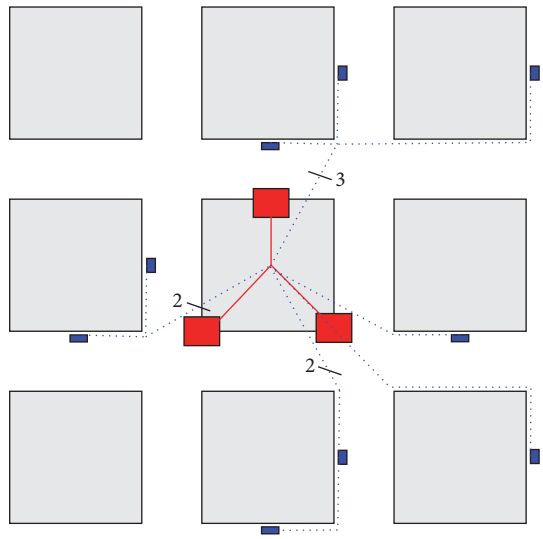


Figure 3: Exemplary backhaul connection.

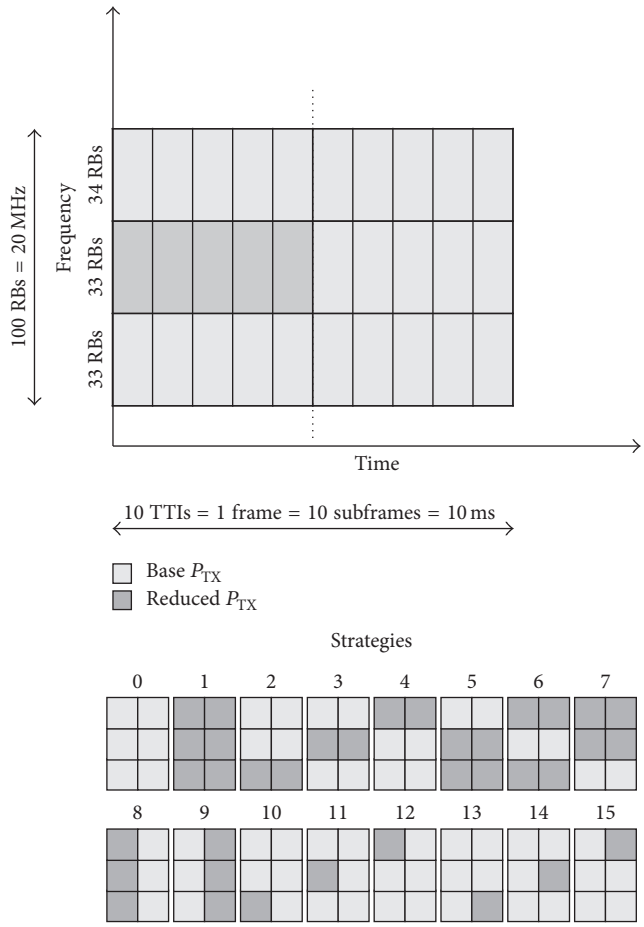


Figure 4: Class of strategies considered in game definitions.

context-information between network nodes (players). For example, the exchange of the exact values of signal-to-noise ratios between the base station and each user could be necessary. In our work, we will also discuss this practical aspect of our solution; that is, a backhaul traffic analysis will be provided, as the inclusion of this part of the network is crucial for a fair overall analysis of energy efficiency.

In order to assess the energy efficiency of the proposed solutions, we express this variable arbitrarily in terms of the average number of bits per Hz per Joule per cell. The second metric used for the assessment of the algorithm quality is the total power consumed in a given period of time.

3. Energy Efficiency Game Definition

3.1. Game Definition. The problem of intercell interference mitigation can be described using the following normal-form game definition:

$$\mathbb{G} = \left(\{M \cup K\}, A, \{U_i\}_{i\in\{M\cup K\}}\right). \tag{3}$$

Let us assume that, at each time instant t, BS i selects its action from a finite set A, following the probability distribution

$$\pi_i(t) = \left(\pi_i^{(\alpha_i^{(1)})}(t), \pi_i^{(\alpha_i^{(2)})}(t), \ldots, \pi_i^{(\alpha_i^{(N)})}(t)\right), \tag{4}$$

where $\pi_i^{(\alpha_i^{(n)})}(t)$ denotes the probability that BS i plays action $\alpha^{(n)}$. As a result of playing one of the strategies, the ith base station will receive a payoff, denoted hereafter as $U_i(\alpha_i^{(k)})$. Such a payoff can be defined as, for example, total throughput observed by the base station reduced by the costs that this base station has to pay for playing this strategy (e.g., energy consumption). Thus, in general, the aim of each BS is to maximize its payoff with or without cooperation with other BSs, achieving the so-called game equilibrium. In what follows, we will extend and adapt this simple model according to the requirements referred to the selected equilibrium.

3.2. Selected Equilibria

3.2.1. Nash Equilibrium. One of the most popular concepts is the one known as the Nash equilibrium. When the base station plays the Nash equilibrium strategy denoted as α_i^* [35, 36], the following relation will hold:

$$U_i(\alpha_i^*, \alpha_{-i}) \geq U_i(\alpha_i, \alpha_{-i}), \quad \forall i \in S, \tag{5}$$

where α_i represents the possible strategy of the ith BS, whereas α_{-i} defines the set of strategies chosen by the other BSs; that is, $\alpha_{-i} = \{\alpha_j\}$, $j \neq i$, and S is the BSs set of cardinality n. The idea behind the Nash equilibrium is to find the point of an achievable rate region (which is related to the selection of one of the available strategies), from which any player cannot increase its utility (increase the total payoff) without reducing other players' payoffs.

Payoff Definition. The achievement of the Nash equilibrium will be considered as an example of a noncooperative game, where base stations optimize their own periods and do not consider the overall system performance. Moreover, we assume that no additional information needs to be sent between the players. In such a case, the utility or better the payoff of the BS i will be defined as the rate observed by this player:

$$U_i(\alpha_i, \boldsymbol{\alpha}_{-i}) \triangleq R_i(\alpha_i, \boldsymbol{\alpha}_{-i}). \tag{6}$$

Boltzmann-Gibbs Distribution. An immediate question arises: how to guarantee the achievement of the Nash equilibrium. One of the well-known learning approaches, effective in terms of the speed of convergence, is to use the logit learning, also known as Boltzmann-Gibbs learning. The equilibrium achieved following the Boltzmann-Gibbs rule is a special case of the ϵ-Nash equilibrium that can be learned in a fully distributed manner [37]. The Boltzmann-Gibbs distribution can be described using

$$\beta_{i,\epsilon}(U_{i,t}(\alpha_i, \boldsymbol{\alpha}_{-i})) = \frac{e^{(1/\epsilon_i)U_{i,t}(\alpha_i, \boldsymbol{\alpha}_{-i})}}{\sum_{\alpha_i^*\in A_i} e^{(1/\epsilon_i)U_{i,t}(\alpha_i^*, \boldsymbol{\alpha}_{-i})}}, \tag{7}$$

where, for player i, $\beta_{i,\epsilon}(U_{i,t}(\alpha_i, \boldsymbol{\alpha}_{-i}))$ is the probability of choosing action α_i at time t, A_i is the set of available actions, and, for $\epsilon_i = \epsilon$, $\forall i$, the value $1/\epsilon$ is interpreted as the temperature parameter that impacts the convergence speed.

Strategy Identification. In our work we are comparing various equilibria. In order to ensure the clarity and readability in each section, we provide a unique identifier used further on in the text with reference to a certain scenario. Hereafter, the Nash equilibrium and the scenario described in this subsection will be referred to as NE.

3.2.2. Correlated Equilibrium. Let us now focus on the idea of the correlated equilibrium, where, in a nutshell, the joint probability of performing selected actions by players is taken into account [38]. Contrary to the Nash equilibrium, the achievement of the correlated equilibrium assumes in our case active information exchange among players. In general, at each time instant, each BS plays one of the N strategies $\alpha^{(n)}$, $1 \leq n \leq N$. Therefore, assuming that set A is discrete and finite, at least one equilibrium exists that represents the system state when a player cannot improve its payoff (utility) when other players do not change their behavior. Such a state is known as the correlated equilibrium (CE), which is defined as follows:

$$\sum_{\boldsymbol{\alpha}_{-i}\in A} \pi(\alpha_i^*, \boldsymbol{\alpha}_{-i})\left(U_i(\alpha_i^*, \boldsymbol{\alpha}_{-i}) - U_i(\alpha_i', \boldsymbol{\alpha}_{-i})\right) \geq 0, \quad \forall \alpha_i', \alpha_i^* \in A, \ \forall i \in \{M \cup K\}. \tag{8}$$

In (8), $\pi(\alpha_i^*, \boldsymbol{\alpha}_{-i})$ is the probability of playing strategy α_i^* in the case when other BSs select their own strategies α_j, $j \neq i$. Probability distribution π is a joint point mass function of the different combinations of BSs strategies. As in [29], the inequality in the correlated equilibrium definition means that when the recommendation for BS i is to choose action α_i^*,

then choosing any other action instead of α_i^* cannot result in a higher expected payoff for this BS.

Payoff Definition. Let us formulate the set of actions selected by all BSs as $\boldsymbol{\alpha} = \{\alpha_i \cup \boldsymbol{\alpha}_{-i}\}$, where $\boldsymbol{\alpha}_{-i}$ is a set of actions selected by all BSs other than i. We can introduce the rate-dependent Vickrey-Clarke-Groves (VCG) [30] auction mechanism, where each BS aims to maximize utility U_i, $\forall i$, defined as

$$U_i\left(\alpha_i, \boldsymbol{\alpha}_{-i}\right) \triangleq R_i\left(\alpha_i, \boldsymbol{\alpha}_{-i}\right) - \zeta_i\left(\alpha_i, \boldsymbol{\alpha}_{-i}\right), \tag{9}$$

where ζ_i denotes the cost (rate loss) introduced by BS i to all other BSs, which is evaluated as follows:

$$\zeta_i\left(\alpha_i, \boldsymbol{\alpha}_{-i}\right) = \sum_{l \neq i} R_l\left(\alpha_l, \boldsymbol{\alpha}_{-i}\right) - \sum_{l \neq i} R_l\left(\alpha_l, \boldsymbol{\alpha}\right). \tag{10}$$

The use of the VCG auction mechanism based on rate leads to the maximization of the overall performance of the system by exploiting cooperation between nodes. However, in modern wireless systems, UE pieces are more interested in fulfilling their quality of service (QoS) requirements than in maximizing their rate. Therefore, as an alternative, one can consider a satisfaction-based VCG auction mechanism, with satisfaction v_i defined as in (19) and (21), that can be formulated as

$$U_i\left(\alpha_i, \boldsymbol{\alpha}_{-i}\right) \triangleq f_i\left(\alpha_i, \boldsymbol{\alpha}_{-i}\right) - \psi_i\left(\alpha_i, \boldsymbol{\alpha}_{-i}\right), \tag{11}$$

where ψ_i denotes the satisfaction-based cost evaluated as follows:

$$\psi_i\left(\alpha_i, \boldsymbol{\alpha}_{-i}\right) = \sum_{l \neq i} f_l\left(\alpha_l, \boldsymbol{\alpha}_{-i}\right) - \sum_{l \neq i} f_l\left(\alpha_l, \boldsymbol{\alpha}\right). \tag{12}$$

Regret-Matching Learning. To achieve CE, a centralized approach can be applied, which is, however, very complex [30]. According to [39], the procedure of regret-matching learning can be used to iteratively achieve CE. In [29–31], a modified regret-matching learning algorithm is proposed to learn in a distributive fashion how to achieve the correlated equilibrium by solving the VCG auction, which aims at minimizing the regret of selecting a certain action. Regret $\mathrm{REG}^{(T)}$ of BS i at time T for playing action $\alpha^{(n)}$ instead of other actions is given as follows:

$$\mathrm{REG}_i^{(T)}\left(\alpha_i^{(n)}, \boldsymbol{\alpha}_i^{(-n)}\right) \triangleq \max\left\{D_i^{(T)}\left(\alpha_i^{(n)}, \boldsymbol{\alpha}_i^{(-n)}\right), 0\right\}, \tag{13}$$

where

$$D_i^{(T)}\left(\alpha_i^{(n)}, \boldsymbol{\alpha}_i^{(-n)}\right) = \max_{j \neq n} \frac{1}{T} \sum_{t \leq T}\left(U_i^t\left(\alpha_i^{(j)}, \boldsymbol{\alpha}_{-i}\right) - U_i^t\left(\alpha_i^{(n)}, \boldsymbol{\alpha}_{-i}\right)\right), \tag{14}$$

where $U_i^t(\alpha_i^{(\cdot)}, \boldsymbol{\alpha}_{-i})$ is the utility at time t. $D_i^T(\alpha_i^{(n)}, \boldsymbol{\alpha}_i^{(-n)})$ is the average payoff that BS i would have obtained if it had played another action compared with $\alpha_i^{(n)}$ every time in the past. Thus, the positive value of $D_i^T(\alpha_i^{(n)}, \boldsymbol{\alpha}_i^{(-n)})$ means that BS i would have obtained a higher average payoff when playing a different action than n. Finally, given the regrets for all N actions, the probability of BS i selecting strategy n can be formulated as follows:

$$\pi_i^{\left(\alpha_i^{(n)}\right)}(T) = 1 - \frac{1}{\mu^{(T-1)}} \mathrm{REG}_i^{(T-1)}\left(\alpha_i^{(n)}, \boldsymbol{\alpha}_i^{(-n)}\right), \tag{15}$$

where

$$\mu^{(T-1)} = \frac{\sum_n \mathrm{REG}_i^{(T-1)}\left(\alpha_i^{(n)}, \boldsymbol{\alpha}_i^{(-n)}\right)}{N-1}. \tag{16}$$

Strategy Identification. The correlated equilibrium achieved through the application of the regret-matching algorithms described above will hereafter be denoted as $\mathrm{CE}_{\mathrm{pure}}$.

3.2.3. Correlated Equilibrium with Reduced Complexity. One of the main burdens related to the practical application of the correlated equilibrium concept described in the previous subsection is the need for a fast exchange of detailed information about channel states for each mobile user or at least payoffs observed by each base station (BS) (player). The exchange of accurate data will result in a high traffic increase observed in the backhaul network, as will be discussed in detail in Section 4. Thus, let us now introduce the concept of achieving equilibrium with the regret-matching algorithm where the complexity is reduced. In this approach, the utility functions as well as the whole regret-matching algorithm and VCG auctions are kept unchanged, except for the fact that in each iteration only the payoff of the strategy selected by each BS is circulated in the network among other players instead of the whole table of payoffs.

Strategy Identification. In order to distinguish this solution from the application of the pure correlated equilibrium we denote this scenario as $\mathrm{CE}_{\mathrm{reduced}}$.

3.2.4. Generalized Nash Equilibrium: Satisfaction Equilibrium. The achievement of the satisfaction equilibrium [40], representing a specific case of the so-called generalized Nash equilibrium, is an example of the goal of a noncooperative game with assumed information exchange. The process of learning the satisfaction equilibrium (SE) can be described using the elements of the following game:

$$\mathbb{G} = \left(\{M \cup K\}, A, \{f_i\}_{i \in \{M \cup K\}}\right), \tag{17}$$

where $\{M \cup K\}$ represents a set of players (BSs), A denotes a set of available actions, and f_i is the satisfaction correspondence of player i, which indicates whether player is satisfied. The correspondence is defined as $f_i(\boldsymbol{\alpha}_{-i}) = \{\alpha_i \in A : U_i(\alpha_i, \boldsymbol{\alpha}_{-i}) \geq \Gamma_i\}$, with $U_i(\alpha_i, \boldsymbol{\alpha}_{-i})$ representing a player's observed utility when playing action α_i and Γ_i denoting the minimum utility level required by player i. A state of the game when all players satisfy their individual constraints simultaneously is referred to as the satisfaction equilibrium (SE), which is defined as follows [40].

Action profile α^+ is an equilibrium for the game

$$\mathbb{G} = \left(\{M \cup K\}, A, \{f_i\}_{i\in\{M\cup K\}}\right) \quad \text{if } \forall i \in \{M \cup K\},\ \alpha_i^+ \in f_k\left(\boldsymbol{\alpha}_{-i}^+\right). \tag{18}$$

Satisfaction Correspondence and Payoff Definition. The existence of SE depends mainly on the set of constraints imposed on the utility function, with the feasibility of the constraints as a necessary condition.

For the considered scenario, where BSs act as game players, the satisfaction correspondence can be defined in relation to the satisfaction level of all users served by the BS, as shown below:

$$f_i\left(\alpha_i, \boldsymbol{\alpha}_{-i}\right) = \frac{1}{|J_i|}\sum_{j\in J_i} s_{i,j}\left(\alpha_i, \boldsymbol{\alpha}_{-i}\right), \tag{19}$$

where $s_{i,j}(\alpha_i, \boldsymbol{\alpha}_{-i})$ is the satisfaction of UE j when BS selects action α_i. Individual UE satisfaction can be defined using the binary representation:

$$s_{i,j}\left(\alpha_i, \boldsymbol{\alpha}_{-i}\right) = \begin{cases} 1, & \text{if } T_j \geq T_{\min}, \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

Alternatively, one can consider a relaxed version of individual UE satisfaction using the sigmoid function:

$$s_{i,j}\left(\alpha_i, \boldsymbol{\alpha}_{-i}\right) = \begin{cases} 1, & \text{if } T_j \geq T_{\min}, \\ \dfrac{1}{1+\exp\left(\beta\cdot\left(T_j - (1-\beta)\cdot T_{\min}\right)\right)}, & \text{otherwise.} \end{cases} \tag{21}$$

Please note that compared to the case of the correlated equilibrium the payoff of each player is strictly defined by the satisfaction correspondence and does not refer explicitly to the rate, as it is the case with the correlated equilibrium.

Learning Satisfaction Equilibrium. We assume that the game players undertake actions in consecutive time intervals, with only one action selected per interval. At each interval, a player also observes whether it is satisfied or not. The selection of actions in each time interval is done based on the probability distribution:

$$\pi_i(t) = \left(\pi_i^{(\alpha_i^{(1)})}(t), \pi_i^{(\alpha_i^{(2)})}(t), \ldots, \pi_i^{(\alpha_i^{(N)})}(t)\right), \tag{22}$$

which is known as the probability distribution of exploration [40]. Under such assumptions, SE can be found using the behavioral rule, which states that the next action taken by player i is as follows:

$$\alpha_i(t) = \begin{cases} \alpha_i(t-1), & \text{if } f_i(t-1) = 1, \\ \alpha_i(t) \sim \pi_i(t), & \text{otherwise.} \end{cases} \tag{23}$$

The choice of probability distribution $\pi_i(t)$ may impact the convergence time and should also allow for the exploration of all actions (thus all actions should have nonzero probability). A simple choice may be to use uniform probability distribution $\pi_i^{(\alpha_i^{(k)})}(t) = 1/|A|$. On the other hand, more sophisticated probability distribution update methods may be used that increase the convergence speed, for example, based on the number of times an action has been selected previously [40].

The main problem with the learning solution presented above is that it neglects the utilities observed by players in the process of updating the probability distribution. An alternative approach has been proposed in [41], where the decentralized optimization is performed using the modified behavioral rule that accounts for observed utilities. This approach, known as the satisfaction equilibrium search algorithm (SESA), utilizes the knowledge of individual utilities to increase the probability of selecting actions that provide a higher payoff.

Strategy Identification. Solutions based on the use of the satisfaction equilibrium will be denoted hereafter as $\text{SE}_{\text{binary}}$ and $\text{SE}_{\text{sigmoid}}$, where the former describes a case where the satisfaction is represented by a binary function as in (20) and the latter identifies a case where the sigmoid function defined in (21) is used.

3.3. Energy Efficiency Factor in Game Definitions. So far, we have discussed strategies that guarantee rate maximization through appropriate resource and interference management. Let us note that the achievement of the satisfaction equilibrium can be interpreted as a game where energy efficiency is taken into account—once a player is satisfied, it will not be considered in the ongoing process of resource allocation for other players. It means that the wastage of redundantly assigned resources will be minimized, thus leading to better energy utilization in the system. However, the solutions presented in the previous section that utilize the correlated equilibrium have to be modified in order to lead to an overall energy efficiency improvement in the system. Thus, we propose including the cost of energy consumption in the game definition, and in particular we propose modifying the payoff functions.

Payoff Definition. In order to achieve better energy utilization by the base stations while keeping the average cell rate unchanged, we propose defining the payoff of the ith BS as follows:

$$U_i\left(\alpha_i, \boldsymbol{\alpha}_{-i}\right) \triangleq \frac{R_i\left(\alpha_i, \boldsymbol{\alpha}_{-i}\right)}{1 + p_i^{(T)}} - \widehat{\zeta}_i\left(\alpha_i, \boldsymbol{\alpha}_{-i}\right), \tag{24}$$

where $p_i^{(T)}$ stands for the total transmit power of the ith BS and $\widehat{\zeta}_i$ denotes the cost (rate loss) introduced by BS i to all other BSs, which is evaluated as follows:

$$\widehat{\zeta}_i\left(\alpha_i, \boldsymbol{\alpha}_{-i}\right) = \sum_{l\neq i}\frac{R_l\left(\alpha_l, \boldsymbol{\alpha}_{-i}\right)}{1 + P_l^{(T)}} - \sum_{l\neq i}\frac{R_l\left(\alpha_l, \boldsymbol{\alpha}\right)}{1 + P_l^{(T)}}. \tag{25}$$

Table 1: Comparison of strategies.

#	Strategy name	ID	Payoff definition	Information exchanged
1	Nash equilibrium	NE	Cell rate achieved by base station: see (6)	None
2	Correlated equilibrium	CE_{pure}	It is as for NE, but the cost of rate reduction observed by other players is considered; see (9)	Channel information or payoff table
3	CE reduced	$CE_{reduced}$	As for number 2, but reduced backhaul traffic, only the chosen payoff is updated	Payoff value for selected strategy
4	Satisfaction equilibrium binary	SE_{binary}	Wastage of redundantly assigned resources is minimized; satisfaction correspondence is calculated as (20)	None
5	Satisfaction equilibrium sigmoid	$SE_{sigmoid}$	It is as for 4, but satisfaction correspondence is calculated as (21)	Satisfaction table
6	Energy efficiency	$(\cdot)_E$	Selected strategies have been modified to include energy efficiency in the payoff definition as in (24) in Section 3.3	As in 2–5
7	Quantization	$(\cdot)_Q$	16-level quantization of the information exchanged among players is applied, Section 3.4	As in 2–5, but with reduced binary representation to 4 bits
8	Energy efficiency and quantization	$(\cdot)_{E,Q}$	Mixed strategies utilize energy efficiency increasing payoff definition and quantization, Section 3.5	As in 7

Similarly, to account for the energy utilization factor when considering the satisfaction equilibrium, (19) is modified as follows:

$$f_i(\alpha_i, \boldsymbol{\alpha}_{-i}) = \frac{(1/|J_i|)\sum_{j\in J_i} s_{i,j}(\alpha_i, \boldsymbol{\alpha}_{-i})}{1 + P_l^{(T)}}, \quad (26)$$

with the satisfaction of UE j $s_{i,j}(\alpha_i, \boldsymbol{\alpha}_{-i})$ calculated using (21).

Strategy Identification. The above concept has been applied to the following strategies: CE_{pure}, $CE_{reduced}$, and $SE_{sigmoid}$. In order to uniquely distinguish the new solutions, we use the subscript $(\cdot)_E$; that is, the following identifiers will be used: $CE_{pure,E}$, $CE_{reduced,E}$, and $SE_{sigmoid,E}$.

3.4. Backhaul Traffic Reduction. As has been discussed in Section 3.2.3, the application of each of the solutions described in this section requires the exchange of a relatively high amount of traffic. Although its influence on the energy consumed by the backhaul network will be discussed later, it is obvious that any reduction of the amount of control information is beneficial. Thus, following the solutions discussed in Section 3.3, we propose a further simplification of the algorithm by the application of the quantization procedure to the payoff values distributed among players. So far, we have assumed that each base station circulates either the full information about the channel between itself and all served users or the table of payoffs. In the strategy denoted as $CE_{reduced,E}$, only the values related to the selected strategy have to be delivered to other players. However, both the channel information and payoff value are double values, which have to be binary represented using, for example, 32 bits. In order to reduce such information overhead, we propose quantizing the information about the payoff value to four bits; that is, the index of one of only sixteen representative values of the payoff can be circulated. One can observe that by the application of such an approach the backhaul traffic will be reduced at least 8 times (if the 32-bit representation is used).

Strategies Identification. As previously, the idea of information quantization has been applied to the strategies: CE_{pure}, $CE_{reduced}$, and $SE_{sigmoid}$, which are now denoted as $CE_{pure,Q}$, $CE_{reduced,Q}$, and $SE_{sigmoid,Q}$.

3.5. Mixed Solution. Finally, we have jointly applied the concepts described in Sections 3.3 and 3.4. The strategies are denoted as $CE_{pure,E,Q}$, $CE_{reduced,E,Q}$, and $SE_{sigmoid,E,Q}$.

3.6. Strategy Comparison. In order to briefly compare the solutions described above, we gathered the concise information in Table 1.

4. Backhaul Traffic Analysis

One of the key problems in the practical realization of this approach is the amount of data that has to be circulated among active players—BSs (e.g., in the correlated equilibrium scenario). This parameter is strictly related to the observed delays in data delivery to the BS; thus it is important to assess the information burden added to the backhaul network. Moreover, it is easy to predict that even highly sophisticated solutions in terms of the guaranteed rate or throughput will not be practically deployed if they either will cause nonacceptable delays in the network or will require high infrastructure investments. In both cases, the application of such an algorithm will result in a direct or indirect cost increase that may not be acceptable for the MNO. As we deal with a highly realistic scenario of dense wireless networks in this paper, the key goal of this chapter is to discuss

the technical feasibility of the considered game-theoretic solutions.

Following the assumption presented in Section 2.2, let us remember that each player has a fiber connection with another in a star topology (i.e., one or two hops are required to deliver data between any two nodes) (fiber connection is our technology of choice, since solutions like fiber-to-the-antenna, FTTA, and/or radio-over-fiber, RoF, based techniques are often of the highest interest from the MNO point of view, e.g., [42–44]; an interested reader is encouraged to also follow the related work on the connection between wireless and optical parts of the communications network [45–47]). Although we select optical fibers as a way of ensuring base station connectivity, other solutions, such as Gigabit Ethernet or wireless backhauling, can also be considered. In the use-case discussed in this paper, we intentionally chose a fiber-based network as a quite mature technology that enables the achievement of high data rates in the network. Once the reference technology is selected, we need to estimate the traffic load due to the application of the proposed solutions. In our calculations, we will focus on the algorithm utilized in the correlated equilibrium case, since it is characterized by the highest needs for data exchange. However, let us again note that backhaul optimization is not part of any of the considered games and is used only for the purpose of comparing the energy consumption.

In the simplest case, every base station needs to exchange with other nodes the information about the channel characteristics to the served users $h_{i,j}^{(s)}$ and the probability distribution for each of the possible playing strategies. In the following, we fix the binary representation of each value exchanged between nodes to 32 bits. Such a matrix with channel information contains $N_{\mathrm{RB}} \cdot \widehat{J}$ entries, where N_{RB} stands for the total number of RBs, and $\widehat{J}$ for the average number of users served by one base station (i.e., $\widehat{J} = J/N$, where N is the total number of base stations). Assuming uniform user deployment in the considered area (i.e., approx. $\widehat{J} = 520/21 \approx 25$), the required number of bits to be transferred is equal to $T_1 = 32 \cdot \widehat{J} \cdot N_{\mathrm{RB}} \cdot N \cdot N = 32_{\mathrm{bits}} \cdot 25_{\mathrm{users}} \cdot 100_{\mathrm{RBs}} \cdot 21_{\mathrm{first\ hop}} \cdot 17 \cdot 21_{\mathrm{second\ hop}} = 440.64 \cdot 10^6$ bits per 10 TTIs, which corresponds to approx. 44 Gbps. Additionally, in order to circulate the selected strategy (one of 16 in our case), each base station needs to send 4 bits resulting in total traffic $T_2 = 4 \cdot N \cdot N = 1296$ bits per 10 TTIs. Such a great number of bits that would have to be exchanged will definitely disqualify such an algorithm from further considerations due to its nonpracticality. Hopefully, such a great burden can be strongly reduced since instead of channel information payoff matrices can be exchanged. In particular, in order to distribute the matrix of payoffs (of the size calculated as the number of strategies times the number of base stations) we need to send $T_3 = 32_{\mathrm{bits}} \cdot 16_{\mathrm{strategies}} \cdot 18_{\mathrm{first\ hop}} \cdot 18_{\mathrm{second\ hop}} = 10512$ bits per 10 TTIs. Thus, the total traffic increase observed in the backhaul network equals approx. $T_t = 1.1$ Mbps.

In order to assess the cost of energy consumption increase due to the higher traffic in the backhaul network, we have evaluated the typical values of power consumed by the contemporary equipment used by operators of fiber networks. A detailed analysis of this problem is presented in [14]. Due to the short distances between the nodes in the network (less than 2 km), there is no need for any in-line amplifiers. Thus, one has to account for the power consumed by the booster, the power amplifier, and, eventually, the optical cross connects with the regenerator deployed on the intermediate node (macro base station). Following [14], two models of energy consumption can be analyzed—one that relies on the measurements of contemporary devices available on the market and another that is based on an analytical model. In the former case, the change in traffic of 1.1 Mbps has, in fact, no measurable effect on the power consumed by the optical devices. It is due to the fact that the values of power consumption refer to particular classes of optical devices or simply do not depend on the traffic load (e.g., optical line amplifier, OLA, used for a short span of 2 km consumes a constant power of 65 W, while a transponder/muxponder for 2.5G traffic and for 10G traffic needs 25 W and 50 W, resp.). A change of the overall traffic by 1.1 Mbps will not result in a change of, for example, transponder class. In other words, following the first approach, backhaul power consumption will be kept unchanged. Thus, let us now discuss the problem of power consumption with the application of analytic models presented in formula (4) in [14]. One can observe that the exact power consumption depends on various parameters, such as power efficiency values (denoted as P/C), cooling and facilities overhead η_c, traffic protection η_{pr}, hop count H, demand capacity D_C, and the number of traffic demands N_d. An exemplary formula for power consumed by, for example, OLA is

$$P_{\mathrm{OLA}} = \eta_c \eta_{\mathrm{pr}} D_C N_D \frac{P_{\mathrm{OLA}}}{C_{\mathrm{OLA}}} H \frac{\alpha_L}{L_{\mathrm{OLA}}}, \tag{27}$$

where L_{OLA} is the optical amplification span length and α_L is the average (lightpath) link length. One can observe that for a given backhaul network topology, all of the components remain the same, except for the number of traffic demands and average demand capacity. Analogous conclusions can be drawn for a power model of any other backhaul network element. Thus, one needs to find the value of the following relation $P_{\mathrm{OLA}}^{(1)}/P_{\mathrm{OLA}}^{(2)} = D_C^{(1)} N_D^{(1)} / D_C^{(2)} N_D^{(2)}$, where the superscripts (1) and (2) represent the states of the system with and without the backhaul traffic generated by the algorithms discussed in this paper. However, based on the discussion and results (Figures 8–10) from [14], it can be concluded that the increase of the total consumed power due to a traffic increase of 1.1 Mbps is rather negligible.

The above discussion is valid for the most demanding solution, that is, the one that utilizes the concept of the correlated equilibrium with full information exchange. Since all other algorithms require less steering data to be exchanged in the backhaul network, it can be concluded that from the point of view of energy consumption by the backhaul network the application of any algorithm discussed in this paper has only a negligible effect. Clearly, such an analysis has to be repeated for certain technologies and solutions applied by MNO.

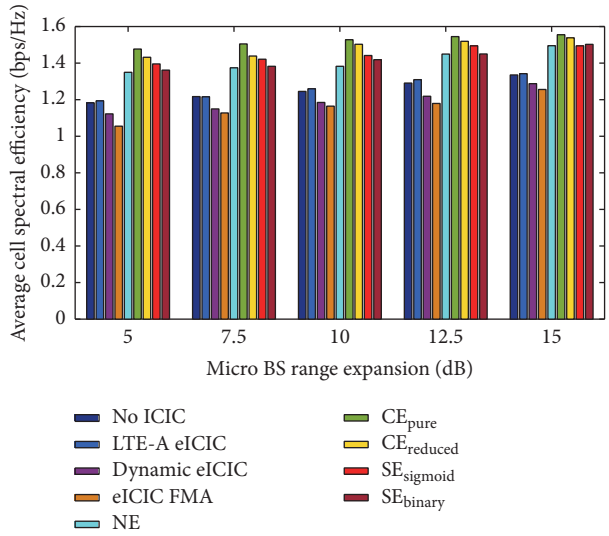


Figure 5: Average cell spectral efficiency for the considered baseline solutions.

5. Simulation Results and Analysis

To investigate the properties and validity of the considered game-based solutions, system-level Monte-Carlo simulations of the system described in Section 2 have been carried out. As a simulation parameter, the micro BS cell range expansion factor has been used, with five values considered {5, 7.5, 10, 12.5, 15} [dB]. As a reference, four configurations have been used:

(i) a plain system with no interference mitigation (denoted as no_ICIC),

(ii) a system utilizing LTE-A fixed eICIC with four ABS specified for macro BSs (hereafter denoted as LTE-A eICIC),

(iii) a system using a dynamic LTE-A eICIC mechanism proposed in [48] (denoted as dynamic eICIC),

(iv) a system using an adaptive mechanism of Fast Muting Adaptation with PF criterion employed, proposed in [49] (denoted as eICIC FMA).

Among the game-based solutions, three of them, namely, CE_{pure}, $CE_{reduced}$, and a proposed version of $SE_{sigmoid}$, have also been considered in the energy-efficient, backhaul-optimized, and mixed forms.

5.1. Baseline Solutions. In Figure 5, the average cell spectral efficiency (number of bits per second per unit bandwidth) for baseline solutions is presented. One can notice that the highest spectral efficiency is achieved in approaches that assume rich information exchange, with CE_{pure} outperforming other solutions. The gain observed for CE_{pure} versus the plain system or the system using LTE-A eICIC is over 20%, with the $CE_{reduced}$ performing only slightly worse (up to a 5% decrease compared to CE_{pure}). Therefore, when analyzing the spectral efficiency, one can state that the game-based approach using the correlated equilibrium is a very promising solution for interference mitigation. One can also notice that there is no improvement or even decrease in the spectral efficiency when eICIC methods are used, compared to the case with no interference mitigation. This indicates that the main victims of interference in the considered case are the macro UE pieces that are affected by micro BSs transmission. In such a situation, eICIC cannot improve the throughput of macro UE pieces, as it specifies almost-blank subframes (ABSFs) for macro BSs only. On the other hand, the proposed approaches using CE or SE optimize the use of resources in both macro and small BSs, thus improving the performance of macro UE pieces experiencing high interference.

The performance of all methods improves with the increase of the range expansion (RE) parameter, which corresponds to a higher number of UE pieces connected to small BSs. In the case of high RE, the UE pieces that would otherwise be assigned to a macro BS and experience high interference are connected to small BSs and benefit from the use of eICIC or other interference mitigation solutions. Moreover, small BSs usually provide services for a smaller number of UE pieces than macro BSs. Thus, offloading users to small BSs results in a higher number of UE pieces being scheduled for transmission.

The solutions based on the SE concept perform more poorly than CE because of the nature of the correspondence function. For UE pieces that achieve the required throughput, any further increase in data rate does not increase their satisfaction. Therefore, the system spectral efficiency is traded for improvement in general user satisfaction represented by achieving certain required throughput.

5.2. Energy-Efficient Solutions. The properties of the considered game-based solutions can also be used to improve the energy efficiency of the system. Therefore, energy-efficient versions of selected algorithms have also been considered in the investigation. Figures 6 and 7 present the comparison of baseline and energy-efficient versions in terms of the

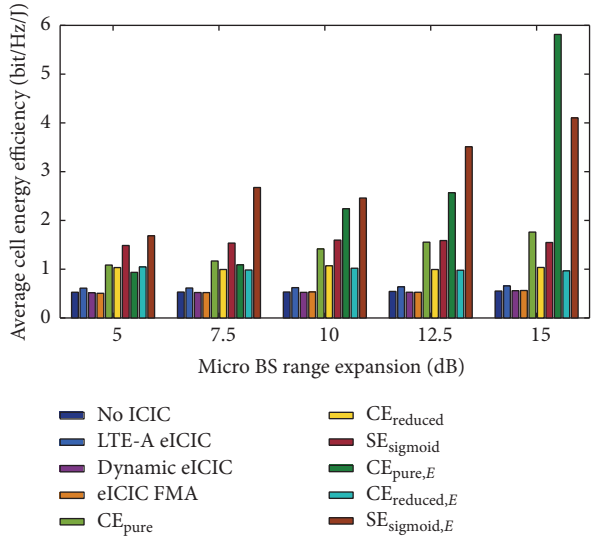


Figure 6: Average cell energy efficiency for the considered baseline and energy-efficient solutions.

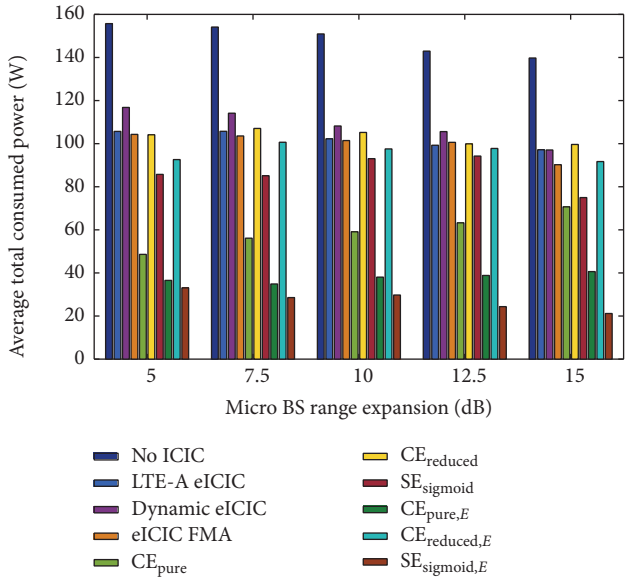


Figure 7: Average consumed power for the considered baseline and energy-efficient solutions.

achieved average cell energy efficiency and total power consumed by the system, respectively. One can notice that a huge gain in energy efficiency can be achieved when using energy-optimized versions of CE_{pure} and $SE_{sigmoid}$ for high RE values, where most of the traffic is offloaded to micro BSs. Moreover, it can be observed in Figure 7 that 3-4 times lower transmit power is used in the case of the energy-optimized solutions when compared to the plain system or the system using eICIC.

In most cases, the highest energy efficiency is achieved when using the $SE_{sigmoid}$ approach. The reason for such behavior is the nature of the SE correspondence function. For UE pieces that achieve the required throughput levels, a further increase in their utility can be achieved by decreasing the transmit power. Therefore, energy efficiency increases at the cost of a slight reduction of spectral efficiency, which is indicated in Figure 8.

A very interesting observation can be made about the behavior of the energy-optimized $CE_{reduced}$ solution. One can notice that due to the reduced exchange of information between the BSs the optimization mechanisms of energy usage cannot determine the gains from the use of other strategies; thus they mostly operate on the basis of throughput analysis. This results in almost identical performance of the energy-optimized approach as of the baseline $CE_{reduced}$.

The introduction of energy efficiency optimization does not degrade the gains of the considered approaches in terms of spectral efficiency. As shown in Figure 8, the spectral efficiency achieved with the energy-optimized solutions is almost the same as that for the baseline algorithms. The only

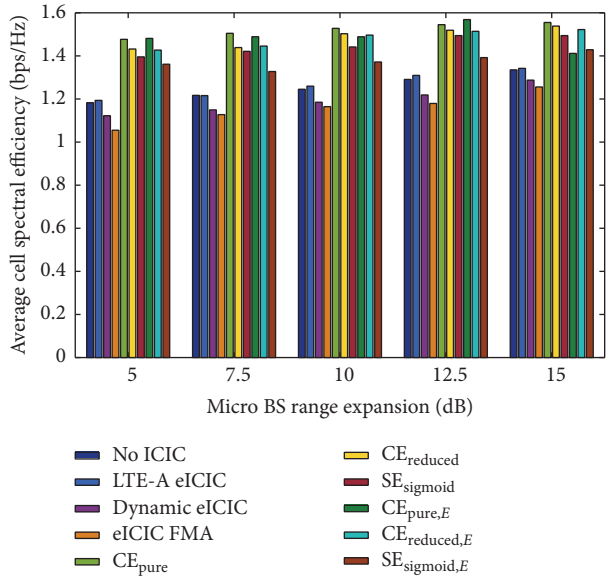


Figure 8: Average cell spectral efficiency for the considered baseline and energy-efficient solutions.

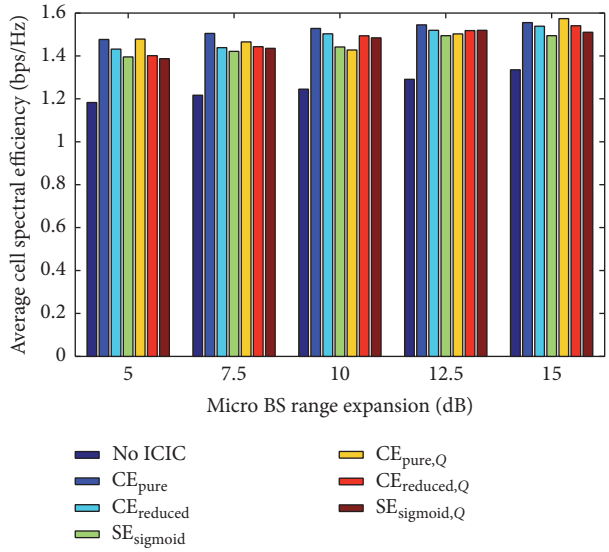


Figure 9: Average cell spectral efficiency for the considered baseline and backhaul-optimized solutions.

exception is the $CE_{pure,E}$ in the case of the RE factor equal to 15 dB, where the very high energy efficiency is achieved at the cost of reduced spectral efficiency, which is still higher than for the plain system.

5.3. Backhaul-Optimized Solutions. The algorithms based on CE or SE can yield high gains in terms of spectral and energy efficiency. However, control information needs to be exchanged between BSs. In order to reduce the burden on the backhaul network, an approach based on payoff quantization has been proposed, where four bits are used to represent each payoff value exchanged in the backhaul. Figure 9 presents the comparison of spectral efficiency achieved with baseline and backhaul-optimized solutions. One can notice that the introduction of quantization results in a minor performance degradation for several cases; however, the gains in terms of spectral efficiency are still significant when compared to the plain system.

An interesting observation is that there is hardly any loss in spectral efficiency when using $CE_{reduced,Q}$ or $SE_{sigmoid,Q}$ compared to $CE_{pure,Q}$. This indicates that quantization has a bigger impact when using solutions based on full exchange of information. When using the SE approach, the characteristics of the correspondence function reduces the cost of quantization. Similarly, for $CE_{reduced,Q}$, the limited exchange of information, and thus slower convergence, mitigates the impact of quantization errors.

Similarly, the mixed approach has been evaluated, where both energy-efficient and backhaul-optimized approaches are applied simultaneously. Figures 10 and 11 present the

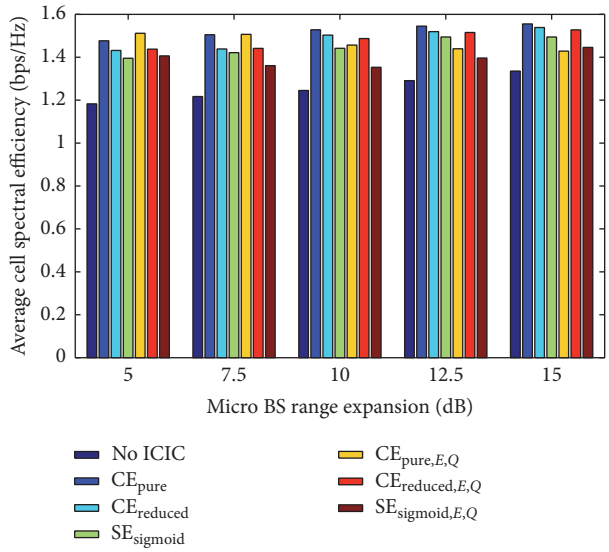


Figure 10: Average cell spectral efficiency for the considered baseline and mixed solutions.

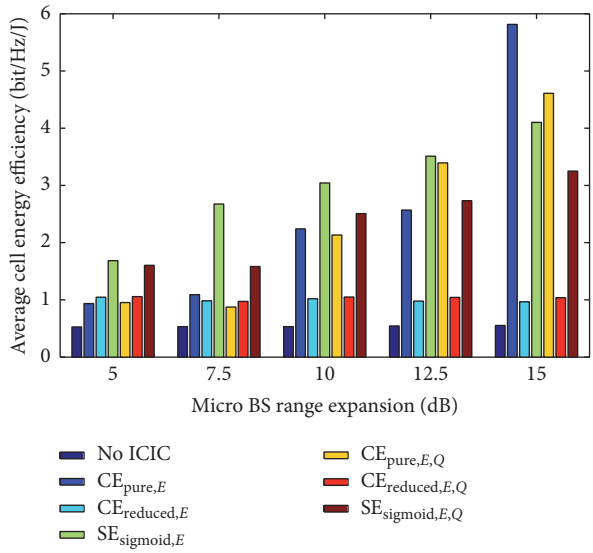


Figure 11: Average cell energy efficiency for the considered baseline and mixed solutions.

achieved average cell spectral efficiency and average cell energy efficiency, respectively. One can notice that the use of the mixed approach does not significantly degrade the performance of the considered algorithms with reference to the baseline solutions. The conclusions are the same as those for the energy- and backhaul-optimized solutions. Thus, both improvements are promising and perfectly applicable approaches that provide the means of practical implementation of the considered game-theoretic schemes.

5.4. Summary. The proposed game-based solutions have been evaluated in terms of spectral efficiency, energy efficiency, and total consumed power. The simulation results clearly indicate that the most promising solutions are the CE_{pure} and $SE_{sigmoid}$ algorithms, with both providing significant increase in terms of spectral and energy efficiency. In practical systems, SE might be a more suitable solution, as usually UE pieces use services that require some minimal or aggregate data rate that can be represented using the satisfaction correspondence function. By using different satisfaction definitions for different services, one can easily distinguish the different payoffs of many UE pieces based on the services they use.

One can notice the significant improvement of CE and SE compared to the eICIC methods, as these are not suitable for scenarios with small- to macro-layer interference. By treating the micro and macro BSs equally in the case of CE and SE, we can improve the performance of UE pieces connected to the macro BS that are victims of strong interference. Moreover, the use of the dynamic eICIC approach does not bring any improvement, as the considered scenario is a low mobility one.

The proposed energy-efficient modification further increases the energy savings of both the CE and SE solutions. High energy efficiency has been observed especially in the case of the SE approach. In the case of CE, a high reduction of energy consumption is achieved only in the case of a high RE parameter, where most of the UE pieces are connected to small BSs.

From the practical point of view, the most suitable solution would be the $CE_{reduced}$ one, as it requires the exchange of the smallest amount of information in the backhaul network. However, one can notice that it benefits less from the energy-optimized approach than other solutions. Furthermore, the practical application of the considered approaches with full information exchange is possible and cost-effective thanks to the robustness of the considered algorithms against the inaccuracy of payoff values caused by quantization. Only a minor reduction in the achieved spectral efficiency has been observed for $CE_{pure,Q}$ and $SE_{sigmoid,Q}$ when compared to their baseline solutions.

Finally, one can state that the use of different power levels in selected subbands by the BSs provides an increase in the spectral efficiency of the system. This increase is independent of the type of users that are affected by the highest interference. For the case when both macro UE pieces or small cell UE pieces are the interference victims, an improvement in system performance can be observed when using the considered game-theoretic solutions based on CE and SE. This is in contrast to the eICIC solutions based on the use of ABSF, as these aim at the reduction of interference from the macro to small cell layer in the downlink.

6. Conclusion

The goal of this work was to propose an efficient and flexible tool for radio resource management in the context of its practical implementation in future wireless networks. Based on the presented results of simulations carried out for a highly accurate model of a dense urban network, it can be concluded that the application of the proposed game-theoretic algorithms guarantees the achievement of high cell throughput, while at the same time minimizing the energy consumed by the base station. All algorithms discussed in this paper concentrate on the optimization of the overall energy consumption and such strategies are preferred that minimize the consumed energy while ensuring high data rates observed by all users. All of the algorithms have been compared with the solution known from the LTE-A standards, eICIC, and have proved their effectiveness. Moreover, based on the detailed discussion of the traffic increase in the backhaul network it can be concluded that the cost of practical implementation of the proposed solutions for RRM will be rather negligible when considering the backhaul requirements.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The work by Paweł Sroka has been partly performed in the framework of the FP7 project ICT-317669 METIS, which was partly funded by the European Union. His work has been also supported by the Polish Ministry of Science and Higher Education funds for the status activity Project DSPB/08/81/0161. The work by Adrian Kliks has been supported by the EU H2020 Project COHERENT under Contract 671639.

References

[1] I. F. Akyildiz, D. M. Gutierrez-Estevez, R. Balakrishnan, and E. Chavarria-Reyes, "LTE-advanced and the evolution to beyond 4G (B4G) systems," *Physical Communication*, vol. 10, pp. 31–60, 2014.

[2] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green cellular networks: a survey, some research issues and challenges," *IEEE Communications Surveys and Tutorials*, vol. 13, no. 4, pp. 524–540, 2011.

[3] M. A. Marsan and M. Meo, "Energy efficient wireless internet access with cooperative cellular networks," *Computer Networks*, vol. 55, no. 2, pp. 386–398, 2011.

[4] L. M. Correia, D. Zeller, O. Blume et al., "Challenges and enabling technologies for energy aware mobile radio networks," *IEEE Communications Magazine*, vol. 48, no. 11, pp. 66–72, 2010.

[5] E. Hossain, V. K. Bhargava, and G. P. Fettweis, Eds., *Green Radio Communication Networks*, Cambridge University Press, Cambridge, Uk, 2012.

[6] K. S. Ko, B. C. Jung, and D. K. Sung, "On the energy efficiency of wireless random access networks with multi-packet reception," in *Proceedings of the 2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC '13)*, pp. 1666–1670, London, UK, September 2013.

[7] Q.-D. Vu, L.-N. Tran, M. Juntti, and E.-K. Hong, "Energy-efficient bandwidth and power allocation for multi-homing networks," *IEEE Transactions on Signal Processing*, vol. 63, no. 7, pp. 1684–1699, 2015.

[8] X. Ge, H. Cheng, M. Guizani, and T. Han, "5G wireless backhaul networks: challenges and research advances," *IEEE Network*, vol. 28, no. 6, pp. 6–11, 2014.

[9] S. Tombaz, K. W. Sung, and J. Zander, "On metrics and models for energy-efficient design of wireless access networks," *IEEE Wireless Communications Letters*, vol. 3, no. 6, pp. 649–652, 2014.

[10] J. Baliga, R. W. A. Ayre, K. Hinton, and R. S. Tucker, "Energy consumption in wired and wireless access networks," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 70–77, 2011.

[11] Y. Zhang, "Performance modeling of energy management mechanism in IEEE 802.16e mobile WiMAX," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '07)*, pp. 3205–3209, IEEE, Hong Kong, March 2007.

[12] T. Adame, A. Bel, B. Bellalta, J. Barcelo, and M. Oliver, "IEEE 802.11AH: the WiFi approach for M2M communications," *IEEE Wireless Communications*, vol. 21, no. 6, pp. 144–152, 2014.

[13] N. Benzaoui, Y. Pointurier, T. Bonald, J.-C. Antona, Q. Wei, and M. Lott, "Performance analysis in a next generation optical mobile backhaul network," in *Proceedings of the 16th International Conference on Transparent Optical Networks (ICTON '14)*, pp. 1–4, July 2014.

[14] W. Van Heddeghem, F. Idzikowski, W. Vereecken, D. Colle, M. Pickavet, and P. Demeester, "Power consumption modeling in optical multilayer networks," *Photonic Network Communications*, vol. 24, no. 2, pp. 86–102, 2012.

[15] U. Bauknecht and F. Feller, "Dynamic resource operation and power model for IP-over-WSON networks," in *Advances in Communication Networking: 19th EUNICE/IFIP WG 6.6 International Workshop, Chemnitz, Germany, August 28–30, 2013. Proceedings*, vol. 8115 of *Lecture Notes in Computer Science*, pp. 1–12, Springer, Berlin, Germany, 2013.

[16] H. Li, "Multi-agent Q-learning of channel selection in multi-user cognitive radio systems: a two by two case," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '09)*, pp. 1893–1898, IEEE, San Antonio, Tex, USA, October 2009.

[17] C. Han and S. Armour, "Energy efficient radio resource management strategies for green radio," *IET Communications*, vol. 5, no. 18, pp. 2629–2639, 2011.

[18] A. Kliks, N. Dimitriou, A. Zalonis, and O. Holland, "WiFi traffic offloading for energy saving," in *Proceedings of the 2013 20th International Conference on Telecommunications (ICT '13)*, pp. 1–5, Casablanca, Morocco, May 2013.

[19] J. Christoffersson, "Energy efficiency by cell reconfiguration: MIMO to non-MIMO and 3-cell sites to omni," in *Proceedings of the IEEE 73rd Vehicular Technology Conference (VTC '11)*, pp. 1–5, IEEE, Yokohama, Japan, May 2011.

[20] P. Kolios, V. Friderikosy, and K. Papadaki, "Switching off low utilization base stations via store carry and forward relaying," in *Proceedings of the IEEE 21st International Symposium on Personal, Indoor and Mobile Radio Communications Workshops (PIMRC '10)*, pp. 312–316, September 2010.

[21] N. Sapountzis, S. Sarantidis, T. Spyropoulos, N. Nikaein, and U. Salim, "Reducing the energy consumption of small cell networks subject to QoE constraints," in *Proceedings of the 2014 IEEE Global Communications Conference (GLOBECOM '14)*, pp. 2485–2491, Austin, Tex, USA, December 2014.

[22] Q. Li, M. Xu, Y. Yang, L. Gao, Y. Cui, and J. Wu, "Safe and practical energy-efficient detour routing in IP networks," *IEEE/ACM Transactions on Networking*, vol. 22, no. 6, pp. 1925–1937, 2014.

[23] J. Zuo, C. Dong, H. V. Nguyen, S. X. Ng, L.-L. Yang, and L. Hanzo, "Cross-layer aided energy-efficient opportunistic routing in ad hoc networks," *IEEE Transactions on Communications*, vol. 62, no. 2, pp. 522–535, 2014.

[24] METIS2020, Mobile and wireless communications Enablers for the Twenty-twenty Information Society, 2015, https://www.metis2020.com/.

[25] D. López-Pérez, I. Güvenç, G. de la Roche, M. Kountouris, T. Q. S. Quek, and J. Zhang, "Enhanced inter-cell interference coordination challenges in heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 22–30, 2011.

[26] R1-104256, *eICIC Solutions Details*, 3GPP Standard, Dresden, Germany, 2010.

[27] R4-10493, Way Forward on Candidate TDM Patterns for Evaluation of eICIC Intra-Frequency Requirements. 3GPP Standard, Jacksonville, Fla, USA, 2010.

[28] 3GPP TS 36.300, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overal description, v. 11.3.0," 2012.

[29] M. Charafeddine, Z. Han, A. Paulraj, and J. Cioffi, "Crystallized rates region of the interference channel via correlated equilibrium with interference as noise," in *Proceedings of the IEEE International Conference on Communications (ICC '09)*, pp. 1–6, IEEE, Dresden, Germany, June 2009.

[30] A. Kliks, P. Sroka, and M. Debbah, "Crystallized rate regions for MIMO transmission," *Eurasip Journal on Wireless Communications and Networking*, vol. 2010, Article ID 919072, 17 pages, 2010.

[31] P. Sroka and A. Kliks, "Distributed interference mitigation in two-tier wireless networks using correlated equilibrium and regret-matching learning," in *Proceedings of the European Conference on Networks and Communications (EuCNC '14)*, June 2014.

[32] N. AbuAli, M. Hayajneh, and H. Hassanein, "Congestion-based pricing resource management in broadband wireless networks," *IEEE Transactions on Wireless Communications*, vol. 9, no. 8, pp. 2600–2610, 2010.

[33] J.-H. Yun and K. G. Shin, "Adaptive interference management of OFDMA femtocells for co-channel deployment," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 6, pp. 1225–1241, 2011.

[34] M. Haddad, S. E. Elayoubi, E. Altman, and Z. Altman, "A hybrid approach for radio resource management in heterogeneous cognitive networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 831–842, 2011.

[35] J. F. Nash, "Equilibrium points in n-person games," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 36, no. 1, pp. 48–49, 1950.

[36] E. G. Larsson, E. A. Jorswieck, J. Lindblom, and R. Mochaourab, "Game theory and the flat-fading gaussian interference channel: analyzing resource conflicts in wireless networks," *IEEE Signal Processing Magazine*, vol. 26, no. 5, pp. 18–27, 2009.

[37] S. Lasaulce and H. Tembine, *Game Theory and Learning for Wireless Networks: Fundamentals and Applications*, Academic Press, Cambridge, Mass, USA, 2011.

[38] R. J. Aumann, "Subjectivity and correlation in randomized strategies," *Journal of Mathematical Economics*, vol. 1, no. 1, pp. 67–96, 1974.

[39] S. Hart and A. Mas-Colell, "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, vol. 68, no. 5, pp. 1127–1150, 2000.

[40] S. M. Perlaza, H. Tembine, S. Lasaulce, and M. Debbah, "Quality-of-service provisioning in decentralized networks: a satisfaction equilibrium approach," *IEEE Journal on Selected Topics in Signal Processing*, vol. 6, no. 2, pp. 104–116, 2012.

[41] S. M. Perlaza, H. Tembine, S. Lasaulce, and M. Debbah, "Satisfaction equilibrium: a general framework for QoS provisioning in self-configuring networks," in *Proceedings of the IEEE Global Communications Conference Configuring Networks (GLOBECOM '10)*, Miami, Fla, USA, December 2010.

[42] K. M. Maamoun and H. T. Mouftah, "A survey and a novel scheme for RoF-PON as FTTx wireless services," in *Proceedings of the 6th International Symposium on High Capacity Optical Networks and Enabling Technologies (HONET '09)*, pp. 246–253, IEEE, Alexandria, Egypt, December 2009.

[43] J. Beas, G. Castanon, I. Aldaya, A. Aragon-Zavala, and G. Campuzano, "Millimeter-wave frequency radio over fiber systems: a survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 1593–1619, 2013.

[44] V. A. Thomas, M. El-Hajjar, and L. Hanzo, "Performance improvement and cost reduction techniques for radio over fiber

communications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 627–670, 2015.

[45] N. Ghazisaidi, M. Maier, and C. M. Assi, "Fiber-wireless (FiWi) access networks: a survey," *IEEE Communications Magazine*, vol. 47, no. 2, pp. 160–167, 2009.

[46] Y. Liu, L. Guo, B. Gong et al., "Green survivability in Fiber-Wireless (FiWi) broadband access network," *Optical Fiber Technology*, vol. 18, no. 2, pp. 68–80, 2012.

[47] B. Kantarci, N. Naas, and H. T. Mouftah, "Energy-efficient DBA and QoS in FiWi networks constrained to metro-access convergence," in *Proceedings of the 14th International Conference on Transparent Optical Networks (ICTON '12)*, pp. 1–4, Coventry, UK, July 2012.

[48] S. Vasudevan, R. N. Pupala, and K. Sivanesan, "Dynamic eICIC—a proactive strategy for improving spectral efficiencies of heterogeneous LTE cellular networks by leveraging user mobility and traffic dynamics," *IEEE Transactions on Wireless Communications*, vol. 12, no. 10, pp. 4956–4969, 2013.

[49] B. Soret and K. I. Pedersen, "Centralized and distributed solutions for fast muting adaptation in LTE-Advanced HetNets," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 1, pp. 147–158, 2015.

Hindawi Publishing Corporation
Mobile Information Systems
Volume 2016, Article ID 1542540, 11 pages
http://dx.doi.org/10.1155/2016/1542540

Research Article

Latent Clustering Models for Outlier Identification in Telecom Data

Ye Ouyang,[1] Alexis Huet,[2] J. P. Shim,[3] and Mantian (Mandy) Hu[4]

[1]*Columbia University, New York, NY, USA*
[2]*Nanjing Howso Technology, Nanjing, China*
[3]*Georgia State University, Atlanta, GA, USA*
[4]*Department of Marketing, The Chinese University of Hong Kong, Shatin, Hong Kong*

Correspondence should be addressed to Alexis Huet; alexis@howso.cn

Received 29 July 2016; Revised 3 November 2016; Accepted 17 November 2016

Academic Editor: Mariusz Głąbowski

Copyright © 2016 Ye Ouyang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Collected telecom data traffic has boomed in recent years, due to the development of 4G mobile devices and other similar high-speed machines. The ability to quickly identify unexpected traffic data in this stream is critical for mobile carriers, as it can be caused by either fraudulent intrusion or technical problems. Clustering models can help to identify issues by showing patterns in network data, which can quickly catch anomalies and highlight previously unseen outliers. In this article, we develop and compare clustering models for telecom data, focusing on those that include time-stamp information management. Two main models are introduced, solved in detail, and analyzed: Gaussian Probabilistic Latent Semantic Analysis (GPLSA) and time-dependent Gaussian Mixture Models (time-GMM). These models are then compared with other different clustering models, such as Gaussian model and GMM (which do not contain time-stamp information). We perform computation on both sample and telecom traffic data to show that the efficiency and robustness of GPLSA make it the superior method to detect outliers and provide results automatically with low tuning parameters or expertise requirement.

1. Introduction

High-speed telecom connections have developed rapidly in recent years, which has resulted in a major increase in data flow through networks. Beyond the issues of storage and management of this flow of data, a major challenge is how to select and use this mass of material to better understand a network. The detection of behaviors that differ from normal traffic patterns is a critical element, since such discrepancies can reduce network efficiency or harm network infrastructures. And because those anomalies can be caused by either a technical equipment problem or a fraudulent intrusion in the network, it is important to identify them accurately and fix them promptly. Data-driven systems have been developed to detect anomalies using machine learning algorithms and can automatically extract information from raw data to promptly alert a network manager when an anomaly occurs.

The data collected in telecom networks contains values for different features (related to network resource and usage) as well as time stamps, and those values can be modeled and processed to seek and detect anomalies using unsupervised algorithms. The algorithms use unlabeled data and assume that information about which data elements are anomalies is unknown (since anomalies in traffic data are rare and may take many forms). They do not directly detect anomalies but instead separate and distinguish data structures and patterns in order to group data from which "zones of anomalies" are deduced. The main advantage of this methodology is the ability to quickly detect previously unseen or unexpected anomalies.

Another component to be taken into consideration for understanding wireless network data behavior is time stamps. This information is commonly collected when data are generated but is not widely used in classic anomaly detection processes. However, since network load fluctuates over the course of a day, adding time-stamp attributes in an evaluation model can allow us to discover periodic behaviors. For example, a normal value during a peak period may be an

anomaly outside that period and thus remain undetected by an algorithm that does not take time stamps into account.

In this article, we use unsupervised models to detect anomalies. Specifically, we focus on algorithms combining both values and dates (time stamps) and introduce two new models to this end. The first one is the time-dependent Gaussian Mixture Model (time-GMM), which is a time-dependent extension of GMM [1] which works by considering each period of time independently. The second one is Gaussian Probabilistic Latent Semantic Analysis (GPLSA), derived from Probabilistic Latent Semantic Analysis (PLSA) [2], which combines values and dates processing together in a unique machine learning algorithm. This latter algorithm is well known in text-mining and recommender systems areas but has been rarely used in other domains such as anomaly detection. In this research, we fully implement these two algorithms with R [3] and test their ability to find anomalies and to adapt to new patterns on both sample and traffic data. We also compare the robustness, complexity, and efficiency of these algorithms.

The rest of the article is organized as follows: in Section 2, we present an overview of techniques to identify anomalies, with an emphasis on unsupervised models. In Section 3, we show different unsupervised anomaly detection models (this section defines two previously introduced unsupervised models: GPLSA and time-GMM). In Section 4, those models are compared to a sample set to highlight the differences of behavior in a simple context. In Section 5, we discuss computations performed on real traffic network data. We finally, in Section 6, draw conclusions about adaptability and robustness of GPLSA.

2. Research Background

Anomaly detection is a broad topic with a large number of previously used techniques. For a broad overview of those methods, we refer to [4].

Previous research focuses mainly on unsupervised statistical based methods such as clustering methods to perform anomaly detection [5–8]. A common assumption for statistical based methods is that the underlying distribution is Gaussian [9], although mixtures of parametric distributions, where normal-points anomalies correspond to two different distributions [10], are also possible. In clustering methods, the purpose is to separate data points and to group objects together that share similarities, and each group of objects is called a cluster. We usually define similarities between objects analytically. Many clustering algorithms that differ on how similarities between objects are measured (using distance measurement, density, or statistical distribution) exist but the most popular and simplest clustering technique is K-means clustering [11].

Advanced methods of detection combine statistical hypotheses and clustering, as seen in the Gaussian Mixture Model (GMM) [1]. This method assumes that all data points are generated from a mixture of K Gaussian distributions; parameters are usually estimated through an Expectation-Maximization (EM) algorithm, where the aim is to iteratively increase likelihood of the set [12]. Some studies have used GMM for anomaly detection problems, as described in [13–15]. Selecting the number of clusters K is not easy: Although methods to automatically select a value of K do exist (a comparison between different algorithms is presented in [16]), the selection is usually chosen manually by researchers and refined after performing different computations for different values.

In telecom traffic data, time stamps are a component to be considered when seeking for traffic anomalies. This information, referred to as contextual attributes in [4], can dramatically change the results of anomaly detection. For example, a value can be considered normal in a certain context (in a peak period) but abnormal in another context (in off-peak periods), and the differentiation can only be made clear when each value has a time stamp associated with it. An overview of outlier detection for temporal data can be found in [17], which comprises ensemble methods (e.g., [18, 19]), time-series models (e.g., with ARIMA or GARCH models in [20]), and correlation analysis [21, 22].

Clustering methods for temporal anomaly detection can automatically take into account and separate different types of behavior from raw time-series data, which allows for some interesting results. One way to incorporate time stamps is to consider the original GMM (i.e., a mixture of K Gaussian distributions), but to weigh each distribution differently, depending on time. This method was first introduced for text-mining [2, 23] with a mixture of categorical distributions and named Probabilistic Latent Semantic Analysis (PLSA). Its actual form (with Gaussian distribution), GPLSA, is used for recommendation systems [24]. No published article that applies GPLSA for anomaly detection has been found.

In the next section, we present five anomaly detection models for traffic data. The first three models are classic models: Gaussian model, time-dependent Gaussian, and GMM, which do not combine clustering and contextual detection and are expected to have several disadvantages. The two remaining models take clustering and time stamps into consideration: the fourth model is a time-dependent GMM, where a GMM is independently determined for each time slot; the fifth model is Gaussian Probabilistic Latent Semantic Analysis (GPLSA) model, which is solved by optimizing all parameters related to clusters and time in a unique algorithm.

3. Presentation of Models

In this section, five different models are defined: Gaussian, time-dependent Gaussian, GMM, time-dependent GMM, and GPLSA. We use the same following notations for all:

(i) W is a traffic data set. This set contains N values indexed with i. N is usually large, that is, from one thousand to one hundred million. Each value is a vector of $\mathbf{R}^p$, where p is the number of features. Furthermore, each feature is assumed to be continuous.

(ii) D is the time-stamp set of classes. This set also contains N values. Since we are expecting a daily cycle, each value d_i corresponds to each hour of the day, consequently standing in $\{1, \ldots, 24\}$.

(iii) $X = (W, D)$ are observed data.

Table 1: Anomaly detection methods compared.

	No date	Date
No clustering	Gaussian	Time-Gaussian
Clustering	GMM	(i) Time-GMM (ii) GPLSA

(iv) For clustering methods, we assume that each value is related to a fixed (although unknown) cluster, named Z. It is a "latent" set, since it is initially unknown. We assume that number of clusters K is known.

An example of traffic data retrieved is shown as follows:

date	Feat. *1*	$\cdots$	Feat. p	W	D
04/13 0:00	1069	$\cdots$	2.4	$(1069, \ldots, 2.4)$	1
04/13 0:30	1004	$\cdots$	2.3	$(1004, \ldots, 2.3)$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
05/04 23:30	997	$\cdots$	2.7	$(997, \ldots, 2.7)$	24.

(1)

For each model, the aim is to estimate parameters with maximum likelihood. When the direct calculation is intractable, an EM algorithm is used to find a local optimum (at least) of the likelihood. A usual hypothesis of independence is added, which is needed to compute the likelihood of the product over the set:

(H) The set of triplets $(W_i, Z_i, D_i)_i$ is an independent vector over the rows i. Note that if the model does not consider D or Z, we remove this set in the hypothesis.

The different models are shown in Table 1, grouped according to their ability to consider time stamps and clustering. In the following, for each model, each hypothesis is listed on the form (X2), where X is current model paragraph followed by the hypothesis number.

3.1. Gaussian Model. In the Gaussian model, the whole data set is assumed to come from a variable that follows a Gaussian distribution. Consequently, each part of the day has a similar behavior and there are no clusters. Mathematically (note that same letter is used for set and variable) the following occurs:

(A1) Each variable W_i follows Gaussian distribution with mean and variance m, Σ. Here, m is a p-vector and Σ is a variance-covariance matrix of size p. They are both independent of i.

Parameters are easily estimated with empirical mean and variance.

3.2. Time-Dependent Gaussian Model. A time component is added to this model, as opposed to the Gaussian model, which does not include a time component. Each time of the day is considered independently, following a particular Gaussian distribution. This allows us to take dependence of time into account:

(B1) For each $s \in \{1, \ldots, 24\}$, each conditional variable W_i such that $D_i = s$ follows a Gaussian distribution with mean and variance m^s, Σ^s.

As for the Gaussian model, parameters are estimated with empirical mean and variance for each class of dates.

3.3. Gaussian Mixture Model. Compared to the Gaussian model, in the GMM, data is assumed to come from a mixture of Gaussian distributions rather than one single Gaussian distribution. The number of clusters K is fixed in advance.

(C1) Each record belongs to a cluster $Z_i = k \in \{1, \ldots, K\}$ with probability α_k.

(C2) Each variable $(W_i \mid Z_i = k)$ follows a Gaussian distribution with mean and variance m_k, Σ_k.

Therefore, each record belongs to an unknown cluster. The task is to estimate both probability for each cluster and the parameters of each Gaussian distribution. To solve this problem, the following decomposition is done:

$$P(W_i) = \sum_k P(W_i \mid Z_i = k) P(Z_i = k). \quad (2)$$

The parameters can be successively updated with an EM algorithm (see [23] for details).

3.4. Time-Dependent Gaussian Mixture Model. Combining the models described in Sections 3.2 and 3.3, we obtain the time-dependent GMM, which includes both clustering and time-dependence. As in Section 3.3, the EM algorithm is used to estimate parameters.

(D1) For each $s \in \{1, \ldots, 24\}$, each record such that $D_i = s$ belongs to a cluster $Z_i = k \in \{1, \ldots, K\}$ with probability $\alpha_{k,s}$.

(D2) For each $s \in \{1, \ldots, 24\}$, each variable $(W_i \mid Z_i = k)$ such that $D_i = s$ follows a Gaussian distribution with mean and variance m_k^s, Σ_k^s.

3.5. Gaussian Probabilistic Latent Semantic Analysis Model. The GPLSA model is based on the classic GMM but introduces a novel link between data values and time stamps. In time-GMM, the different classes of dates are considered independently, whereas GPLSA introduces dependence between latent clusters and time stamps but only within those two variables. That is, in knowing latent cluster Z, we assume there is no more dependence on time. This assumption allows making the problem computationally tractable. Explicitly, the following occurs:

(E1) For each $s \in \{1, \ldots, 24\}$, each record such that $D_i = s$ belongs to a cluster $Z_i = k \in \{1, \ldots, K\}$ with probability $\alpha_{k,s}$.

(E2) Each variable $(W_i \mid Z_i = k)$ follows a Gaussian distribution with mean and variance m_k, Σ_k.

(E3) For all i, $P(W_i \mid D_i, Z_i) = P(W_i \mid Z_i)$.

To solve this problem, the following decomposition is done (the assumption (E3) is used for the first factor of the sum):

$$P\left(W_i \mid D_i = s\right) = \sum_k P\left(W_i \mid Z_i = k\right) P\left(Z_i = k \mid D_i = s\right). \tag{3}$$

The EM algorithm can be adapted in this case to iteratively increase the likelihood and estimate parameters in order to obtain exact update formulas. The complete calculus to derive these formulas is given in the Appendix. We let $f(\cdot \mid m, \Sigma)$ equal the density of a Gaussian with parameters m and Σ. Also, we define E_s as the set of indexes i, where $d_i = s$. The following algorithm describes the steps to get final parameters:

Step 1. At time $t = 1$, let some initial parameters $m_k^{(t-1)}$, $\Sigma_k^{(t-1)}$, and $\alpha_{k,s}^{(t-1)}$ for all k, s.

Step 2. For all k, i, compute the probability $Z_i = k$ knowing $W_i = w_i$, $D_i = d_i$, and parameters

$$T_{k,i}^{(t)} := \frac{f\left(w_i \mid m_k^{(t-1)}, \Sigma_k^{(t-1)}\right) \alpha_{k,d_i}^{(t-1)}}{\sum_{l=1}^{K} f\left(w_i \mid m_l^{(t-1)}, \Sigma_l^{(t-1)}\right) \alpha_{l,d_i}^{(t-1)}}. \tag{4}$$

Step 3. For all k, s, compute (here $\#E_s$ stands for the length of E_s)

$$S_{k,s}^{(t)} = \sum_{j=1}^{\#E_s} T_{k,E_s(j)}^{(t)}. \tag{5}$$

Step 4. For all k, s, update $\alpha_{k,s}$ with

$$\alpha_{k,s}^{(t)} = \frac{S_{k,s}^{(t)}}{\sum_{l=1}^{K} S_{l,s}^{(t)}}. \tag{6}$$

Step 5. For all k, update the means with

$$m_k^{(t)} = \frac{\sum_{i=1}^{N} w_i T_{k,i}^{(t)}}{\sum_{i=1}^{N} T_{k,i}^{(t)}}. \tag{7}$$

Step 6. For all k, update the covariance matrix with (here $'$ refers to the transpose)

$$\Sigma_k^{(t)} = \frac{\sum_{i=1}^{N} \left(w_i - m_k\right)' \left(w_i - m_k\right) T_{k,i}^{(t)}}{\sum_{i=1}^{N} T_{k,i}^{(t)}}. \tag{8}$$

Step 7. Let $t = t + 1$ and repeat Steps 2 to 7 until convergence at a date T. At that date, parameters are estimated.

Step 8. For each i, the chosen cluster is k maximizing $T_{k,i}^{(T)}$.

Step 9. For each i, the likelihood of this point for the estimated parameters is

$$P\left(d_i\right) \sum_{l=1}^{K} f\left(w_i \mid m_l^{(T)}, \Sigma_l^{(T)}\right) \alpha_{l,d_i}^{(T)}. \tag{9}$$

4. Comparison of Models

All five models defined in Section 3 are implemented with R [3] into a framework that is able to perform computations and to show clustering and anomaly identification plots (using ggplot2 [25]). In this section, we apply our framework to a sample set to compare abilities to detect anomalies and check robustness of the methods. The sample set is built to highlight the difference of behaviors between models in a simple and understandable context. Consequently, only one sample feature is considered in addition to time-stamp dates.

In this set, we observe that time-GMM and GPLSA are able to detect anomalies within the set, and those methods are then potential candidates for anomaly detection in a time-dependent context. Furthermore, we show that GPLSA is more robust and allows a higher interpretation level of resulting clusters.

4.1. Sample Definition. The sample is built by superposing the three following random sets:

$$\begin{aligned} t &\longmapsto \cos\left(\frac{2\pi t}{T}\right) + \varepsilon, \\ t &\longmapsto \cos\left(\pi + \frac{2\pi t}{T}\right) + \varepsilon, \\ t &\longmapsto -2.5 + \varepsilon, \end{aligned} \tag{10}$$

where ε is independent random variables for each t sampled according to the continuous uniform distribution on $[0, 1]$ and where T has a daily period. The range of the two first functions is 24 hours, whereas the third one is only defined from 0:00 to 15:00.

Three anomalies are added on this set, defined, respectively, at 6:00, 12:00, and 18:00 with values −1.25, 0.5, and 1.65. The resulting set is shown in Figure 1.

4.2. Anomaly Identification. All five models are trained and the likelihood of each point is computed for each model. Since we expect 3 anomalies to be found in this sample set, the 3 lowest likelihood values are defined as anomalies for each model. For the clustering process, the chosen number of clusters is $K = 5$.

The results are shown in Figure 1. In (a), the whole data set is modeled as one Gaussian distribution and we found no expected anomalies. In (b), each period is determined with a Gaussian distribution, and we only discovered the anomaly at 18:00. In (c), the whole set is clustered and we only discovered the anomaly at 6:00. Finally, in (d), the time-GMM and GPLSA models are trained and the same results obtained: the 3 anomalies were successively detected.

Thus, time-GMM and GPLSA are both able to detect expected anomalies contrary to other methods.

4.3. Comparison between Time-GMM and GPLSA. The same anomalies have been detected with time-GMM and GPLSA. However, they are detected differently. We offer a summary of the comparison in Table 2.

First, GPLSA evaluates time stamps and values at once; that is, all parameters are estimated at the same time.

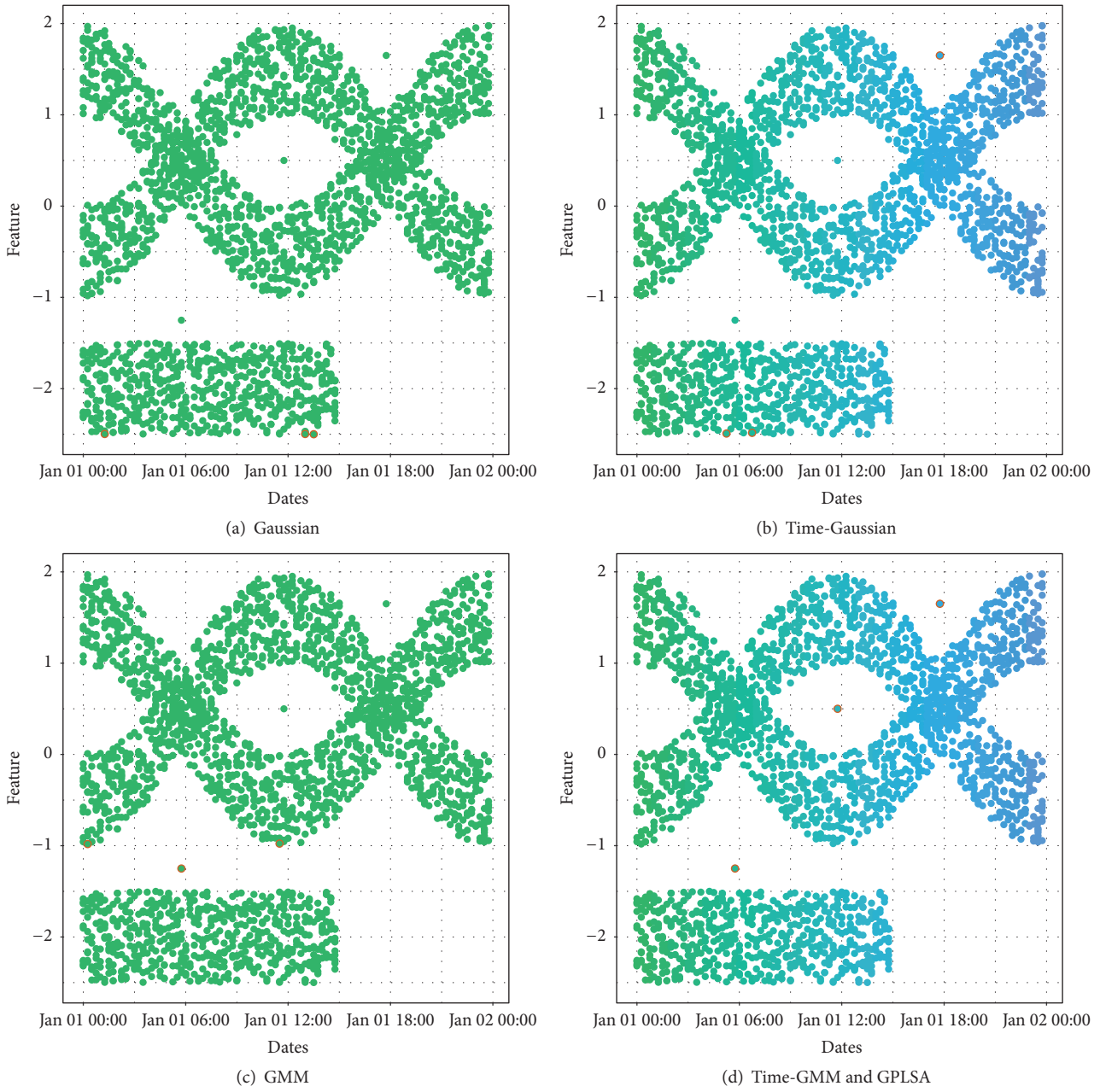


Figure 1: Anomaly detection for 5 different models in the sample set defined in Section 4. The three values with the lowest likelihood are circled in orange. Each color represents a different time-stamp class (only 1 class for (a) and (c); 24 classes for (b) and (d)).

Table 2: Comparison between time-GMM and GPLSA.

	Time-GMM	GPLSA
Cluster number	Fixed number of clusters at each date	Number of clusters can adapt to each date
Cluster relations	No relation between clusters of each date	Homogeneity of clusters across dates
Interpretability	Low	High
Data used	Only a part of data is used at each date	All data is used for each date
Nb: of param.	$(3K-1)D$	$(D+2)K$
Robustness	Medium	High

Consequently, consecutive dates can share similar clustering behaviors. With time-GMM, parameters are trained independently for each class of dates, and no relation exists between the clusters of different classes.

Second, the number of clusters in each class is soft for GPLSA (i.e., it can be different to the specified number of clusters for some class of dates). This allows the model to automatically adapt the number of clusters depending on which cluster is needed in the model. In time-GMM, each class has a specified number of clusters. This is shown in Figure 2, where the first seven hours are plotted in identified clusters for time-GMM (a) and GPLSA (b).

Third, the model is trained with the whole data for GPLSA, whereas only a fraction of data is used for each time-GMM computation. If there is a limited number of data in a class of dates, this can cause a failure to correctly estimate time-GMM parameters.

Fourth, the number of parameters needed for estimation is $(D + 2) \times K$ for GPLSA and $(3K - 1) \times D$ for time-GMM (with D number of classes and K number of clusters, and in dimension $p = 1$). Consequently, there are fewer parameters to estimate with GPLSA.

On the whole, GPLSA implies a better interpretation level (first and second points) of resulting clusters over time-GMM, combined with a higher robustness (third and fourth points).

5. Results and Discussion

In this section, anomaly detection is performed on real traffic network data. Based on the comparison of models done in Section 4, we select GPLSA to deduce anomalies and compare results with time-GMM. In Section 5.1, the collected data set is described and preprocessed; then, we apply GPLSA and show the results in Section 5.2. This Section 5.2 specifically focuses on behavior observed after applying the algorithm. Those results are compared with time-GMM results in Section 5.3. Finally, Section 5.4 highlights the ability of GPLSA to perform anomaly detection.

5.1. Data Description and Preprocessing. Data have been gathered from a Chinese mobile operator. They comprise a selection of 24 traffic features collected for 3,000 cells in the city of Wuxi, China. The features are only related to cell sites and do not give information about specific users. They represent, for example, the average number of users within a cell or the total data traffic for the last quarter of hour. The algorithm is trained over two weeks, with one value for each quarter of hour and for each cell.

We discarded the rows of data containing missing values. Only values and time stamps were taken into consideration for computations, and the identification number of cells was discarded. Some features only take nonnegative values and have a skewed behavior, and consequently, some features are preprocessed by applying the logarithm. To maintain interpretability, we do not apply feature normalization on variables. We expect that GPLSA can manage this set, even though some properties of the model are not verified, such as normality assumptions.

5.2. Computations and Results. We used the GPLSA model for the feature corresponding to the "average number of users within cell" and selected $K = 3$ clusters. Anomalies are values with the lowest resulting likelihood, computed to get (on average) 2 alerts and 8 warnings each day. Visual results are shown on Figure 3.

In (a), the three clusters are identified, whereas, in (b), a different color is used for each class of dates. In (c), the different log-likelihood values are shown. Finally, in (d), the estimation of the probability $\alpha_{k,s}$ to be in each cluster k knowing $D = s$ is plotted.

Anomalies are shown in (a), (b), and (c) and the extreme values related to each class of dates are correctly detected. In (a) and (d), identified clusters are shown in three distinct colors. The probability to be in each cluster varies across class as expected, with a lower probability in the upper cluster during off-peak hours. Also, as shown in (a), the upper cluster has a symmetric shape and the mean value is relatively similar across dates.

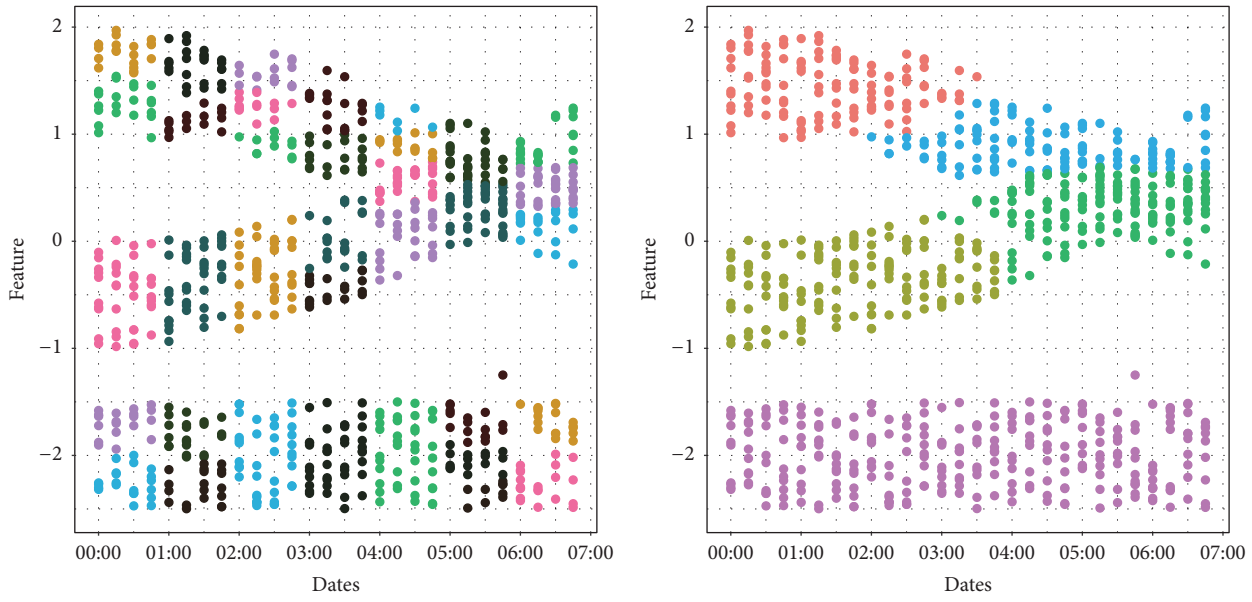
5.3. Comparison with Time-GMM. We compare results obtained in Section 5.2 with time-GMM, using the same number of clusters $K = 3$, and the same number of alerts and warnings each day. We show results on Figure 4. In (a), the three clusters are identified for each class D (between 1 and 24) and in (b), the different log-likelihood values are shown.

We observe that time-GMM correctly detects most of extreme values. Each class is related to a specific likelihood function and has its own way to represent data. We see that the cluster extents related to the highest values have a similar width for all classes on Figure 4(a) ($D = 1$ to 24). By comparing Figure 4(b) with Figure 3(c), we observe a larger "bump" (located in green during off-peak hours) for time-GMM. For these reasons, and contrary to GPLSA, anomalies are overrepresented in some classes (e.g., 3 warnings are detected for $D = 8$ for the first two days) whereas others do not contain anomalies for this time period ($D = 6$). Those results endorse the higher level of interpretation and robustness of GPLSA over time-GMM.

5.4. Discussion. According to the results, GPLSA is able to detect anomalies in a time-dependent context. We identified global outliers (e.g., on Figure 3(b) at Apr. 15 16:00 in red) as well as context-dependent anomalies (e.g., at Apr. 15 5:00 in orange). Off-peak periods are taken into consideration, and unusual values specific to those periods detected.

Gaussian hypothesis on GPLSA is not really constraining. As shown in Figure 3(a), clusters are adaptable and try to fit Gaussian distributions. They are appropriate to represent the value distribution for each class of dates and cluster.

Cluster adaptation is shown in Figure 3(d). The three clusters represent different level of values. The upper cluster represents higher values, which are more probable during peak periods. The lower cluster represents lower values, with a roughly constant probability. The third cluster in the middle is also useful to obtain a good anomaly detection behavior (results with $K = 2$ clusters are unable to correctly detect anomalies).



(a) Time-GMM (b) GPLSA

FIGURE 2: Identified clusters for 2 models in the sample set defined in Section 4 between 0:00 and 7:00. In (a), each class of one hour contains 5 clusters, and clusters are not related across hours. In (b), the whole set contains 5 clusters.

About anomaly detection itself, a threshold indicating the number of alerts to be detected can be set. This method of detection is static and relatively simple. Improving this method of detection is possible and straightforward through likelihood computations: inside a cell, an anomaly could be detected with a repetition of low likelihood scores.

6. Conclusion

In this paper, we present and compare unsupervised models to detect anomalies in wireless network traffic and demonstrated the robustness, interpretability, and ability of the GPLSA model to detect anomalies, as compared to other methods such as time-GMM. Anomaly detection was also performed and analyzed on real traffic data. We highlighted the adaptability of the GPLSA in this context to detect anomalies, even those with new patterns that are difficult to manually predict. As a result, mobile operators can have a versatile way to identify and detect anomalies, which would reduce the cost of possible aftermaths (e.g., network failure).

Improvement of this methodology could be operated. Currently, once the model is computed, anomaly detection is only based on punctual detection through likelihood values. A dynamic detection from consecutive values of likelihood could increase credibility of each alert and reduce the number of false alarms.

Furthermore, the model is only trained from a fixed data set in this research. But this could be extended by considering real-time stream data dealt with in an online context. Thus, new patterns could be updated quickly, to improve responsiveness and anomaly identification.

Appendix

A. Recall of Hypotheses for GPLSA

We recall that $X = (W, D)$ are observed data and Z are latent unobserved data.

Observed values are $(x_i)_{i\in\{1,\dots,N\}}$, where $x_i = (d_i, w_i)$. Each traffic value w_i is a vector of $\mathbf{R}^P$, where p is the number of features. Each time stamp d_i is an integer in $\{1, \dots, S\}$, with $S = 24$. In the following, levels of $\{1, \dots, S\}$ are indexed with s.

Latent values are $(z_i)_{i\in\{1,\dots,N\}}$. They take a finite number of states $k \in \{1, \dots, K\}$, where K is the defined number of clusters.

We recall the different hypotheses for GPLSA:

(H) The set of triplets $(W_i, Z_i, D_i)_i$ is an independent vector over the rows i.

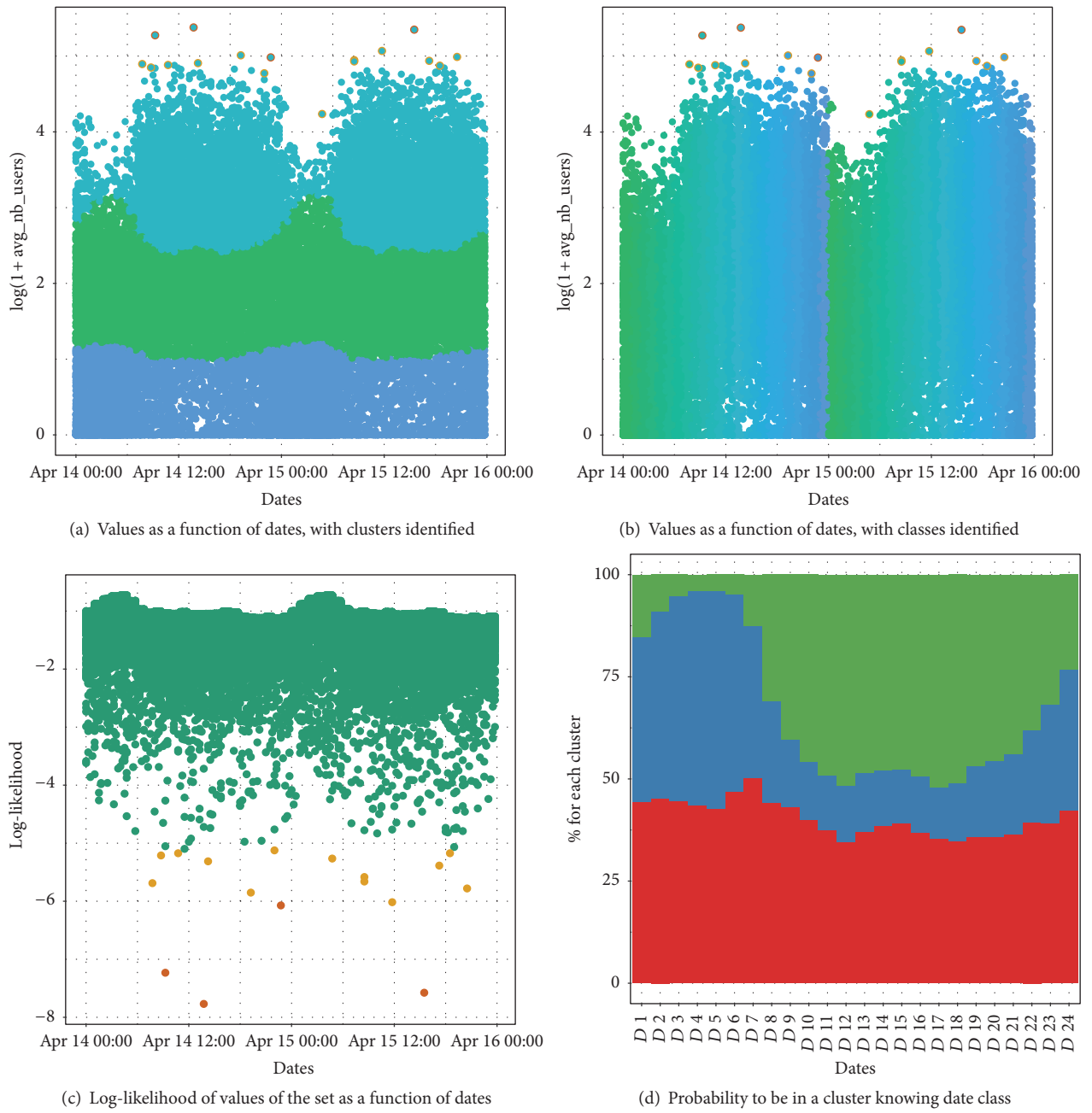
(E1) For each $s \in \{0, \dots, 23\}$, each record such that $D_i = s$ belongs to a cluster $Z_i = k \in \{1, \dots, K\}$ with probability $\alpha_{k,s}$.

(E2) Each variable $(W_i \mid Z_i = k)$ follows a Gaussian distribution with mean and variance m_k, Σ_k.

(E3) For all i, $P(W_i \mid D_i, Z_i) = P(W_i \mid Z_i)$.

Unknown parameters of the model are grouped together into $\theta := (\alpha_{k,s}, m_k, \Sigma_k;\ k \in \{1, \dots, K\},\ s \in \{1, \dots, S\})$.

Initial estimated parameters $\theta^{(0)}$ are defined as follows: all terms $\alpha_{k,s}^{(0)}$ are equal to $1/K$, and $(m_k^{(0)}, \Sigma_k^{(0)})$ are initialized using K-means clustering.



(a) Values as a function of dates, with clusters identified

(b) Values as a function of dates, with classes identified

(c) Log-likelihood of values of the set as a function of dates

(d) Probability to be in a cluster knowing date class

FIGURE 3: Anomaly detection with GPLSA from traffic data set presented in Section 5. Plots are restricted to two days in (a), (b), and (c). Red and orange points are related to the lowest likelihood values obtained, with an average of 2 red points and 8 orange points each day.

We define at each iteration t

$$\theta^{(t)} := \left(\alpha_{k,s}^{(t)}, m_k^{(t)}, \Sigma_k^{(t)};\ k \in \{1,\dots,K\},\ s \in \{1,\dots,S\}\right). \tag{A.1}$$

Estimated parameters $\theta^{(t)}$ are updated from $\theta^{(t-1)}$ iteratively using the EM algorithm. The algorithm stops when convergence of the related likelihood is reached.

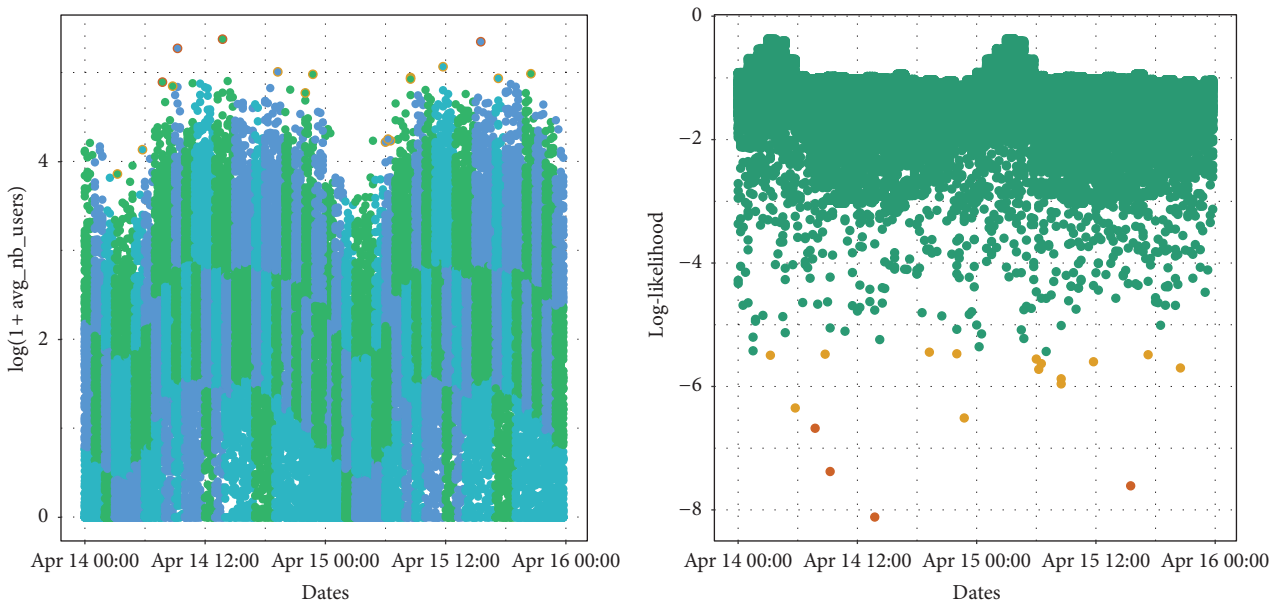
We use our hypotheses to express useful probabilities using θ. We recall that $f(\cdot \mid m, \Sigma)$ is density of a Gaussian with parameters m and Σ. Let $w_i \in \mathbf{R}^p$, $k \in \{1,\dots,K\}$ and $s \in \{1,\dots,S\}$.

From (E3), we know that $P(W_i = w_i \mid \theta,\ D_i = s,\ Z_i = k) = P(W_i = w_i \mid \theta,\ Z_i = k)$. Applying (E2), we deduce that $P(W_i = w_i \mid \theta,\ Z_i = k) = f(w_i \mid m_k, \Sigma_k)$. On the whole, we obtain

$$P\left(W_i = w_i \mid \theta,\ D_i = s,\ Z_i = k\right) = f\left(w_i \mid m_k, \Sigma_k\right). \tag{A.2}$$

Also, from (E1),

$$P\left(Z_i = k \mid D_i = s, \theta\right) = \alpha_{k,s}. \tag{A.3}$$



(a) Values as a function of dates, with clusters identified (cluster identification is independent for each time slot)

(b) Log-likelihood of values of the set as a function of dates

FIGURE 4: Anomaly detection with time-GMM from traffic data set as presented in Section 5. Plots are restricted to two days. Red and orange points are related to the lowest likelihood values obtained, with an average of 2 red points and 8 orange points each day.

This probability follows a discrete multinomial distribution that is proportional to $\alpha_{k,s}$ (where for each d, the coefficients sum to 1 over all k).

B. Recall about EM

The chosen strategy to estimate parameters θ is to find some parameters that maximize the marginal likelihood of the observed data X, as defined by

$$L(\theta; X) = P(X \mid \theta) = \sum_{Z} P(X, Z \mid \theta). \tag{B.1}$$

As the direct computations are intractable, we use EM to update parameters iteratively:

(1) Set some initial parameters $\theta^{(0)}$. For t from 0 until convergence, repeat the following steps (2) and (3).

(2) Perform the expectation step (E step):

$$Q\left(\theta \mid \theta^{(t)}\right) \coloneqq \mathrm{E}_{Z \mid X, \theta^{(t)}}\left[\log\left(L(\theta; X, Z)\right)\right] \tag{B.2}$$

which can be rewritten as

$$Q\left(\theta \mid \theta^{(t)}\right) = \sum_{Z} \log P(X, Z \mid \theta)\, P\left(Z \mid X, \theta^{(t)}\right). \tag{B.3}$$

(3) Perform the maximization step (M step):

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q\left(\theta \mid \theta^{(t)}\right). \tag{B.4}$$

A theoretical reason to update the expected value function $Q(\cdot \mid \theta^{(t)})$ is that likelihood $L(\theta; X)$ will increase or remain constant at each step [26]. However, after convergence, the parameters can be stuck in a local maximum of the likelihood function.

C. Expectation Step of EM in the GPLSA Context

We assume we are in step t, and we want to update $m_k^{(t)}$, $\Sigma_k^{(t)}$, $\alpha_{k,s}^{(t)}$ for all k and s. From (B.3) and using hypothesis (H), we get

$$Q\left(\theta \mid \theta^{(t)}\right) = \sum_{i=1}^{N} \sum_{k=1}^{K} \log P\left(X_i = x_i,\ Z_i = k \mid \theta\right) \cdot P\left(Z_i = k \mid X_i = x_i, \theta^{(t)}\right). \tag{C.1}$$

For the left term, since $X_i = (D_i, W_i)$ and using equations (A.2) and (A.3),

$$P\left(D_i = d_i,\ W_i = w_i,\ Z_i = k \mid \theta\right) = f\left(w_i \mid m_k, \Sigma_k\right) \alpha_{k,d_i} P\left(D_i = d_i\right). \tag{C.2}$$

For the right term, using (A.2) and (A.3) for parameter $\theta^{(t)}$,

$$P\left(Z_i = k \mid X_i = x_i, \theta^{(t)}\right) = \frac{P\left(W_i = w_i \mid Z_i = k\right) P\left(Z_i = k \mid D_i = d_i\right)}{\sum_{l=1}^{K} P\left(W_i = w_i \mid Z_i = l\right) P\left(Z_i = l \mid D_i = d_i\right)}, \tag{C.3}$$

$$P\left(Z_i = k \mid X_i = x_i, \theta^{(t)}\right) = \frac{f\left(w_i \mid m_k^{(t)}, \Sigma_k^{(t)}\right)\alpha_{k,d_i}^{(t)}}{\sum_{l=1}^{K} f\left(w_i \mid m_l^{(t)}, \Sigma_l^{(t)}\right)\alpha_{l,d_i}^{(t)}}. \tag{C.4}$$

Then we define $T_{k,i}^{(t)}$ as $P(Z_i = k \mid X_i = x_i, \theta^{(t)})$, which is explicitly computable from (C.4).

Seen in the whole,

$$Q\left(\theta \mid \theta^{(t)}\right) = \sum_{i=1}^{N}\sum_{k=1}^{K} \log\left[f\left(w_i \mid m_k, \Sigma_k\right)\alpha_{k,d_i} P\left(D_i = d_i\right)\right] T_{k,i}^{(t)}, \tag{C.5}$$

$$Q\left(\theta \mid \theta^{(t)}\right) = \sum_{i=1}^{N}\sum_{k=1}^{K}\left[-\frac{K}{2}\log 2\pi - \frac{1}{2}\cdot \log\left|\Sigma_k\right| \frac{1}{2}\left(x_i - m_k\right)' \Sigma_k^{-1}\left(x_i - m_k\right) + \log\alpha_{k,d_i} + \log P\left(D = d_i\right)\right] T_{k,i}^{(t)}. \tag{C.6}$$

Finally, we obtain an explicit formula for $Q(\theta \mid \theta^{(t)})$ which can be maximized.

D. Expectation Step of EM in the GPLSA Context

From the shape (C.6) of $Q(\cdot \mid \theta^{(t)})$, we can separate maximization of (m_k, Σ_k) for each k and weights $(\alpha_{k,s})_k$ for each s.

(1) For the Weights $\alpha_{k,s}$. For each fixed time stamp s, we update $(\alpha_{k,s})_k$. These are considered all together since there is a constraint: the sum over k has to be 1. From (C.6), we only have to maximize

$$G\left((\alpha_{k,s})_k\right) = \sum_{i=1}^{N}\sum_{k=1}^{K}\log\alpha_{k,d_i} T_{k,i}^{(t)} = \sum_{k=1}^{K}\sum_{i=1}^{N}\log\alpha_{k,d_i} T_{k,i}^{(t)}. \tag{D.1}$$

For each $s \in \{1, \ldots, S\}$, we let E_s the set of indexes $i \in \{1, \ldots, N\}$ such that $d_i = s$. Therefore

$$G\left((\alpha_{k,s})_k\right) = \sum_{k=1}^{K}\sum_{s=1}^{S}\sum_{j=1}^{\#E_s}\log\alpha_{k,s} T_{k,E_s(j)}^{(t)}, \qquad G\left((\alpha_{k,s})_k\right) = \sum_{k=1}^{K}\sum_{s=1}^{S}\log\alpha_{k,s}\sum_{j=1}^{\#E_s} T_{k,E_s(j)}^{(t)}. \tag{D.2}$$

We let for all k, s: $S_{k,s}^{(t)} := \sum_{j=1}^{\#E_s} T_{k,E_s(j)}^{(t)}$ to obtain

$$G\left((\alpha_{k,s})_k\right) = \sum_{k=1}^{K}\sum_{s=1}^{S} S_{k,s}^{(t)}\log\alpha_{k,s}. \tag{D.3}$$

Since s is fixed, all terms except one are constant. Consequently, we only have to maximize

$$F\left((\alpha_{k,s})_k\right) := \sum_{k=1}^{K} S_{k,s}^{(t)}\log\alpha_{k,s}. \tag{D.4}$$

Finally, we compute the derivative with respect to $\alpha_{k,s}$. Here, we remember that $\sum_{k=1}^{K}\alpha_{k,s} = 1$. To remove this constraint, we let $\alpha_{K,s} = 1 - (\alpha_{1,s} + \cdots + \alpha_{K-1,s})$. We rewrite this as

$$F\left((\alpha_{k,s})_k\right) = \sum_{k=1}^{K-1} S_{k,s}^{(t)}\log\alpha_{k,s} + S_{K,s}^{(t)}\log\left(1 - \left(\alpha_{1,s} + \cdots + \alpha_{K-1,s}\right)\right). \tag{D.5}$$

By differentiation,

$$\frac{\partial F\left((\alpha_{k,s})_k\right)}{\partial\alpha_{k,s}} = \frac{s_{k,s}^{(t)}}{\alpha_{k,s}} - \frac{s_{K,s}^{(t)}}{\alpha_{K,s}}, \tag{D.6}$$

$$\frac{\partial F\left((\alpha_{k,s})_k\right)}{\partial\alpha_{k,s}} = \frac{S_{k,s}^{(t)}}{\alpha_{k,s}} - \frac{S_{K,s}^{(t)}}{\alpha_{K,s}}. \tag{D.7}$$

If we want this value (D.6) to be zero, we get

$$\alpha_{k,s} = \alpha_{K,s}\frac{S_{k,s}^{(t)}}{S_{K,s}^{(t)}}. \tag{D.8}$$

Now, using the constraint

$$1 = \alpha_{K,s}\frac{\sum_{k=1}^{K} S_{k,s}^{(t)}}{S_{K,s}^{(t)}} + \alpha_{K,s} \tag{D.9}$$

then,

$$\alpha_{K,s} = \frac{S_{K,s}^{(t)}}{\sum_{k=1}^{K} S_{k,s}^{(t)}}. \tag{D.10}$$

This follows for all $k \in \{1, \ldots, K\}$

$$\alpha_{k,s} = \frac{S_{k,s}^{(t)}}{\sum_{l=1}^{K} S_{l,s}^{(t)}}. \tag{D.11}$$

By computing the Hessian matrix, we find that the obtained extremum is the maximum value.

(2) For the Means and Variances (m_k, Σ_k). From (C.6), we can perform computations for each fixed cluster k. Since some terms of this sum have no dependence on k, we have to maximize

$$-\frac{1}{2}\sum_{i=1}^{N}\left[\log\left|\Sigma_k\right| + \left(x_i - m_k\right)'\Sigma_k^{-1}\left(x_i - m_k\right)\right] T_{k,i}^{(t)} \tag{D.12}$$

or minimize

$$\sum_{i=1}^{n}\left[\log\left|\Sigma_k\right| + \left(x_i - m_k\right)'\Sigma_k^{-1}\left(x_i - m_k\right)\right] T_{k,i}^{(t)}. \tag{D.13}$$

We obtain the same formula as for GMM and then give the update rules with

$$m_k = \frac{\sum_{i=1}^{n} x_i T_{k,i}^{(t)}}{\sum_{i=1}^{n} T_{k,i}^{(t)}}, \qquad \Sigma_k = \frac{\sum_{i=1}^{n}\left(x_i - m_k\right)'\left(x_i - m_k\right) T_{k,i}^{(t)}}{\sum_{i=1}^{n} T_{k,i}^{(t)}}. \tag{D.14}$$

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] G. J. McLachlan and K. E. Basford, "Mixture models. Inference and applications to clustering," in *Statistics: Textbooks and Monographs*, Dekker, New York, NY, USA, 1988.

[2] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, 1999.

[3] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016, https://www.R-project.org/.

[4] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: a survey," *ACM Computing Surveys*, vol. 41, no. 3, article 15, 2009.

[5] P. Laskov, P. Düssel, C. Schäfer, and K. Rieck, "Learning intrusion detection: supervised or unsupervised?" in *Image Analysis and Processing—ICIAP 2005: 13th International Conference, Cagliari, Italy, September 6–8, 2005. Proceedings*, vol. 3617 of *Lecture Notes in Computer Science*, pp. 50–57, Springer, Berlin, Germany, 2005.

[6] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004.

[7] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: classification of skewed data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 50–59, 2004.

[8] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: existing solutions and latest technological trends," *Computer Networks*, vol. 51, no. 12, pp. 3448–3470, 2007.

[9] V. Bamnett and T. Lewis, *Outliers in Statistical Data*, 1994.

[10] D. Agarwal, "Detecting anomalies in cross-classified streams: a Bayesian approach," *Knowledge and Information Systems*, vol. 11, no. 1, pp. 29–44, 2007.

[11] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B: Methodological*, vol. 39, no. 1, pp. 1–38, 1977.

[13] H. Hajji, "Statistical analysis of network traffic for adaptive faults detection," *IEEE Transactions on Neural Networks*, vol. 16, no. 5, pp. 1053–1063, 2005.

[14] Y. Ouyang, M. H. Fallah, S. Hu et al., "A novel methodology of data analytics and modeling to evaluate LTE network performance," in *Proceedings of the 13th Annual Wireless Telecommunications Symposium (WTS '14)*, IEEE, Washington, DC, USA, April 2014.

[15] M. J. Desforges, P. J. Jacob, and J. E. Cooper, "Applications of probability density estimation to the detection of abnormal conditions in engineering," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 212, no. 8, pp. 687–703, 1998.

[16] M. M. T. Chiang and B. Mirkin, "Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads," *Journal of Classification*, vol. 27, no. 1, pp. 3–40, 2010.

[17] M. Gupta, J. Gao, C. Aggarwal, and J. Han, "Outlier detection for temporal data," *Synthesis Lectures on Data Mining and Knowledge Discovery*, vol. 5, pp. 1–129, 2014.

[18] S. Rayana and L. Akoglu, "Less is more: building selective anomaly ensembles," *ACM Transactions on Knowledge Discovery from Data*, vol. 10, no. 4, article 42, 2016.

[19] Y. Ouyang and T. Yan, "Profiling wireless resource usage for mobile apps via crowdsourcing-based network analytics," *IEEE Internet of Things Journal*, vol. 2, no. 5, pp. 391–398, 2015.

[20] J. Guo, W. Huang, and B. M. Williams, "Real time traffic flow outlier detection using short-term traffic conditional variance prediction," *Transportation Research Part C: Emerging Technologies*, vol. 50, pp. 160–172, 2015.

[21] A. Y. Lokhov, N. Lemons, T. C. McAndrew, A. Hagberg, and S. Backhaus, "Detection of cyber-physical faults and intrusions from physical correlations," https://arxiv.org/abs/1602.06604.

[22] Y. Ouyang and H. M. Fallah, "A performance analysis for UMTS packet switched network based on multivariate KPIs," in *Proceedings of the IEEE Wireless Telecommunications Symposium*, Tampa, Fla, USA, April 2010.

[23] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, Berkeley, Calif, USA, 1999.

[24] Z. Lu, W. Pan, E. W. Xiang, Q. Yang, L. Zhao, and E. H. Zhong, "Selective transfer learning for cross domain recommendation," https://arxiv.org/abs/1210.7056.

[25] H. Wickham, *Elegant Graphics for Data Analysis*, Springer Science & Business Media, 2009.

[26] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, New York, NY, USA, 1987.

Hindawi Publishing Corporation
Mobile Information Systems
Volume 2016, Article ID 5894752, 10 pages
http://dx.doi.org/10.1155/2016/5894752

Research Article

A Categorized Resource Sharing Mechanism for Device-to-Device Communications in Cellular Networks

Jie Chen,[1] Chang Liu,[2] Husheng Li,[3] Xulong Li,[1] and Shaoqian Li[1]

[1]*University of Electronic Science and Technology of China, Chengdu, China*
[2]*Dalian University of Technology, Dalian, China*
[3]*Department of Electrical Engineering and Computer Science, The University of Tennessee, Knoxville, TN, USA*

Correspondence should be addressed to Jie Chen; jiechenuestc@163.com

Received 12 July 2016; Revised 30 September 2016; Accepted 24 October 2016

Academic Editor: Piotr Zwierzykowski

Copyright © 2016 Jie Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Device-to-Device (D2D) communications are considered one of the key technologies for 5G wireless communication systems. In this paper, a resource sharing mechanism, which applies different policies for different cases (thus being categorized), is proposed. In this scheme, all D2D pairs are divided into three groups by comparing the minimum transmit power with the maximum transmit power of each cellular UE. The proposed mechanism enables multiple D2D pairs in the second group to share the resource with cellular user equipment (UE) simultaneously, by adjusting the transmit powers of these D2D transmitters. At the same time, D2D pairs in the first group and the third group share resource with cellular UE based on the transmit power minimization principle. Simulation results show that the proposed scheme can achieve relatively higher network throughput and lower transmit power consumption of the D2D system.

1. Introduction

The increasing demand for higher data rates within local area and the gradually increasing spectrum congestion have triggered research activities on improving the spectral efficiency and interference management. In recent years, D2D (Device-to-Device) communications have gained much attention [1, 2]. D2D communications enhance the spectral efficiency by spatially reusing radio resource and prolong the battery life of user equipment (UE) by reducing the transmission power. Due to these advantages, the D2D communications have been actively discussed in standardization bodies for the next generation cellular systems such as Long Term Evolution-Advanced (LTE-A) [3].

However, D2D links may yield significant interference to the communication system. Therefore, resource management scheme, which supports the resource reuse by taking the intracell interference into account, has a great impact on the overall network performance. So far, major efforts are to be aimed at interference control through resource sharing mode selection [4], power control [5–7], and resource allocation [8–10]. The previous works [5–7] are mainly concentrated on the power control in a specific resource block without the consideration of resource allocation, while the previous works in [8–10] have focused on the case that only one D2D link can share resource with one cellular link. Additionally, in [11], the authors study joint channel and power allocation to improve the energy efficiency of user equipment. In [12–15], the authors design the power-based resources allocation scheme and achieved relatively better performance, while they do not consider the factor of geography distribution, which may affect the power seriously. In [16], the authors proposed an interference-aware graph based resource sharing algorithm that can effectively obtain the near-optimal resource assignment solutions at the base station but with low computational complexity. In [17], a resource allocation scheme based on a column generation method is proposed. In [18], the authors aimed to optimize resource sharing for D2D communication to better utilize uplink resources in a multiuser cellular system with guaranteed quality of normal cellular communications.

In this paper, to further improve spectrum utilization and system capacity, a resource sharing method, which enables multiple D2D links to share resource with cellular UE

simultaneously, is proposed. Firstly, the minimum transmit power of a D2D transmitter is calculated by the required minimum Signal-to-Interference-plus-Noise Ratio (SINR) and the interference from cellular UE which shares the resource with the D2D link. Secondly, based on the interference threshold of eNodeB, the maximum transmit power of a D2D transmitter is calculated. Thirdly, by comparing the minimum transmit power with the maximum one on the resource of each cellular UE, a set of cellular UE devices that can share resource with the D2D link is attained. Then, the D2D pairs are divided into three groups according to the comparison result.

Finally, when many D2D links can only share resource with some special cellular UE or the number of D2D links is larger than the number of cellular UE devices, the transmit powers of these D2D transmitters are adjusted to ensure that the cumulative interference to eNodeB is below a threshold; and the minimum SINR value of each D2D link is met as well. After that, they can share resource with the cellular UE simultaneously. Simulation results show that the proposed scheme can achieve a higher network capacity and lower transmit power consumption of the D2D system.

Then, we can summarize the main contributions in the following:

(i) Different from the existing works, we analyze the resource sharing from the feasibility of transmit power and design the power bound principle by the requirement of minimum SINR and maximum interference, which provides a novel power management scheme.

(ii) According to the different geometry distribution (distances of various nodes in the system), the resources are divided into 3 groups based on the power bound principle, thus fully utilizing the UE resources.

(iii) Taking advantage of the adjustable transmit power of D2D, the proposed mechanism enables multiple D2D pairs in the second group to share the resource with cellular UE simultaneously. In addition, D2D pairs in the first group and the third group can share resource with cellular UE based on the transmit power minimization principle, which further improves the power efficiency.

The remainder of this paper is organized as follows. The next section describes the system model. The resource sharing method between cellular UE and D2D pairs is presented in Section 3. The performance of the proposed method is evaluated in Section 4 and the paper is concluded with Section 5.

2. System Model

Considering an OFDMA based cellular network, which is frequency division duplex (FDD), and concentrating on a single cell served by eNodeB as depicted in Figure 1, it is assumed that the eNodeB knows the path-loss components between any two UE devices and between any UE and the eNodeB, based on the locations of the UE or the average

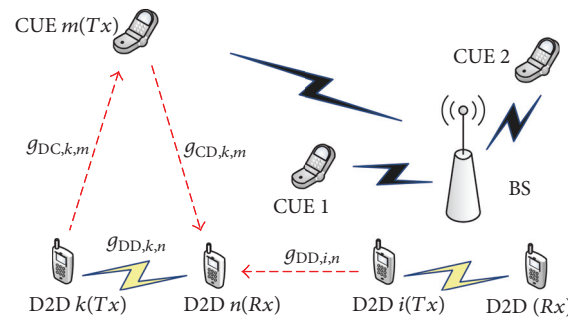


Figure 1: Device-to-Device communication scenarios in a cellular network.

channel qualities [19]. In the cell, there are m cellular UE devices, where $m = 1, \ldots, M$. In addition to m cellular UE devices, there are k pairs of D2D UE devices which communicate directly with each other, where $k = 1, \ldots, K$. Consider the notion that D2D UE devices only share uplink resource with cellular UE. Furthermore, it is also assumed that exclusive resources are reserved for N pairs of D2D UE devices.

In this work, we focus on the factor of path loss due to the different distance between the D2D pairs and UE. The different distances lead to the different channel gains and thus can further affect the transmit power. According to Figure 1, $g_{DD,k,n}$ represents the channel gain between D2D k and D2D i, while $g_{CD,m,k}$ signifies the channel gain between cellular UE m and D2D k, and $g_{DC,k,m}$ denotes the channel gain between D2D k and cellular UE m

3. The Proposed Resource Sharing Mechanism

3.1. Conditions for Sharing Resource between a Cellular Link and D2D Links. We define u as a set of D2D links, which can share uplink resources with the cellular UE m if and only if the following condition is met:

$$\mathrm{SINR}_k = \frac{P_k g_{DD,k,n}}{N_k + I_{m,k} + \sum_{i \in u, i \neq k} P_i g_{DD,i,k}} \geq T_0, \quad \forall k \in u, \tag{1}$$

$$\sum_{k \in u} P_k g_{DC,k,m} \leq I_0, \tag{2}$$

$$I_{m,k} = P_m g_{CD,m,k}. \tag{3}$$

Here, T_0 and I_0 are defined as the minimal SINR value of D2D k and the interference threshold value of cellular UE, respectively. P_k and P_m denote the transmit power of D2D link k and transmit power of cellular UE m, respectively.

3.2. Resource Sharing and Power Adjustment

3.2.1. Acquiring Transmit Power Matrix. We define a $K \times M$ matrix X, of which the (k, m)th element $x_{k,m}$ denotes the

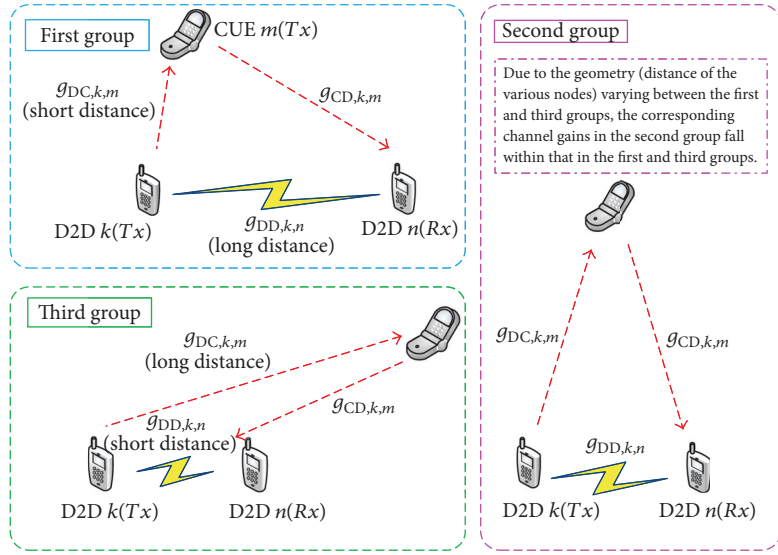


Figure 2: Grouping illustration for a cellular network.

minimum transmitter power of D2D k that shared resources with cellular UE m, where $x_{k,m}$ can be calculated by (4). We define a $K \times M$ matrix Y, of which the (k, m)th element $y_{k,m}$ means the maximum transmitter power of D2D k that shared resources with cellular UE m, where $y_{k,m}$ can be calculated by (5). Hence,

$$\mathrm{SINR}_k = \frac{x_{k,m} g_{\mathrm{DD},k,m}}{N_k + P_m g_{\mathrm{CD},m,k}} \geq T_0 \tag{4}$$

$$I_k = y_{k,m} g_{\mathrm{DB},k,m} \leq I_0. \tag{5}$$

Here, T_0 and I_0 are defined as the minimal SINR value of D2D k and the interference threshold value of cellular UE, respectively.

3.2.2. Grouping D2D UE. According to the condition for sharing resource between the cellular link m and D2D link k, k cannot share resource with m if $x_{k,m} > y_{k,m}$. Then, D2D pairs are divided into the following three groups by comparing each element of matrix X and matrix Y:

(i) First group (if all $x_{k,m}$ are more than all $y_{k,m}$): D2D pairs in the first group cannot share resource with any cellular UE.

(ii) Second group (if part of $x_{k,m}$ are more than part of $y_{k,m}$): D2D pairs in the second group can share resource with some cellular UE.

(iii) Third group (if all $x_{k,m}$ are less than all $y_{k,m}$): D2D pairs in the third group can share resource with all cellular UE.

For further understanding, according to the geography distribution (distances of the various nodes), we then draw Figure 2 to address the categorization problem. It is shown that if the distance between D2D $k(Tx)$ and CUE $m(Tx)$ is short and the distance between D2D $k(Tx)$ and D2D $n(Rx)$ is long, then the channel gain $g_{\mathrm{DC},k,m}$ turns to large value and $g_{\mathrm{DD},k,n}$ turns to small value. In this case, the corresponding minimum transmit power in (4) may be larger than the maximum transmit power in (5), which makes the first group and D2D pairs unable to share resource with any cellular UE. Otherwise, the corresponding minimum transmit power in (4) may be smaller than the maximum transmit power in (5) and this situation belongs to the third group. Hence, as for the second group, due to the geometry (distance of the various nodes) varying between the first and third groups, the corresponding channel gains in the second group fall within that in the first and third groups.

Now, the resource sharing model, which enables multiple D2D UE devices to reuse the resource of cellular UE and minimizes the total transmit power of D2D system, is presented. We concentrate on resource sharing between cellular UE and D2D links in the second group. We construct an L-by-M matrix P by selecting rows corresponding to the D2D links in the second group from matrix X, that is, selecting the ith row from matrix X if D2D link belongs to the second group. Similarly, we construct an L-by-M matrix P by selecting rows from matrix Y. L denotes the number of D2D pairs in the second group. Then, two algorithms are provided for seeking the minimum transmit power value in matrix P under the following two cases.

Case 1 ($L < M$). In Algorithm 1, X and Y denote the row and the column of P. Lines (2)–(9) in Algorithm 1 are used for seeking the minimum transmit power value p in matrix P, then marking the row and column that p belongs to, and continuing searching for the minimum transmit power value in the unmarked rows and columns in matrix P; the algorithm repeats the above operation about L times, thus obtaining L minimum transmit power values. Lines (10)–(18) are explained as follows: if each of these transmit power

Input:
The matrix P;
Output:
A set of minimum transmit power $p = (p_1, p_2, \ldots, p_L)$;
(1) $X \leftarrow \{1, 2, \ldots, L\}$,
(2) do $\{Y \leftarrow \{1, 2, \ldots, M\}$
(3) **for** $k = 1 : L$
(4) **for all** $l \in X$, $m \in Y$
(5) $p_k \leftarrow \min(P_{l,m})$, record l and m
(6) $q_k \leftarrow q_{l,m}$
(7) $X \leftarrow X - l$, $Y = Y - m$
(8) **end for**
(9) **end for**
(10) **for** $k = 1 : L$
(11) **if** $p_k \leq q_k$
(12) **return** p_k
(13) **else**
(14) $p_k \Rightarrow p_{l,m}$, $p_{l,m}$ belongs to matrix P
(15) add the subscript l to X
(16) **end if**
(17) **end for**}
(18) **while** $(X \neq \emptyset)$
(19) $P = (p_1, p_2, \ldots, p_L)$

Algorithm 1: Seeking the minimum transmit power value.

values is less than the corresponding maximum transmit power value, the algorithm returns them. If some of them are larger than the corresponding maximum transmit power, Algorithm 1 continues searching for the minimum transmit power of these D2D links according to lines (2)–(9) until each of them is less than the corresponding maximum transmit power. In line (19), $p_k \Rightarrow p_{l,m}$ means that p_k is the (l, m)th element of matrix P.

Case 2 ($L > M$). In Algorithm 2, first, it searches for M minimum transmit power values from matrix P by calling Algorithm 1. Then, it updates all the columns in matrix P to unmarked and searches for the minimum transmit power values of the other $L - M$ D2D links by calling Algorithm 1. When a feasible set p is obtained, a set of transmit power values in matrix X can be also attained by $p_k \Rightarrow p_{l,m} \Rightarrow x_{k,m}$. If these transmit power values ($x_{k,m}$) belong to different rows and different columns, the D2D links share resource with cellular UE by the subscript of the minimum transmit power; that is, $x_{i,j}$ means that D2D link i shares resource with cellular UE j. However, some minimum transmit power values belong to the same column, which means that multiple D2D links compete for the resource of cellular UE.

Thus, according to (1) and (3), the algorithm adjusts the transmitter power of these D2D links, such that they can share resource with the cellular UE when (2) holds. Otherwise, a D2D link, of which the transmit power value is minimal to share resource with the cellular UE, is selected.

As mentioned above, the target of the resource sharing method is not only to improve the system capacity but also to minimize the transmit power of D2D system. In order to minimize the total transmit power of D2D system, D2D links in the second group and the third group share resource with cellular UE based on the power minimization principle. The exclusive resources prefer to be allocated to the D2D UE of smaller transmit power in the first group, while each of the remaining cellular UE devices prefers to share resource with the D2D UE of smaller transmit power in the third group.

Table 1: Simulation parameters.

Cell radius	500 m
Maximal distance between one D2D pair	50 m
Number of cellular users	25
Number of D2D pairs	25
Bandwidth per RB W_{RB}	180 kHz
Path-loss model	$35.3 + 37.6 \log_{10} d$ dB
Target bit error rate BER[1]	0.01
The probability threshold θ	0.01
Power level of thermal noise N_0	−174 dBm/Hz
The mean μ_φ of multipath fading	1 dB
The standard deviation σ_n of shadowing	8 dB

[1]Note that BER is an important effect factor on the throughput for systems; that is, the larger the BER, the smaller the achievable throughput. Thus, we add it to the simulation parameters for the calculation of achievable throughput.

4. Performance Analysis

In this section, the performance of the proposed resource sharing mechanism is evaluated. First, the simulation parameters are set and then the simulation results are presented and analyzed. All the simulations are operated under the MATLAB environment.

4.1. Simulation Setup. There are M cellular UE devices and K D2D pairs within a single circular cell with the radius of 500 m. It is assumed that the bandwidth of an RB is $W_{\mathrm{RB}} = 180$ kHZ, and the noise spectral density is $N_0 = -174$ dBm/Hz. The path loss, shadowing, and Rayleigh fading are considered. Based on LTE system models [19], the path-loss model is $z = 35.3 + 37.6 \log_{10} d$ (dB), where d is the distance between the transmitter and the receiver. According to [20], it is assumed that the shadowing components of all links are i.i.d., and the shadowing component η follows a log-normal distribution with zero mean and standard deviation σ_n, $\sigma_n = 8$ dB. It is also assumed that the multipath components of all links are independent of each other, and all of them are exponentially distributed with the same mean of μ_φ, $\mu_\varphi = 1$. The minimal SINR threshold value of each D2D pair for normal communication is $T_0 = 0$ dB, and the transmit power of each cellular UE is 20 dBm. We consider the number of D2D pairs and the interference threshold value of eNodeB I_0 as variables of the simulation. Simulation parameters are summarized in Table 1.

4.2. Simulation Results. The simulation results are plotted in Figures 3–8. Before discussing the results, we shortly describe

Input:
The matrix P;
Output:
A set of minimum transmit power $p = (p_1, p_2, \ldots, p_M)$;
(1) $X \leftarrow \{1, 2, \ldots, M\}, Y \leftarrow \{1, 2, \ldots, M\}$
(2) call Algorithm 1
(3) $X \leftarrow \{M + 1, M + 2, \ldots, L\}$, $Y \leftarrow \{1, 2, \ldots, M\}$
(4) call Algorithm 1
(5) $P = (p_1, p_2, \ldots, p_M)$

Algorithm 2: Seeking the minimum transmit power value.

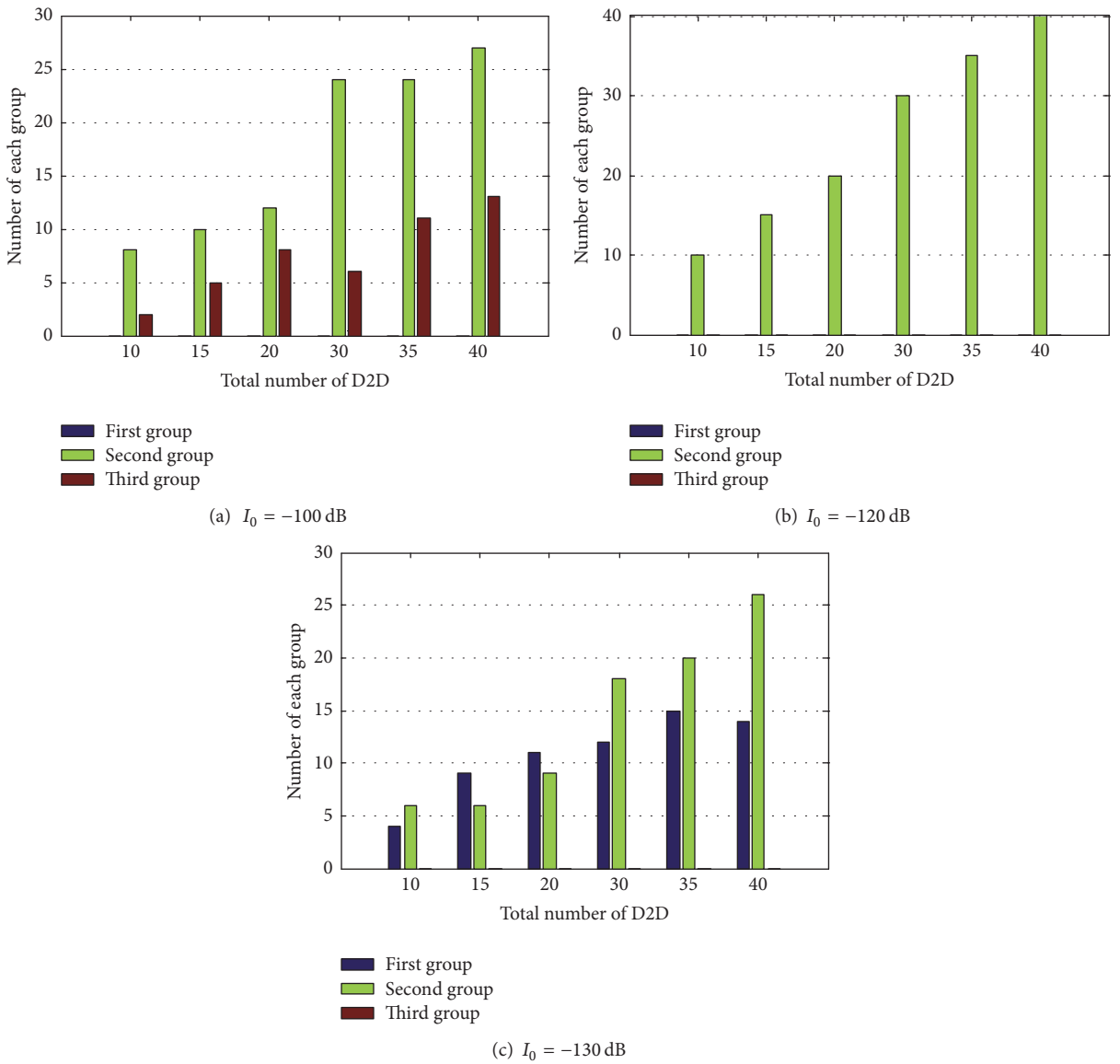


Figure 3: The number of D2D pairs in each group.

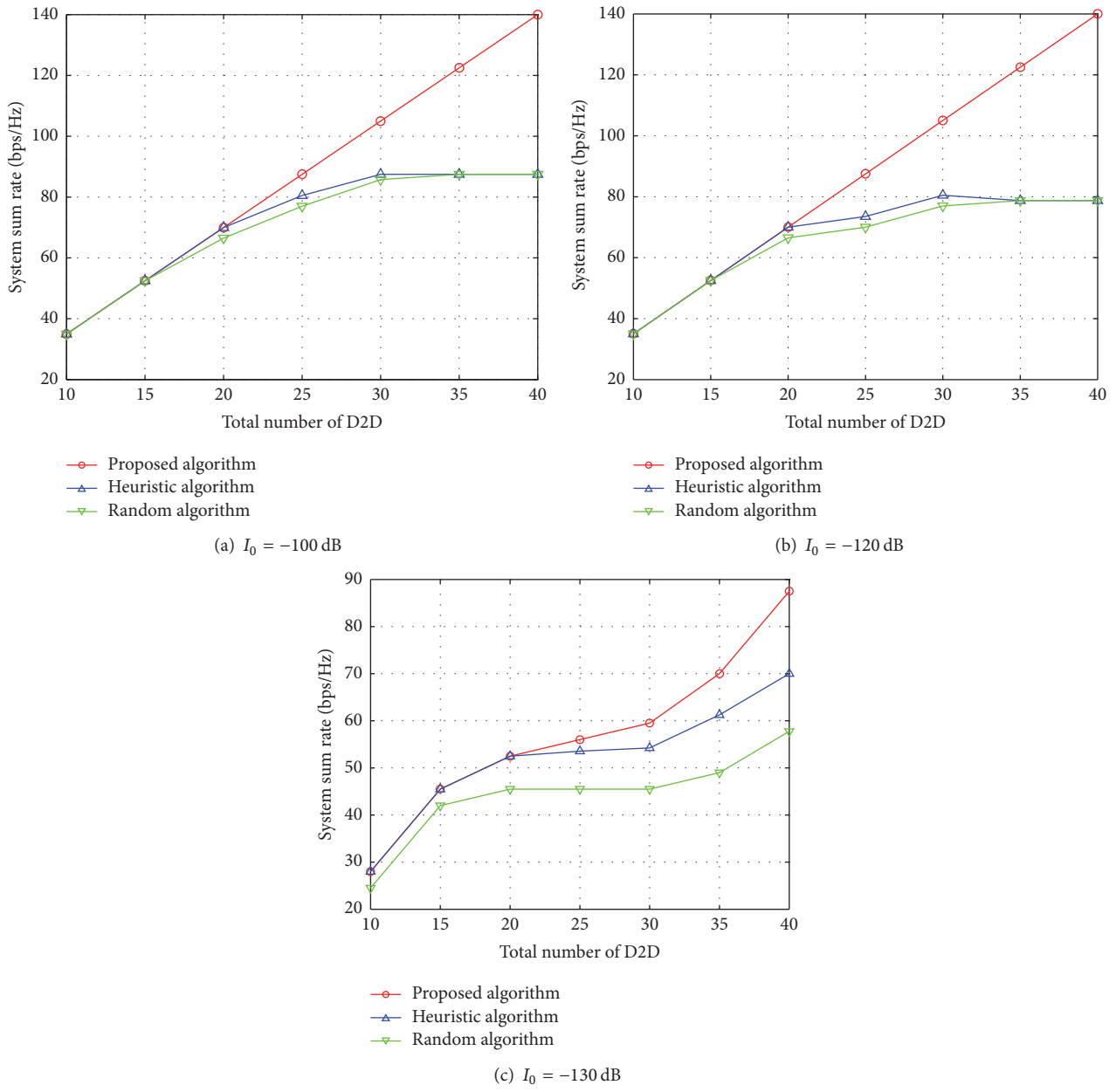


Figure 4: System sum rate with the number of D2D under different interference thresholds.

the random algorithm [21] and the heuristic algorithm [22]; both of them assume that cellular UE can only share resource with a D2D pair. In the random algorithm, a D2D pair shares resource with random cellular UE in the cell, while cellular UE of priority shares resource with a D2D pair which causes the least interference to the eNodeB in the heuristic algorithm.

In Figure 3, the number of D2D pairs in each group is presented. It can be observed that the number of D2D pairs in the first group is increasing while the number of D2D pairs in the third group is decreasing with the interference threshold value of eNodeB becoming smaller. Since D2D pairs in the first group can only use exclusive resource, some of them cannot access the system when exclusive resource is insufficient. Next, we analyze the performance of the proposed scheme.

The system sum rate with the number of D2D under different interference thresholds can be found in Figure 4. It can be seen that the system sum rate of the proposed algorithm is better than the other two algorithms.

In Figure 5, since the proposed algorithm enables multiple D2D UE devices to share resource with the same cellular UE simultaneously, it achieves a higher network capacity than the other two algorithms. From Figure 5(a), we can

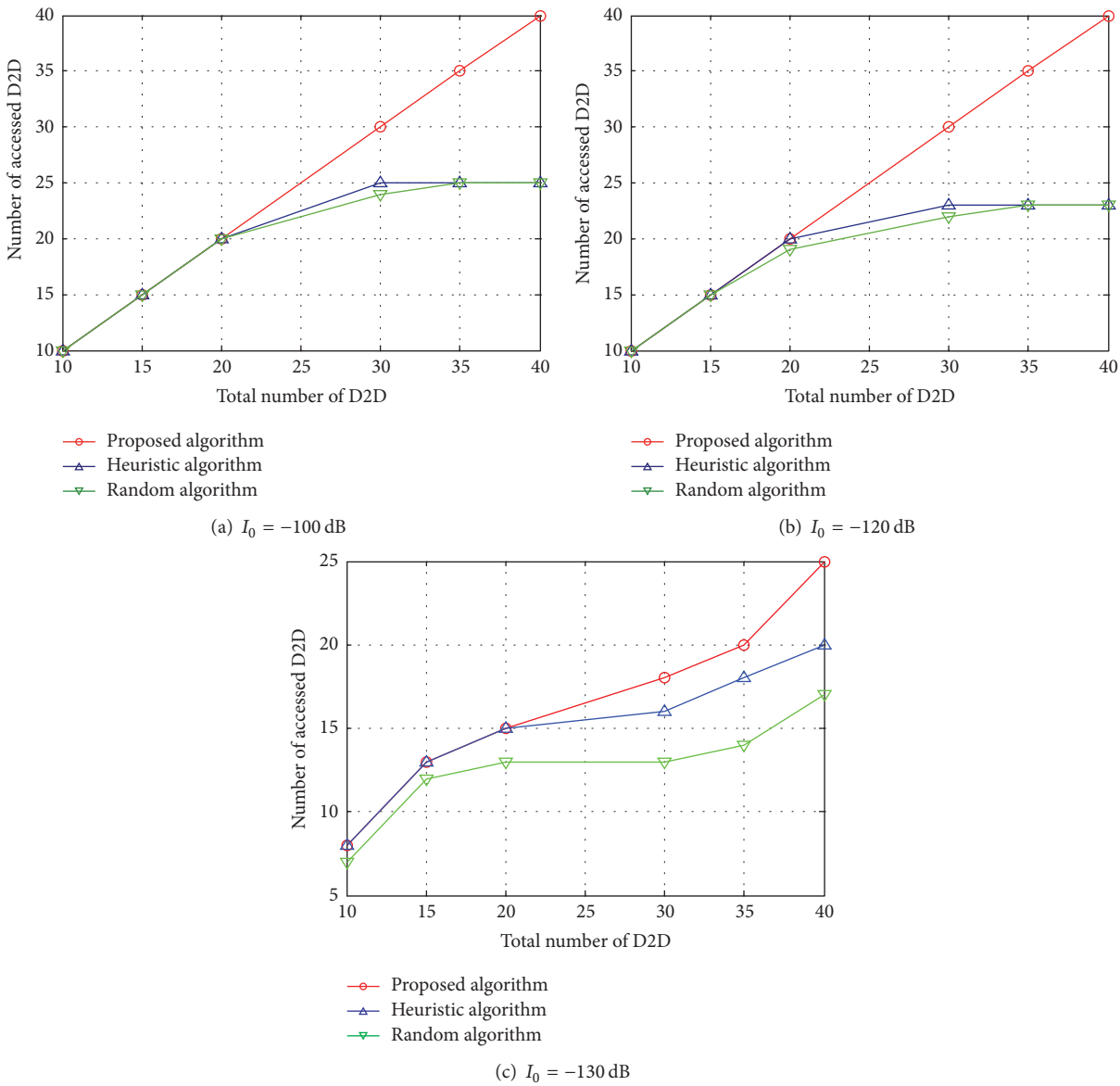


Figure 5: The number of accessed D2D pairs.

see that the maximum number of accessed D2D pairs is 25 which is equal to cellular UE. The reason is that cellular UE can only share resource with a D2D pair in the other two algorithms. At the same time, it can be observed that the number of accessed D2D pairs is decreasing with the interference threshold value of eNodeB becoming smaller. When many D2D pairs belong to the first group and they can only use exclusive limited resources, as a result, some D2D pairs have no resource to use and thus cannot access the system.

The average transmit power of D2D transmitter is presented in Figure 6. It shows that the proposed resource sharing method has a better performance in the aspect of transmit power over the random algorithm and heuristic algorithm. The reason is as follows. In the random algorithm, a D2D pair shares resource with random cellular UE, while cellular UE of priority shares resource with D2D pair which causes the least interference to the eNodeB in the heuristic algorithm. However, in the proposed algorithm, cellular UE devices prefer to share resource with the D2D pairs of smaller transmit power.

At the end, we have compared the proposed method with the methods proposed in [11], as shown in Figures 7 and 8. As for literatures [16–18], the proposed methods indeed achieve relatively better performance and lay the foundation in the area of the D2D resource sharing. However, it is shown that the CA algorithm in [11] can achieve a relatively higher sum rate compared with our proposed method, while the average transmit power is rather higher than the proposed method in our paper. Hence, our proposed method can be regarded as a

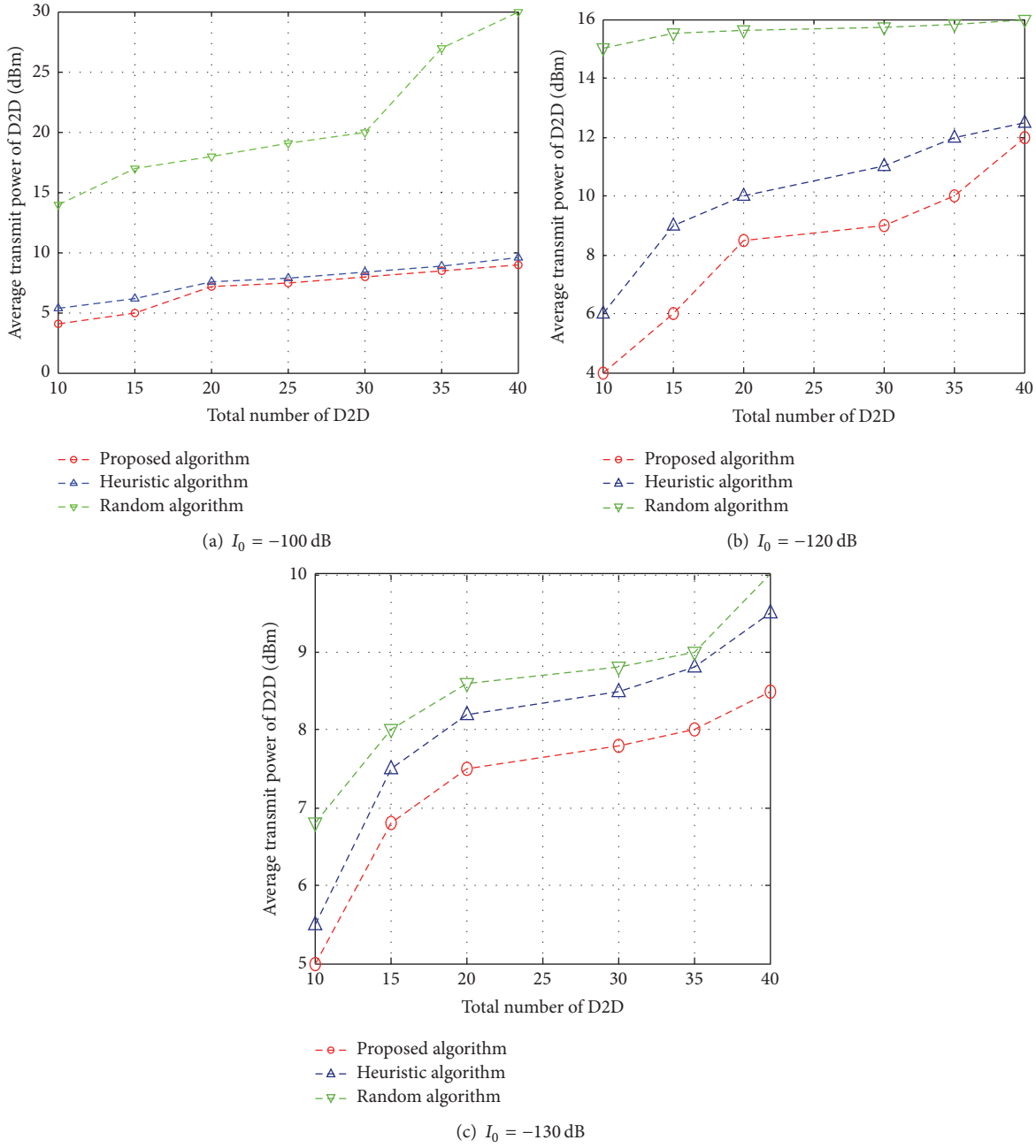


Figure 6: Transmit power of D2D transmitter.

kind of sum rate-power tradeoff scheme and can be used as an alternative in reality.

5. Conclusion

In this paper, a mechanism where multiple D2D links share resource with a cellular link in cellular network is proposed. Firstly, a minimum transmit power matrix and a maximum transmit power matrix are constructed, and D2D UE devices are divided into three groups by comparing each element of matrix. We construct a minimum transmit power matrix according to D2D pairs in the second group. By circular searching of the minimum value in matrix and comparing it with the corresponding maximum transmit power, the minimum transmit power values are obtained. Then, we adjust the transmit power of D2D transmitters that belong to the same column, and the D2D pairs share resource with cellular UE by the subscript of the selected minimum transmit power. In order to minimize the transmit power of D2D system, D2D UE devices in the first group and third group share resource with cellular UE based on the transmit power minimization principle. Finally, the simulation results

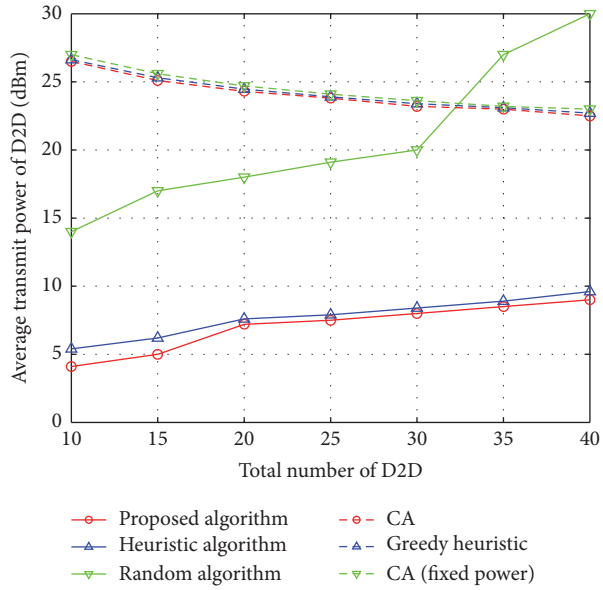


Figure 7: Average transmit power with the number of D2D for different methods.

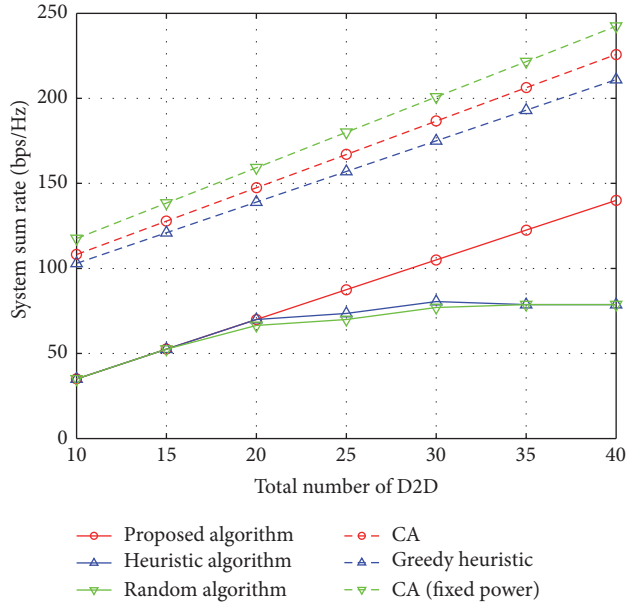


Figure 8: System sum rate with the number of D2D for different methods.

show that the proposed policy can achieve a relatively higher network capacity and lower transmit power of D2D system.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

This work was partially supported by the National 863 Program under Grant 2015AA01A705 and China Scholarship Council. The work of H. Li was supported by the National Science Foundation under Grants ECCS-1407679, CNS-1525226, CNS-1525418, and CNS-1543830.

References

[1] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hug, "Device-to-device communication as an underlay to LTE-advanced networks," *IEEE Communications Magazine*, vol. 47, no. 12, pp. 42–49, 2009.

[2] G. Fodor, E. Dahlman, G. Mildh et al., "Design aspects of network assisted device-to-device communications," *IEEE Communications Magazine*, vol. 50, no. 3, pp. 170–177, 2012.

[3] 3GPP, "Feasibility study for proximity services (ProSe) (release 12)," TR 22.803, 2015.

[4] C.-H. Yu, K. Doppler, C. B. Ribeiro, and O. Tirkkonen, "Resource sharing optimization for device-to-device communication underlaying cellular networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2752–2763, 2011.

[5] J. Gu, S. J. Bae, B.-G. Choi, and M. Y. Chung, "Dynamic power control mechanism for interference coordination of device-to-device communication in cellular networks," in *Proceedings of the 3rd International Conference on Ubiquitous and Future Networks (ICUFN '11)*, pp. 71–75, IEEE, Dalian, China, June 2011.

[6] C.-H. Yu, O. Tirkkonen, K. Doppler, and C. Ribeiro, "On the performance of device-to-device underlay communication with simple power control," in *Proceedings of the IEEE 69th Vehicular Technology Conference (VTC Spring '09)*, pp. 1–5, April 2009.

[7] C.-H. Yu, O. Tirkkonen, K. Doppler, and C. Ribeiro, "Power optimization of device-to-device communication underlaying cellular communication," in *Proceedings of the IEEE International Conference on Communications (ICC '09)*, pp. 1–5, IEEE, Dresden, Germany, June 2009.

[8] P. Jänis, V. Koivunen, Ć. Ribeiro, J. Korhonen, K. Doppler, and K. Hugl, "Interference-aware resource allocation for device-to-device radio underlaying cellular networks," in *Proceedings of the IEEE 69th Vehicular Technology Conference (VTC '09)*, pp. 1–5, IEEE, Barcelona, Spain, April 2009.

[9] T. Peng, Q. Lu, and H. Wang, "Interference avoidance mechanisms in the hybrid cellular and device-to-device systems," in *Proceedings of the IEEE 20th International Symposium on PIMRC*, pp. 617–621, 2009.

[10] M. Zulhasnine, C. Huang, and A. Srinivasan, "Efficient resource allocation for device-to-device communication underlaying LTE network," in *Proceedings of the IEEE 6th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob '10)*, pp. 368–375, IEEE, Ontario, Canada, October 2010.

[11] F. Wang, C. Xu, L. Song, and Z. Han, "Energy-efficient resource allocation for device-to-device underlay communication," *IEEE Transactions on Wireless Communications*, vol. 14, no. 4, pp. 2082–2092, 2015.

[12] X. Lin, J. G. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3GPP device-to-device proximity services," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 40–48, 2014.

[13] H. Min, J. Lee, S. Park, and D. Hong, "Capacity enhancement using an interference limited area for device-to-device uplink underlaying cellular networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 12, pp. 3995–4000, 2011.

[14] M. Belleschi, G. Fodor, D. D. Penda, A. Pradini, M. Johansson, and A. Abrardo, "Benchmarking Practical RRM Algorithms for D2D Communications in LTE Advanced," *Wireless Personal Communications*, vol. 82, no. 2, pp. 883–910, 2015.

[15] M. G. S. Rêgo, T. F. Maciel, H. H. M. Barros, F. R. P. Cavalcanti, and G. Fodor, "Performance analysis of power control for device-to-device communication in cellular MIMO systems," in *Proceedings of the 2nd International Workshop on Self Organizing Networks (IWSoN '12)*, August 2012.

[16] R. Zhang, X. Cheng, L. Yang, and B. Jiao, "Interference-aware graph based resource sharing for device-to-device communications underlaying cellular networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '13)*, pp. 140–145, April 2013.

[17] P. Phunchongharn, E. Hossain, and D. Kim, "Resource allocation for device-to-device communications underlaying LTE-advanced networks," *IEEE Wireless Communications*, vol. 20, no. 4, pp. 91–100, 2013.

[18] J. Wang, D. Zhu, C. Zhao, J. C. F. Li, and M. Lei, "Resource sharing of underlaying device-to-device and uplink cellular communications," *IEEE Communications Letters*, vol. 17, no. 6, pp. 1148–1151, 2013.

[19] D. H. Lee, K. W. Choi, W. S. Jeon, and D. G. Jeong, "Two-stage semi-distributed resource management for device-to-device communication in cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 1908–1920, 2014.

[20] 3GPP, "Physical layer aspects for evolved universal terrestrial radio access (UTRA) (release 7)," TS 25.814, 2006.

[21] B. Wang, L. Chen, X. Chen, X. Zhang, and D. Yang, "Resource allocation optimization for device-to-device communication underlaying cellular networks," in *Proceedings of the IEEE 73rd Vehicular Technology Conference (VTC '11)*, pp. 1–6, IEEE, Budapest, Hungary, May 2011.

[22] Y. Xu, R. Yin, T. Han, and G. Yu, "Dynamic resource allocation for Device-to-Device communication underlaying cellular networks," *International Journal of Communication Systems*, vol. 27, no. 10, pp. 2408–2425, 2014.

Hindawi Publishing Corporation
Mobile Information Systems
Volume 2016, Article ID 2185714, 15 pages
http://dx.doi.org/10.1155/2016/2185714

Research Article

Convolution Model of a Queueing System with the cFIFO Service Discipline

Sławomir Hanczewski, Adam Kaliszan, and Maciej Stasiak

Faculty of Electronics and Telecommunications, Poznan University of Technology, Ul. Polanka 3, 60-965 Poznan, Poland

Correspondence should be addressed to Sławomir Hanczewski; slawomir.hanczewski@put.poznan.pl

Received 29 July 2016; Accepted 20 October 2016

Academic Editor: Ioannis Moscholios

Copyright © 2016 Sławomir Hanczewski et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article presents an approximate convolution model of a multiservice queueing system with the continuous FIFO (cFIFO) service discipline. The model makes it possible to service calls sequentially with variable bit rate, determined by unoccupied (free) resources of the multiservice server. As compared to the FIFO discipline, the cFIFO queue utilizes the resources of a multiservice server more effectively. The assumption in the model is that the queueing system is offered a mixture of independent multiservice Bernoulli-Poisson-Pascal (BPP) call streams. The article also discusses the results of modelling a number of queueing systems to which different, non-Poissonian, call streams are offered. To verify the accuracy of the model, the results of the analytical calculations are compared with the results of simulation experiments for a number of selected queueing systems. The study has confirmed the accuracy of all adopted theoretical assumptions for the proposed analytical model.

1. Introduction

In recent years there has been a rapid increase in development of networks, in particular of mobile networks. While the evidence indicates that increasing competition is bringing down the cost of mobile services and user equipment, this implicates that more and more of the percentage of total network traffic is generated by mobile devices. According to the report [1], there were 7.3 billion active mobile devices in 2015, while the identified global trends are reporting much increased smartphone-generated traffic with each smartphone generating 1.4 Gb per month on average. It is also worth noticing that the ever-increasing number of users is using wireless broadband access networks. The authors of the report estimate that the number of active devices operating on LTE networks will reach 4.3 billion in 2021.

The dynamic development of telecommunications (mobile) networks, the growing number of offered online services with strictly defined Quality of Service (QoS) parameters, and the ever-increasing number of network users cause network operators to introduce a number of different traffic management mechanisms that increase the effectiveness of the network. Good examples of the above mechanisms are threshold compression mechanisms [2–4] and thresholdless compression mechanism that allow elastic and adaptive traffic to be supported [5, 6]. In the case of systems with real time services it is necessary to ensure an appropriate delay level and/or deviation from true periodicity of a presumed periodic signal (*jitter*). As a result, analyses of these systems are not feasible unless appropriate models of queueing systems that would take into account the multiservice nature of traffic are constructed. A great number of works, including [7–9], discuss a solution that is based on a single-service Erlang C model with the accompanying assumption that all call classes have similar characteristics and service conditions. Another solution proposes an application of a recurrent formula that describes the occupancy distribution in full-availability multiservice systems with elastic traffic [5]. The model presented in [6] has an accompanying assumption that all streams of serviced traffic—if there is an instance of a lack of free resources in the system—can be compressed without limit, which eventually leads to lossless traffic service. Here, the level of compression can be interpreted as a measure of delay for all call classes. These solutions, however, do not allow us to determine delay parameters for individual call classes. In [10–13] new

queueing models are proposed that make it possible to determine individual delay characteristics for particular call classes offered to the system with a queueing discipline called state-dependent FIFO (SD-FIFO). This discipline ensures access to a multiservice server to be available for all classes of calls, while the division of the resources of a server between individual call classes depends on the number of all calls that are currently in the queueing system and corresponds to the balanced fairness algorithm [14, 15] for the division of resources in multiservice systems.

This article proposes an approximate model of a queueing system with continuous FIFO (cFIFO) service discipline for a mixture of multiservice Bernoulli-Poisson-Pascal (BPP) call streams. In general, the cFIFO discipline assumes that calls that are in the queue are serviced according to the FIFO discipline. However, when the server has a lower amount of resources than demanded by the first call in the queue, then it can start servicing this call with a lower bitrate than the one demanded by the call. In the proposed model, a convolution algorithm is used to determine the occupancy distribution in the queueing system. Appropriate convolution algorithms to model full-availability multiservice systems with losses and multiservice access network systems (the so-called multiservice tree network) are proposed in [16–18]. Papers [19, 20] propose convolution algorithms to model multiservice full-availability systems with resource reservation. Paper [21] develops and discusses a two-dimensional convolution model for a multiservice overflow system. Yet another work [22] proposes a generalized version of a convolution algorithm that makes it possible to model different multiservice state-dependent systems. The advantage of convolution algorithms is that they provide a possibility to model approximately systems with call streams with different distributions, provided they are not mutually dependent. Hence, in order to present the capabilities of the proposed model, this article also presents the results of a comparison of the analytical model with the data obtained by the digital simulation experiments for traffic streams other than BPP call streams. It is worthwhile to stress that convolution algorithms have never been used in the analysis of queueing systems before.

The article is organized as follows. Section 2 provides a description of a multiservice queueing system with the SD-FIFO discipline. Section 3 presents a description of a multiservice queueing system with the continuous discipline cFIFO. Section 4 proposes an analytical convolution model of the considered queueing system. A number of exemplary results of a comparison of the analytical model with the results of a simulation are provided and discussed in Section 5. The last section, Section 6, is a short summary of the article.

2. Queueing System with the SD-FIFO Discipline

The multiservice queueing system with the SD-FIFO queue service discipline is described in [10, 11, 13]. The analytical model developed for this system allows the most important parameters that characterize the queueing system (such as the average time for a call to be in the queue or the average queue length) to be determined for each call class offered to the system. An exemplary SD-FIFO system is shown in Figure 1. The system is composed of a multiservice server with the capacity V allocation units (AUs) (allocation unit is the capacity unit for broadband systems [23, 24]. Typically, it is defined as the greatest common divisor of the maximum bitrates of individual calls [25] or equivalent bandwidths [26] determined for offered call streams. A method for a determination of allocation units in telecommunications systems is provided in the Appendix) and a buffer with the capacity U AUs. The notion of the multiservice server is understood to be a server with the capacity that enables concurrent service to a number of calls with different demanded bitrates to be effected. The system (Figure 1) is offered 3 call classes. In the considered SD-FIFO queueing system, virtual queues for each call class serviced by the system are created within one buffer. Depending on the number of calls that are currently in the system, appropriate resources of the server are allocated to individual classes. The amount of these resources varies every time the number of calls of individual classes that are currently in the system is changed (resulting by a termination of service or admittance of a new call). According to this discipline, in each occupancy state of the system, appropriate resources of the service are allocated to calls of all classes. It should be stressed that the method for resource allocation in the server for individual call classes is compatible with the balanced fairness algorithm, well known from the literature of the subject [14, 15].

The system is composed of a multirate server with the capacity V AUs and a buffer with the capacity U AUs. The system is offered a set of $\mathbf{M}$ traffic classes of the type T ($T \in \{\mathrm{Er}, \mathrm{En}, \mathrm{Pa}, G\}$), where

(i) Er denotes the Erlang traffic (Poisson call stream),

(ii) En denotes the Engset traffic (Bernoulli call stream),

(iii) Pa denotes the Pascal traffic (Pascal call stream),

(iv) G denotes a type of traffic, with non-Poissonian call stream whose properties have been determined either empirically or in a simulation.

The cardinality of the set $\mathbf{M}$, that is, the number of traffic classes offered to the system, is equal to m ($|\mathbf{M}| = m$). Any randomly chosen traffic class c demands $t_{T,c}$ AUs for service. This notation method will be also used in the remaining part of the article to identify the parameters related to the considered call classes. We notice that the mixture of Erlang, Engset, and Pascal traffics is denoted as BPP traffic.

Let us consider now a queueing system with the SD-FIFO discipline composed of a multiservice server with the capacity V AUs and a buffer with the capacity U AUs. The system is offered a set of $\mathbf{M}$ traffic classes of the Erlang-type call classes. Individual call classes are described by the following parameters:

(i) $\lambda_{\mathrm{Er},i}$ is the intensity of a call stream of class i ($0 < i \leq m$),

(ii) $\mu_{\mathrm{Er},i}$ is the intensity of a service stream of class i,

(iii) $t_{\mathrm{Er},i}$ is the number of AUs demanded by a call of class i,

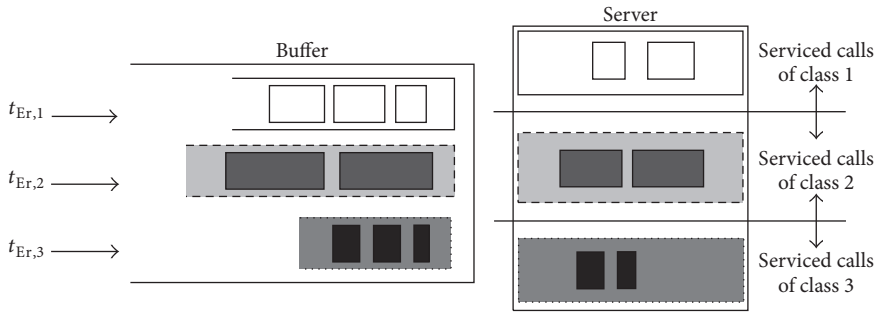


Figure 1: Queueing system with the SD-FIFO service discipline.

(iv) $a_{\mathrm{Er},i}$ is the traffic intensity of class i:

$$a_{\mathrm{Er},i} = \frac{\lambda_{\mathrm{Er},i}}{\mu_{\mathrm{Er},i}}. \tag{1}$$

The occupancy distribution in the system with the SD-FIFO discipline is determined on the basis of the following recurrence [11]:

$$[Q(n)]^{\mathbf{M}}_{V,U} = \begin{cases} \dfrac{1}{\min(n,V)} \displaystyle\sum_{i=1}^{m} a_{\mathrm{Er},i} t_{\mathrm{Er},i} \left[Q\left(n - t_{\mathrm{Er},i}\right)\right]^{\mathbf{M}}_{V,U} & \text{for } 0 \le n \le V + U, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $[Q(n)]^{\mathbf{M}}_{V,U}$ is the probability that there are n occupied AUs in the system with server capacity V AUs and buffer capacity U AUs. To the system a set of $\mathbf{M}$ traffic classes is offered.

It is possible to determine on the basis of the distribution $[Q(n)]^{\mathbf{M}}_{V,U}$ the parameter $[y_{\mathrm{Er},i}(n)]_{V,U}$, that is, the number of calls of class i serviced in the server in state n:

$$\left[y_{\mathrm{Er},i}(n)\right]_{V,U} = \frac{a_{\mathrm{Er},i} t_{\mathrm{Er},i} \left[Q\left(n - t_{\mathrm{Er},i}\right)\right]^{\mathbf{M}}_{V,U}}{[Q(n)]^{\mathbf{M}}_{V,U}}. \tag{3}$$

The $[y_{\mathrm{Er},i}(n)]_{V,U}$ parameter determines then the resources occupied in the server by a given traffic class and is dependent on the occupancy distribution $[Q(n)]^{\mathbf{M}}_{V,U}$. This means that the resources occupied in the server by a given call class are dependent on the occupancy state of resources n by all traffic classes in the system (serviced in the server and those in the queue).

Thus, Formula (3) defines the queueing service discipline in the system determined by distribution (2), that is, that from a virtual queue of class i; in a given occupancy state of system n, this number of calls of this class is taken for service that satisfies (3). Because of these dependencies, this discipline is labeled SD-FIFO [11].

On the basis of distribution (2) one can determine the appropriate queueing characteristics for the system under consideration [11], for example, average waiting time, average queue length, or blocking probability.

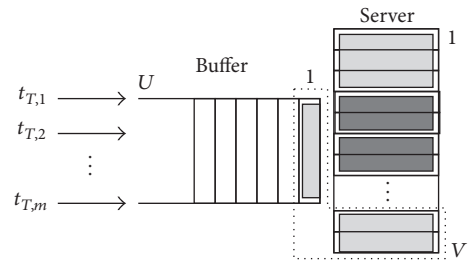


Figure 2: Queueing system with cFIFO queueing discipline.

3. Queueing System with Continuous cFIFO Discipline

Let us consider the multiservice queueing system presented in Figure 2. Calls that are in the queue are serviced according to the proposed cFIFO discipline. In the case when the multiservice server has a lower throughput than the bitrate required by the first call that is in the queue to be serviced, then the server can start servicing this call with a lower bitrate than the demanded $t_{T,c}$ AUs. Figure 2 shows a call, marked with the hyphenated line, that demands 3 AUs. The server only has two free AUs; hence—in line with the cFIFO service discipline—2 AUs of the considered call will be serviced in the server, while one AU will be waiting for service in the buffer. Only one call in the system can be serviced with a lower number of AUs than that demanded for service. Such a concept ensures the maximum usage of the resources of the server providing at the same time a simple algorithm for buffer service.

Let us consider now the operation of a queueing system with the system with the parameters $V = 2$ AUs and $U = 2$ AUs, presented in Figure 3(a), as an example. This system can be analyzed either at the microstate level (Figure 3(b)) or at the macrostate level (Figure 3(c)). Microstate of the multiservice service process is defined by the number of calls of individual classes serviced in the server and waiting in the buffer [11]. Macrostate is then defined by the total number of AUs that are occupied by calls that are in the system, that is, those that are being serviced in the server and those waiting in the buffer [11]. The macrostate does not take into consideration the distribution of occupied AUs between individual classes of calls. The system presented in Figure 3(a)

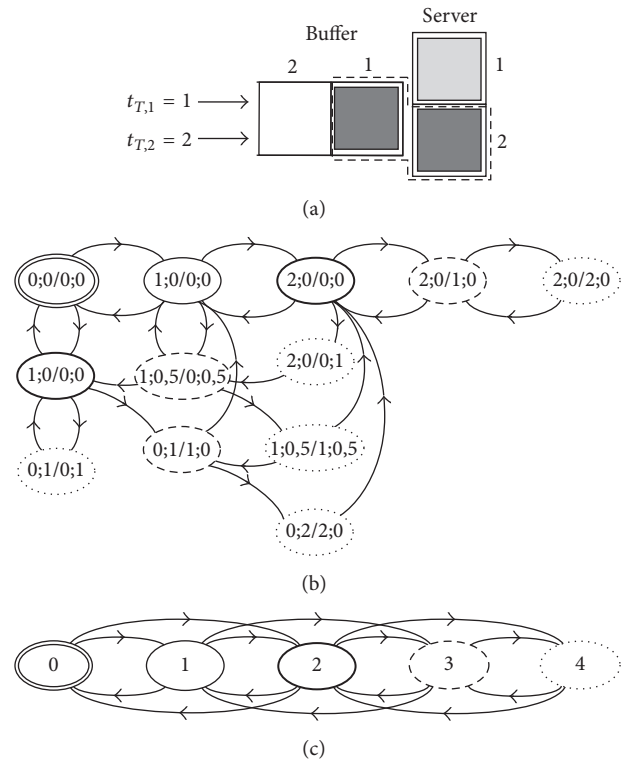


Figure 3: Queueing system with cFIFO discipline ($V = 2$ AUs, $U = 2$ AUs): (a) visualization of macrostate $\{1; 0.5/0; 0.5\}$, (b) service process, microstate level, and (c) service process, macrostate level.

is offered two call classes that demand $t_{T,1} = 1$ and $t_{T,2} = 2$ AUs, respectively. Each microstate of the process is represented by the ordered set $X/Z = \{x_{T,1}; x_{T,2}/z_{T,1}; z_{T,2}\}$, where the values $x_{T,1}$ and $x_{T,2}$ define the number of calls of classes 1 and 2 that are serviced in the server and the values $z_{T,1}$ and $z_{T,2}$ define then the number of calls of classes 1 and 2 that are waiting in the buffer. In Figure 3(b), the microstate $\{1; 0.5/0; 0.5\}$ determines such an occupancy state of the considered queueing system in which a call of class 2 is serviced in the server with a lower number of AUs than that demanded; that is, one AU ("0.5 of the call") is serviced in the server and one AU ("0.5 of the call") is waiting in the buffer. Let us consider two possible ways of service termination for a call of class 2 in the microstate under consideration. In the first case, after the termination of service for a call of class 1, the server starts to service a call of class 2 with the demanded number of $t_{T,2} = 2$ AUs (in Figure 3(b) it is the transition from microstate $\{1; 0.5/0; 0.5\}$ to microstate $\{0; 1/0; 0\}$). In the second case, because of long service time for the call of class 1, the considered call of class 2 is serviced with $t_{T,2} = 1$ AU throughout the service (in Figure 3(b) it is the transition from microstate $\{1; 0.5/0; 0.5\}$ to microstate $\{1; 0/0; 0\}$).

4. Model of a Queueing System with cFIFO Discipline

The model of a queueing system proposed in the article has been developed at the macrostate level. This means that the occupancy state is described by just one parameter, namely, the total number of occupied AU in the system (i.e., together in the server and buffer). Each macrostate $\Omega(n)$ is the sum of such microstates X/Z that satisfies the following condition:

$$\Omega(n) = \left\{ \frac{X}{Z} : \sum_{c=1}^{m} x_{T,c} t_{T,c} + \sum_{c=1}^{m} z_{T,c} t_{T,c} = n \right\}. \quad (4)$$

A description of a system at the macrostate level greatly simplifies its analysis in that it limits the analysis to a lower number of states. In the system presented in Figure 3(a), and defined by the parameters $V = 2$ AUs, $U = 2$ AUs, the number of microstates is equal to 12 (Figure 3(b)), whereas the number of macrostates is equal to 5 (Figure 3(c)). In Figure 3(b), the contours of the microstates that belong to one macrostate are marked by identical line. For example, macrostate $\Omega(3)$ includes the following microstates: $\{2; 0/1; 0\}$, $\{1; 0.5/0; 0.5\}$, and $\{0; 1/1; 0\}$.

4.1. Convolution Algorithm. Multiservice telecommunications systems can be generally modelled by algorithms that analyze either the dependencies between microstates or macrostates. The one approach is characterized by a large computational complexity [27], whereas the other approach is effective in modelling systems that are offered Poisson call streams. Papers [16, 17] propose convolution algorithms that make it possible to model systems with multiservice traffic on the basis of mutually independent occupancy distributions. These distributions are determined for all call streams offered to the system and can be determined on the basis of

appropriate models and theoretical single-service systems. They can also directly result from conducted measurements in real systems or from simulation experiments for a number of selected classes of calls.

Let us consider now the operation of a convolution algorithm with a multiservice system that is composed of a server with the capacity V and a buffer with zero capacity ($U = 0$) as an example [28, 29]. In the literature of the subject this system is frequently labelled as the full-availability group (FAG) [30]. Our assumption is that the system is offered a set of $\mathbf{M}$ ($|\mathbf{M}| = m$) traffic classes of the type T ($T \in \{\mathrm{Er}, \mathrm{En}, \mathrm{Pa}, G\}$) and any randomly chosen traffic class c demands $t_{T,c}$ AUs for service.

The input data for the convolution algorithm are the occupancy distributions for single classes $[p]_{V,0}^{\{T,c\}}$, where $c \in \mathbf{M}$. In the adopted notation for the distribution $[p]_{V,0}^{\{T,c\}}$ the first expression in the lower index V defines the capacity of the server, while the second element U defines the capacity of the buffer (in the system under investigation, the buffer with zero capacity $U = 0$ is considered). The distributions $[p]_{V,0}^{\{T,c\}}$ are determined with the assumption that the system with the capacity V is offered only one class of calls. For Erlang, Engset, or Pascal traffic classes, to determine this distribution one can use appropriate theoretical distributions [31]; for example, in the case of class i of Erlang traffic, the distribution for a single class can be determined on the basis of the Erlang distribution:

$$[p(x)]_{\nu_{\mathrm{Er},i}}^{\mathrm{Er},i} = \frac{(a_{\mathrm{Er},i})^x / x!}{\sum_{l=0}^{\nu_{\mathrm{Er},i}} (a_{\mathrm{Er},i})^l / l!}, \tag{5}$$

where $x \in (0 \le x \le \nu_{\mathrm{Er},i})$ and $a_{\mathrm{Er},i}$ is the traffic intensity of class i, whereas the parameter $\nu_{\mathrm{Er},i}$ denotes the maximum number of calls that can be serviced in the server with the capacity V:

$$\nu_{\mathrm{Er},i} = \left\lfloor \frac{V}{t_{\mathrm{Er},i}} \right\rfloor. \tag{6}$$

In the case of Engset and Pascal traffic, the distributions of single classes can be determined on the basis of the following formulas:

$$[p(x)]_{\nu_{\mathrm{En},j}}^{\mathrm{En},j} = \frac{\binom{S_{\mathrm{En},j}}{x} (\gamma_{\mathrm{En},j})^x}{\sum_{l=0}^{\nu_j} \binom{S_{\mathrm{En},j}}{l} (\gamma_{\mathrm{En},j})^l}, \tag{7}$$

$$[p(x)]_{\nu_{\mathrm{Pa},k}}^{\mathrm{Pa},k} = \frac{\binom{-S_{\mathrm{Pa},k}}{x} (-\gamma_{\mathrm{Pa},k})^x}{\sum_{l=0}^{\nu_k} \binom{-S_{\mathrm{Pa},k}}{l} (-\gamma_{\mathrm{Pa},k})^l}, \tag{8}$$

where the parameters $\gamma_{\mathrm{En},j}$ and $\gamma_{\mathrm{Pa},k}$ define the average intensity of calls offered by a single free Engset and Pascal traffic source. The parameters $S_{\mathrm{En},j}$ and $S_{\mathrm{Pa},k}$ are the total number of Engset and Pascal traffic sources. The relationship between the average traffic offered to the system by Engset traffic stream $a_{\mathrm{En},j}$ and Pascal traffic stream $a_{\mathrm{Pa},k}$ and the parameters $\gamma_{\mathrm{En},j}$ and $\gamma_{\mathrm{Pa},k}$ is as follows:

$$a_{\mathrm{En},j} = S_{\mathrm{En},j} \frac{\gamma_{\mathrm{En},j}}{1 + \gamma_{\mathrm{En},j}}, \tag{9}$$

$$a_{\mathrm{Pa},k} = -S_{\mathrm{Pa},k} \frac{(-\gamma_{\mathrm{Pa},k})}{1 + (-\gamma_{\mathrm{Pa},k})}. \tag{10}$$

The occupancy distribution $[p]_{V,0}^{\{T,c\}} = \{[p(0)]_{V,0}^{\{T,c\}}, [p(1)]_{V,0}^{\{T,c\}}, [p(2)]_{V,0}^{\{T,c\}}, \ldots, [p(n)]_{V,0}^{\{T,c\}}, \ldots, [p(V)]_{V,0}^{\{T,c\}}\}$ of a single class c, determining the occupancy probabilities for all macrostates n (such that $n = x \cdot t_{T,c}$ AUs) in the server with the capacity V, is related to distribution (5) by the following dependence:

$$[p(n)]_{V,0}^{\{T,c\}} = \begin{cases} [p(x)]_{\nu_{T,c}}^{\{T,c\}} & \text{for } n = x \cdot t_{T,c},\ 0 \le x \le \nu_{T,c} \\ 0 & \text{for remaining } n. \end{cases} \tag{11}$$

In the case of other traffic streams (of the type G), distributions of single classes can be determined on the basis of measurements taken in real systems or on the basis of relevant simulation experiments.

By having the single distributions for all classes offered to the system $[p]_{V,0}^{\{T,c\}}$ at our disposal we can now determine the aggregated distribution $[P]_{\bullet,0}^{\mathbf{M}}$ which is the result of the convolution operation of all distributions for single classes offered to the system:

$$[P]_{\bullet,0}^{\mathbf{M}} = [p]_{V,0}^{\{T,1\}} * [p]_{V,0}^{\{T,2\}} * \cdots * [p]_{V,0}^{\{T,m\}}. \tag{12}$$

The distribution $[P]_{\bullet,0}^{\mathbf{M}}$ is not a normalized distribution. After an appropriate normalization process, the normalized distribution $[P]_{V,0}^{\mathbf{M}}$ can be written in the following form:

$$[P]_{V,0}^{\mathbf{M}} = \left\{k[P(0)]_{\bullet,0}^{\mathbf{M}}, k[P(1)]_{\bullet,0}^{\mathbf{M}}, \ldots, k[P(V)]_{\bullet,0}^{\mathbf{M}}\right\}, \tag{13}$$

where k is the normalization constant:

$$k = \frac{1}{\sum_{n=0}^{V} [P(n)]_{\bullet,0}^{\mathbf{M}}}. \tag{14}$$

Let us note that, for example, as a result of the convolution operation of two normalized distributions with the length V, a distribution with the length $2V$ is obtained. From the point of view of the considered system with the capacity of V AUs, all states $n > V$ will never occur. Therefore, such states, in the distribution with the length $2V$ must be removed and distribution must be normalized. This process is discussed in detail in [16, 17, 19], among others.

In (12), the symbol $*$ denotes the convolution operation which is defined in the following way:

$$[P]_{\bullet,0}^{\mathbf{A}\cup\mathbf{B}} = [P]_{V,0}^{\mathbf{A}} * [P]_{V,0}^{\mathbf{B}} = \Big\{ [P(0)]_{V,0}^{\mathbf{A}} [P(0)]_{V,0}^{\mathbf{B}}, [P(0)]_{V,0}^{\mathbf{A}} [P(1)]_{V,0}^{\mathbf{B}} + [P(1)]_{V,0}^{\mathbf{A}} [P(0)]_{V,0}^{\mathbf{B}}, \ldots, \sum_{l=0}^{n} [P(l)]_{V,0}^{\mathbf{A}} [P(n-l)]_{V,0}^{\mathbf{B}}, \ldots, \sum_{l=0}^{V} [P(l)]_{V,0}^{\mathbf{A}} [P(V-l)]_{V,0}^{\mathbf{B}} \Big\}, \tag{15}$$

where the symbol $[P]_{V,0}^{\mathbf{A}}$ denotes the normalized distribution of a single class or the aggregated distribution of a number of classes that belong to a given set $\mathbf{A}$ ($\mathbf{A} \in \mathbf{M}$), whereas the symbol $[P]_{V,0}^{\mathbf{B}}$ denotes the normalized distribution of a single class or the aggregated normalized distribution of a number of classes that belong to a given set $\mathbf{B}$ ($\mathbf{B} \in \mathbf{M}$ and $\mathbf{A} \cap \mathbf{B} = 0$). The symbol $[P]_{\bullet,0}^{\mathbf{A}\cup\mathbf{B}}$ defines the aggregated nonnormalized distribution that is the result of a convolution of the distributions from the sets $\mathbf{A}$ and $\mathbf{B}$.

The convolution operation (15) enables us to determine the average number $[y_{T,c}(n)]_{V,0}$ of serviced calls of class c in the server with the capacity V AUs with zero buffer that is in the occupancy state n AUs. A determination of this parameter should take into consideration the following reasoning [30]. First, the aggregated nonnormalized distributions of all classes, except class c, are to be determined:

$$[P]_{\bullet,0}^{\mathbf{M}\setminus\{\mathbf{c}\}} = [p]_{V,0}^{\{1\}} * \cdots * [p]_{V,0}^{\{c-1\}} * [p]_{V,0}^{\{c+1\}} * \cdots * [p]_{V,0}^{\{m\}}. \tag{16}$$

Then, on the basis of the distributions $[P]_{\bullet,0}^{\mathbf{M}\setminus\{\mathbf{c}\}}$ and $[p]_{V,0}^{\{c\}}$ the value of the parameter $[y_{\mathrm{Er},c}(n)]_{V,0}$ can be determined, that is, the average number of calls of class c in occupancy state n of the server:

$$[y_{T,c}(n)]_{V,0} = \frac{\sum_{l=0}^{n} l\,[p(l)]_{V,0}^{\{c\}} [P(n-l)]_{\bullet,0}^{\mathbf{M}\setminus\{\mathbf{c}\}}}{t_{T,c} \sum_{l=0}^{n} [p(l)]_{V,0}^{\{c\}} [P(n-l)]_{\bullet,0}^{\mathbf{M}\setminus\{\mathbf{c}\}}}. \tag{17}$$

4.2. Model of the cFIFO Queueing System for Poisson Traffic Streams. Let us consider now the queueing system with the cFIFO service discipline the operation of which is presented in Section 3. Our assumption is that this system is offered m Erlang traffic streams for which call arrival intensities of new calls are described by Poisson distributions with the parameters $\lambda_{\mathrm{Er},1}, \ldots, \lambda_{\mathrm{Er},i}, \ldots, \lambda_{\mathrm{Er},m}$ for each traffic class, respectively. Service time is described with exponential distributions with the parameters $\mu_{\mathrm{Er},1}, \ldots, \mu_{\mathrm{Er},i}, \ldots, \mu_{\mathrm{Er},m}$. Individual call classes demand, respectively, $t_{\mathrm{Er},1}, \ldots, t_{\mathrm{Er},i}, \ldots, t_{\mathrm{Er},m}$ AUs for service. In the case of Erlang traffic, it is assumed that the number of traffic sources is infinite. This means that the call arrival intensity for new call arrivals $\lambda_{\mathrm{Er},i}$ of a given class i is independent of the number of calls of this class that are currently in the system:

$$\lambda_{\mathrm{Er},i}(n) = \lambda_{\mathrm{Er},i} \quad \text{for } 0 \leqslant n \leqslant V + U. \tag{18}$$

The intensity of offered Erlang traffic of class i can be determined on the basis of Formula (1):

$$a_{\mathrm{Er},i} = a_{\mathrm{Er},i}(n) = \frac{\lambda_{\mathrm{Er},i}}{\mu_{\mathrm{Er},i}} \quad \text{for } 0 \leqslant n \leqslant V + U. \tag{19}$$

By expressing the call intensity for a given class i in allocation units per time unit (the product $\lambda_{\mathrm{Er},i} t_{\mathrm{Er},i}$), we can express traffic offered to the system in the following form:

$$A_{\mathrm{Er},i} = a_{\mathrm{Er},i} t_{\mathrm{Er},i}. \tag{20}$$

Paper [30] analyzes the dependence between the occupancy distribution in a multiservice full-availability system and the occupancy distribution in a multiservice state-dependent system, with the assumption of identical capacity and traffic offered to both systems. A similar approach will be now applied to the queueing model proposed in the article to determine the dependence between the occupancy distribution $[P(n)]_{V+U,0}^{\mathbf{M}}$ in the server with zero buffer and the distribution $[Q(n)]_{V,U}^{\mathbf{M}}$ in the considered queueing system:

$$[Q(n)]_{V,U}^{\mathbf{M}} = \delta(n)\,[P(n)]_{V+U,0}^{\mathbf{M}}, \tag{21}$$

where $\delta(n)$ defines the relation between the corresponding probabilities $[Q(n)]_{V,U}^{\mathbf{M}}$ and $[P(n)]_{V+U,0}^{\mathbf{M}}$, that is, the probability that the queueing system (V, U) is in state n AUs and the occupancy probability n AUs in the server with zero buffer $(V + U, 0)$, with the assumption that traffic offered to both systems is identical. To determine the parameter $\delta(n)$ we can use the recurrence dependencies derived for multiservice SD-FIFO queueing systems. The distribution $[Q(n)]_{V,U}^{\mathbf{M}}$ in the cFIFO system is then approximated by distribution (2) that, with (20) taken into consideration, will be rewritten in the following way:

$$[Q(n)]_{V,U}^{\mathbf{M}} = \begin{cases} \dfrac{1}{\min(n,V)} \displaystyle\sum_{i=1}^{m} A_{\mathrm{Er},i} [Q(n - t_{\mathrm{Er},i})]_{V,U}^{\mathbf{M}} & \text{for } 0 \leq n \leq V + U, \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

Assuming, then, that the method for a determination of the occupancy distribution in the queueing system and in the system with zero buffer (when the system is offered Erlang call classes) is known, it is possible to determine the value of the parameter $\delta(n)$ by dividing both sides of (22) by the probabilities $[P(n)]_{V+U,0}^{\mathbf{M}}$:

$$\frac{[Q(n)]_{V,U}^{\mathbf{M}}}{[P(n)]_{V+U,0}^{\mathbf{M}}} = \frac{1}{\min(n,V)} \sum_{i=1}^{m} \frac{[Q(n - t_{\mathrm{Er},i})]_{V,U}^{\mathbf{M}}}{[P(n)]_{V+U,0}^{\mathbf{M}}} A_{\mathrm{Er},i}. \tag{23}$$

Equation (23), with (21) taken into consideration, can be easily transformed into an iterative formula for the parameter $\delta(n)$:

$$\delta(n) = \frac{1}{\min(n,V)} \sum_{i=1}^{m} \delta(n - t_{\mathrm{Er},i}) \frac{[P(n - t_{\mathrm{Er},i})]^{\mathbf{M}}_{V+U,0} A_{\mathrm{Er},i}}{[P(n)]^{\mathbf{M}}_{V+U,0}}. \quad (24)$$

The Markov process in the server with zero buffer to which a mixture of Erlang traffic classes is offered is a reversible process [28, 29]. For the neighboring states $\Omega(n)$ and $\Omega(n - t_i)$, because of class i, it is possible then to write local balance equations:

$$A_{\mathrm{Er},i} [P(n - t_{\mathrm{Er},i})]^{\mathbf{M}}_{V+U,0} = t_{\mathrm{Er},i} [y_{\mathrm{Er},i}(n)]_{V+U,0} [P(n)]^{\mathbf{M}}_{V+U,0}. \quad (25)$$

Now, taking into consideration (25), (24) can be rewritten in the following form:

$$\delta(n) = \frac{\sum_{i=1}^{m} \delta(n - t_{\mathrm{Er},i}) [y_{\mathrm{Er},i}(n)]_{V+U,0} t_{\mathrm{Er},i}}{\min(n,V)}, \quad (26)$$

The convolution operation makes a determination of the occupancy distribution $[P]^{\mathbf{M}}_{V+U,0}$ and the average number of calls of individual classes $[y_{\mathrm{Er},i}(n)]_{V+U,0}$ serviced in the server with zero buffer (Formulas (16) and (17)) possible. These results enable us then to assess the value of the parameter $\delta(n)$ (Formula (26)) and eventually to evaluate the occupancy distribution in the queueing system $[Q(n)]^{\mathbf{M}}_{V,U}$ on the basis of (21).

4.3. Commentary. In Section 4.2, to find the general dependence $\delta(n)$ between the cFIFO queueing system described by the parameters (V,U) and the system with zero buffer $(V + U, 0)$ the occupancy distribution $[Q(n)]^{\mathbf{M}}_{V,U}$ in the SD-FIFO system is used. The possibility of an approximation of the cFIFO system by the SD-FIFO system results from significant similarities in both queue service disciplines involved. In the cFIFO discipline considered in the article only one call can be partly serviced in the server and partly waiting in the queue. The SD-FIFO discipline, in turn, allows even a number of calls to be partly serviced simultaneously. In both instances, however, resources of the server are used maximally and it is just this fact that determines the probability of occupancy distributions for systems with cFIFO and SD-FIFO queues. Figures 4 and 5 show the occupancy distribution $[Q(n)]_{V,U}$ in a queueing system with the structural parameters $V = 20$ AUs and $U = 10$ AUs, respectively. The occupancy distributions in the SD-FIFO system and the cFIFO system obtained in a simulation were then compared. The systems were offered three Erlang traffic streams with the following demands: $t_{\mathrm{Er},1} = 1$ AU, $t_{\mathrm{Er},2} = 2$ AUs, and $t_{\mathrm{Er},3} = 3$ AUs. Traffic was offered in the following proportions: $A_{\mathrm{Er},1} : A_{\mathrm{Er},2} : A_{\mathrm{Er},3} = 1 : 1 : 1$. Figure 4 shows the occupancy distributions for the case of small system loads ($a = 0.6$ Erl/AU) and Figure 5 for large system loads ($a = 1.2$ Erl/AU), where a is the average traffic offered per AU of the system:

$$a = \frac{A_{\mathrm{Er},1} + A_{\mathrm{Er},2} + A_{\mathrm{Er},3}}{V}. \quad (27)$$

The simulation results are shown with 95% confidence interval determined on the basis of the Student distribution for 5 series, 100,000 calls each. The results for the presented comparison indicate a very good convergence of occupancy distributions in the SD-FIFO and cFIFO queueing systems. The simulation experiments for other systems differentiated by their capacity, number, and demands of offered traffic classes conducted by the authors earlier confirm strong convergence of occupancy distributions of both queueing systems.

It should be stressed that queueing characteristics (such as the average number of calls of particular classes in the queue in corresponding occupancy states of the system) are not characterized by just as good convergence as the occupancy distributions. This would be the first reason why the SD-FIFO model cannot be directly applied to determine queueing characteristics for the cFIFO system. The other reason is the fact that to determine the parameter $\delta(n)$ (Formula (26)) the occupancy distribution of the SD-FIFO queueing system, derived for Erlang traffic, is used. In the case of other traffic streams, we do not observe such a good convergence between distributions for the SD-FIFO and cFIFO systems any longer. However, Formula (26) turns out to be very universal and versatile and approximates cFIFO queueing systems that service traffic with any, mutually independent, call streams very well. Appropriate numerical examples are presented in Section 5.

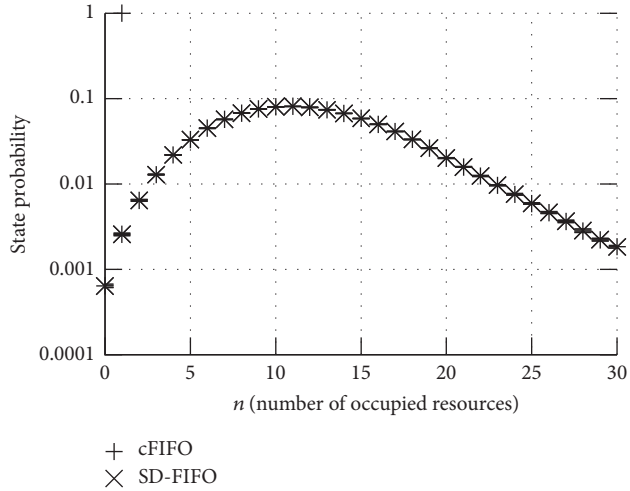
4.4. Characteristics of cFIFO System with Erlang Traffic. By having the knowledge of the coefficients $\delta(n)$ at hand we are in position to determine the occupancy distribution $[Q]^{\mathbf{M}}_{V,U}$ in a queueing system with the cFIFO discipline. This distribution can also provide a basis for a determination of appropriate QoS parameters, such as the blocking probability E_i and the average number of occupied AUs in queue L. The phenomenon of blocking for calls of class i in the queueing system occurs when the queue lacks $t_{\mathrm{Er},i}$ free AUs to store a new call of class i. Therefore,

$$E_{\mathrm{Er},i} = \sum_{n=V+U-t_{\mathrm{Er},i}+1}^{V+U} [Q(n)]^{\mathbf{M}}_{V,U}. \quad (28)$$

The average queue length L (for calls of all classes) is a mean value for the resources (expressed in AUs) occupied by calls of all classes that are waiting in the queue:

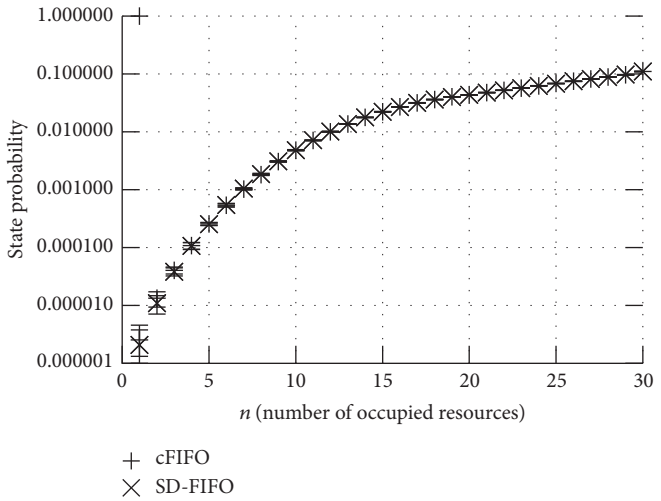
$$L = \sum_{n=1}^{U} n [Q(V+n)]^{\mathbf{M}}_{U,V}. \quad (29)$$

Let us consider now a method for a determination of the average number $L_{\mathrm{Er},i}$ of busy AUs in a queue occupied by calls of class i. This article proposes an approximate determination



n	cFIFO	±	SD-FIFO	±
0	0,00064291	$2{,}78985E-05$	0,000636624	$2{,}18414E-05$
5	0,0329501	0,00022368	0,0329734	0,000142919
10	0,0801712	0,00030422	0,0801458	0,000197831
15	0,0586448	0,000261642	0,0586142	0,000287113
20	0,0200488	0,000182976	0,020198	0,000166521
25	0,0058959	0,000103815	0,00593507	0,000119229
28	0,00281098	$7{,}29451E-05$	0,00284266	0,000127392
29	0,00221662	$6{,}85938E-05$	0,00223206	0,00009094
30	0,00184949	$6{,}73381E-05$	0,00182969	$5{,}44463E-05$

Figure 4: Occupancy distributions in SD-FIFO and cFIFO queueing systems (a = 0.6 Erl/AU).



n	cFIFO	±	SD-FIFO	±
5	0,000256823	$1{,}36216E-05$	0,000254323	$1{,}52558E-05$
10	0,00480267	0,000102197	0,00482289	0,000111694
15	0,0222231	0,000164597	0,0222044	0,000205997
20	0,0431998	0,000170365	0,0433649	0,000166231
25	0,0682773	0,000251808	0,06816	0,000218717
27	0,0817774	0,000403868	0,0817644	0,000242521
28	0,0883987	0,000276499	0,0888262	0,000234295
29	0,0965358	0,000410187	0,0972971	0,000323555
30	0,110553	0,000691369	0,109987	0,00039381

Figure 5: Occupancy distributions in SD-FIFO and cFIFO queueing systems (a = 1.2 Erl/AU).

of this parameter on the basis of the following reasoning. First, it is the average number of $L_{\mathrm{Er},i}$ AUs occupied by calls of class i that are waiting in the queue that has to be determined:

$$L_{\mathrm{Er},i} = \sum_{n=V+1}^{V+U} z_{\mathrm{Er},i}(n)\, t_{\mathrm{Er},i} \left[Q(n)\right]_{U,V}^{\mathbf{M}}, \tag{30}$$

where $z_{\mathrm{Er},i}(n)$ is the average number of calls of class i waiting in the queue in state n of the system. The product $z_{\mathrm{Er},i}(n)t_{\mathrm{Er},i}$ is the average number of AUs occupied in the buffer by waiting calls of class i in state n. In states n such that $n \le V$, the queue is empty; therefore the product $z_{\mathrm{Er},i}(n)t_{\mathrm{Er},i}$ is equal to zero.

In states n such that $n > V$, the total number of occupied AUs in the queue (by calls of all classes) can be determined as follows:

$$\sum_{i\in M} z_{\mathrm{Er},i}(n)\, t_{\mathrm{Er},i} = n - V. \tag{31}$$

Between the neighboring states $n-1$ and n there is a difference in the number of busy AUs occupied by calls of class i waiting in the queue, which, in the article, is denoted by the symbol $\Delta_{\mathrm{Er},c}(n)$. We can thus write

$$z_{\mathrm{Er},i}(n) \cdot t_{\mathrm{Er},i} = z_{\mathrm{Er},i}(n-1) \cdot t_{\mathrm{Er},i} + \Delta_{\mathrm{Er},i}(n), \tag{32}$$

where

$$\sum_{i\in M} \Delta_{\mathrm{Er},i}(n) = 1. \tag{33}$$

Let us assume that the parameter $\Delta_{\mathrm{Er},i}(n)$ is proportional to the intensity in which the buffer is occupied by calls of a given class i. We can therefore determine the value of the parameter $\Delta_{\mathrm{Er},i}(n)$ in the following way:

$$\Delta_{\mathrm{Er},i}(n) = \frac{\left[Q\left(n - t_{\mathrm{Er},i}\right)\right]_{V,U}^{\mathbf{M}} \lambda_{\mathrm{Er},i}\varphi_{\mathrm{Er},i}\left(n - t_{\mathrm{Er},i}\right)}{\sum_{j\in M}\left[Q\left(n - t_{\mathrm{Er},j}\right)\right]_{V,U}^{\mathbf{M}} \lambda_{\mathrm{Er},j}\varphi_{\mathrm{Er},j}\left(n - t_{\mathrm{Er},j}\right)}, \tag{34}$$

where

$$\varphi_{\mathrm{Er},i}\left(n-t_{\mathrm{Er},i}\right) = \begin{cases} t_{\mathrm{Er},i} & \text{for } \left(n-t_{\mathrm{Er},i}\right) \geq V, \\ n-V & \text{for } \left(V-t_{\mathrm{Er},i}\right) < \left(n-t_{\mathrm{Er},i}\right) < V. \end{cases} \quad (35)$$

Equation (35) is the result of a particular service discipline in the cFIFO queue. If in system state $(n - t_{\mathrm{Er},i})$ a new call of class i arrives, then in the occupancy states of the server that are determined by the condition $(V - t_{\mathrm{Er},i}) < (n - t_{\mathrm{Er},i}) < V$ a number of AUs of the considered call of class i will be immediately serviced by the server, while the remaining part will be directed to the queue. If $(n - t_{\mathrm{Er},i}) > V$, then all AUs of a new call of class i will be immediately placed in the queue. Equations (34) and (35) can be written as just one equation in the following way:

$$\Delta_{\mathrm{Er},i}(n) = \frac{\left[Q\left(n-t_{\mathrm{Er},i}\right)\right]^{\mathbf{M}}_{V,U}\lambda_{\mathrm{Er},i}\cdot\min\left(t_{\mathrm{Er},i},n-V\right)}{\sum_{j\in m}\left[Q\left(n-t_{\mathrm{Er},j}\right)\right]^{\mathbf{M}}_{V,Q}\lambda_{\mathrm{Er},j}\cdot\min\left(t_{\mathrm{Er},j},n-V\right)}. \quad (36)$$

Ultimately, (32), taking into account (36), can be written as follows:

$$z_{\mathrm{Er},i}(n)\,t_{\mathrm{Er},i} = z_{\mathrm{Er},i}(n-1)\,t_{\mathrm{Er},i} + \frac{\left[Q\left(n-t_c\right)\right]^{\mathbf{M}}_{V,U}\lambda_{\mathrm{Er},i}\cdot\min\left(t_{\mathrm{Er},i},n-V\right)}{\sum_{j\in m}\left[Q\left(n-t_{\mathrm{Er},j}\right)\right]^{\mathbf{M}}_{V,U}\lambda_{\mathrm{Er},j}\cdot\min\left(t_{\mathrm{Er},j},n-V\right)}, \quad (37)$$

where $n > V$. A determination of all values $z_{\mathrm{Er},i}(n)$ will make it possible, on the basis of (30), to evaluate and assess all queue lengths for calls of individual classes.

4.5. Model of a Queueing System with cFIFO Discipline and State-Dependent Call Streams. To determine the occupancy distribution in a queueing system with the cFIFO queueing discipline and offered BPP traffic (or traffic with any distribution of the call stream) we will use the general recurrence dependencies derived for a system with Erlang traffic (Section 4.4). To determine the occupancy distribution in the cFIFO queueing system with BPP traffic we use then dependencies (21) and (25) that, in their generalized form, can be rewritten as follows:

$$[Q(n)]^{\mathbf{M}}_{V,U} = \delta(n)\,[P(n)]^{\mathbf{M}}_{V+U,0}, \quad (38)$$

$$\delta(n) = \frac{\sum_{c=1}^{m}\delta\left(n-t_{T,c}\right)\left[y_{T,c}(n)\right]_{V+U,0}\,t_{T,c}}{\min(n,V)}, \quad (39)$$

where the parameter $[y_{T,c}(n)]_{V+U,0}$ is determined on the basis of a convolution operation (Formulas (16) and (17)). The occupancy distribution for the server with zero buffer $[P(n)]^{\mathbf{M}}_{V+U,0}$ in Formula (38) is determined on the basis of a convolution algorithm for which the input data are the distributions of single classes, determined for Erlang traffic by Formula (5) and for Engset and Pascal traffic by Formulas (7) and (9) as well as (8) and (10), respectively.

It should be emphasized that the application of the convolution algorithm in the proposed model makes it possible to determine characteristics of queueing systems for call streams other than BPP streams. Distributions of single classes that are the input data for the convolution algorithm can be then determined empirically on the basis of measurements or by simulation experiments.

On the basis of the occupancy distribution $[Q]^{\mathbf{M}}_{V,U}$, according to (28) and (29), it is possible to determine the blocking probability and the average queue length for calls of all classes.

Let us consider now the method for a determination of the average queue length for calls of a single class c. We will first determine the average number of occupied resources in the buffer $z_{T,c}(n)t_{T,c}$ in state n of the system by calls of class c. By the adoption of the same initial assumptions (30)–(33) that were earlier adopted for the Erlang call stream, (32) can be rewritten in such a way as to include any type traffic T:

$$z_{T,c}(n)\cdot t_{T,c} = z_{T,c}(n-1)\cdot t_{T,c} + \Delta_{T,c}(n). \quad (40)$$

Equation (36) will be also rewritten in our adopted notation to include the dependence between the call stream and the state:

$$\Delta_{T,c}(n) = \frac{\left[Q\left(n-t_{T,c}\right)\right]^{\mathbf{M}}_{V,U}\lambda_{T,c}(n)\cdot\min\left(t_{T,c},n-V\right)}{\sum_{j\in m}\left[Q\left(n-t_{T,j}\right)\right]^{\mathbf{M}}_{V,U}\lambda_{T,j}(n)\cdot\min\left(t_{T,j},n-V\right)}, \quad (41)$$

where $\lambda_{T,c}(n)$ determines the call arrival intensity of new calls of any traffic class c of type T in state n of the system.

In the case of Engset traffic, the parameter $\lambda_{\mathrm{En},j}(n)$ in state n depends on the total number of traffic sources $S_{\mathrm{En},j}$ and on the number of currently serviced calls in the server $x_{\mathrm{En},j}(n)$ and the number of calls waiting in the buffer $z_{\mathrm{En},j}(n)$:

$$\lambda_{\mathrm{En},j}(n) = \gamma_{\mathrm{En},j}\cdot\left(S_{\mathrm{En},j} - x_{\mathrm{En},j}(n) - z_{\mathrm{En},j}(n)\right) = \lambda_{\mathrm{En},j}\frac{S_{\mathrm{En},j} - x_{\mathrm{En},j}(n) - z_{\mathrm{En},j}(n)}{S_{\mathrm{En},j}}, \quad (42)$$

where $\lambda_{\mathrm{En},j}$ is the intensity of the call arrival process for calls of class j, determined with the assumption that all sources are not occupied:

$$\lambda_{\mathrm{En},j} = \gamma_{\mathrm{En},j}S_{\mathrm{En},j}. \quad (43)$$

By approximating the number of occupied resources $x_{\mathrm{En},j}(n) + z_{\mathrm{En},j}(n)$ in the considered queueing system with nonzero buffer by calls of class j in state n by the average number $[y_{\mathrm{En},j}(n)]_{V+U,0}$ of occupied resources in state n in the server with zero buffer, Formula (42) can be rewritten as follows:

$$\lambda_{\mathrm{En},j}(n) = \lambda_{\mathrm{En},j}\frac{S_{\mathrm{En},j} - \left[y_{\mathrm{En},j}(n)\right]_{V+U,0}}{S_{\mathrm{En},j}}. \quad (44)$$

Using the analogous simplifying assumptions as for Engset traffic we are in position to determine the intensity of the call arrival process for Pascal calls in state n of the system in the following way:

$$\lambda_{\mathrm{Pa},k}(n) = \lambda_{\mathrm{Pa},k} \frac{S_{\mathrm{Pa},k} + [y_{\mathrm{Pa},k}(n)]_{V+U,0}}{S_{\mathrm{Pa},k}}. \quad (45)$$

The call arrival intensities that have been described by (44) for calls of Engset class, and (45) for Pascal class, make it possible to determine, on the basis of (40) and (41), the average number of occupied resources of a queue by calls of individual classes.

Independently of the considered call streams, the proposed model can be written in the form of the following method, henceforth called the cFIFO method.

cFIFO Method

(1) Determination of occupancy distributions for individual classes: $[p]^{\{1\}}_{V+U,0}, [p]^{\{2\}}_{V+U,0}, \ldots, [p]^{\{m\}}_{V+U,0}$ in multiservice server with the capacity $V+U$ and with zero buffer.

(2) Determination of nonnormalized aggregated occupancy distributions of all classes, except a class c ($c \in \mathbf{M}$): $[P]^{\mathbf{M}\setminus\{1\}}_{\bullet,V+U,0}, [P]^{\mathbf{M}\setminus\{2\}}_{\bullet,V+U,0}, \ldots, [P]^{\mathbf{M}\setminus\{\mathbf{c-1}\}}_{\bullet,V+U,0}, [P]^{\mathbf{M}\setminus\{\mathbf{c}\}}_{\bullet,V+U,0}$, $[P]^{\mathbf{M}\setminus\{\mathbf{c+1}\}}_{\bullet,V+U,0}, \ldots, [P]^{\mathbf{M}\setminus\{\mathbf{m}\}}_{\bullet,V+U,0}$.

(3) Determination of the average number $[y_{T,c}(n)]_{V+U,0}$ of serviced calls in state n ($n \in \langle 0, V+U\rangle$) for each class c ($c \in \mathbf{M}$).

(4) Determination, by convolution algorithm, of the occupancy distribution $[P]^{\mathbf{M}}_{V+U,0}$ for multiservice server with the capacity $V+U$ and with zero buffer.

(5) Determination of transformation coefficients $\delta(n)$.

(6) Determination of the occupancy distribution $[Q(n)]^{\mathbf{M}}_{V,U}$ for queueing system with the cFIFO discipline.

(7) Determination of the average queue length L and the blocking probability $E_{T,c}$ ($c \in \mathbf{M}$) for calls of individual classes.

(8) Determination of the average number $z_{T,c}(n)t_{T,c}$ of occupied AUs by calls of class c ($c \in \mathbf{M}$) waiting in the queue in such states n of the system that $V < n \leq V+U$.

(9) Determination of the average number $L_{T,c}$ of occupied AUs by calls of class c, ($c \in \mathbf{M}$), waiting in the queue.

5. Numerical Results

In order to verify the proposed analytical method the results of the calculations were compared with the results provided by the simulation experiments. For this purpose, a dedicated simulation program for a cFIFO queueing system evaluation was developed. The simulator uses an experiment with steady system time in which the process interaction method is used. The classes with finite number of sources were implemented according to the principles described in [32]. The program was written in the C++ language in the Qt environment. The software architecture and process interaction method is described in [33]. The obtained results are presented in graphs as the function of traffic offered to one AU in the server:

$$a = \sum_{c=1}^{m} \frac{A_{T,c}}{V}. \quad (46)$$

Each simulation experiment to determine the characteristics of the system under investigation for particular values a involved 10 series. The prerequisite condition for a single series to be completed was call service for 10,000,000 calls or a loss of 10,000 calls. The adopted length of a single series in the simulation made it possible to determine confidence intervals at the 95% level. The results obtained on the basis of the analytical model are shown in the graphs with lines, whereas the results of the simulation are presented with appropriate symbols. In most cases, the values for the confidence intervals in the graphs are so small that they do not exceed the value of the symbol that denotes the result of a simulation. The proportions of offered traffic in all presented cases were $A_{T,1} : A_{T,2} : A_{T,3} = 1 : 1 : 1$. For each of the considered queueing systems, the results for the blocking probability and the average queue lengths for individual classes, expressed in the number of occupied AUs, will be presented.

Figures 6–8 show the results for three queueing systems that service Erlang traffic. The assumption in the systems presented in Figures 6 and 7 was that the intensities of service streams of particular call classes were, respectively, equal to $\mu_{\mathrm{Er},1} = 1$, $\mu_{\mathrm{Er},2} = 1$, $\mu_{\mathrm{Er},3} = 1$, whereas in the case of the results shown in Figure 8 they were, respectively, $\mu_{\mathrm{Er},1} = 1$, $\mu_{\mathrm{Er},2} = 2$, $\mu_{\mathrm{Er},3} = 3$.

The next stage in the process of verification of the proposed analytical model was to test its accuracy for systems to which a mixture of Erlang, Engset, and Pascal traffic was offered (Figure 9). In this case, the intensities of service streams of particular call classes were, respectively, equal to $\mu_{\mathrm{En},1} = 1$, $\mu_{\mathrm{Pa},2} = 1$, $\mu_{\mathrm{Er},3} = 1$. The number of traffic sources for Engset and Pascal classes is equal to $S_{\mathrm{En},1} = 30$, $S_{\mathrm{Pa},2} = 20$, respectively.

In the first case of considered types of non-Poissonian offered traffic (denoted by symbol NE), each call stream is described by normal distribution $N(\mathrm{mean}_c, \mathrm{var}_c)$ and each service stream is described by exponential distribution. It is assumed that the mean value mean_c is equal to the inverse of offered traffic intensity of a given call class ($a^{-1}_{N,c}$) and variance is equal to mean^2_c. So, we consider the system with normally distributed call stream $N(a^{-1}_{N,c}, a^{-2}_{N,c})$. The intensities of service streams are equal to $\mu_{N,1} = \mu_{N,2} = \mu_{N,3} = 1$ (similar to previously considered queueing systems). The results for this queueing system are presented on Figure 10.

Figure 11 shows the results for a queueing system that, for each serviced traffic class, the call stream and service stream are described by normal distribution. This type of traffic is denoted by symbol NN. Similar to previous case, each call stream is described by the distribution $N(a^{-1}_{\mathrm{NN},c}, a^{-2}_{\mathrm{NN},c})$. Each

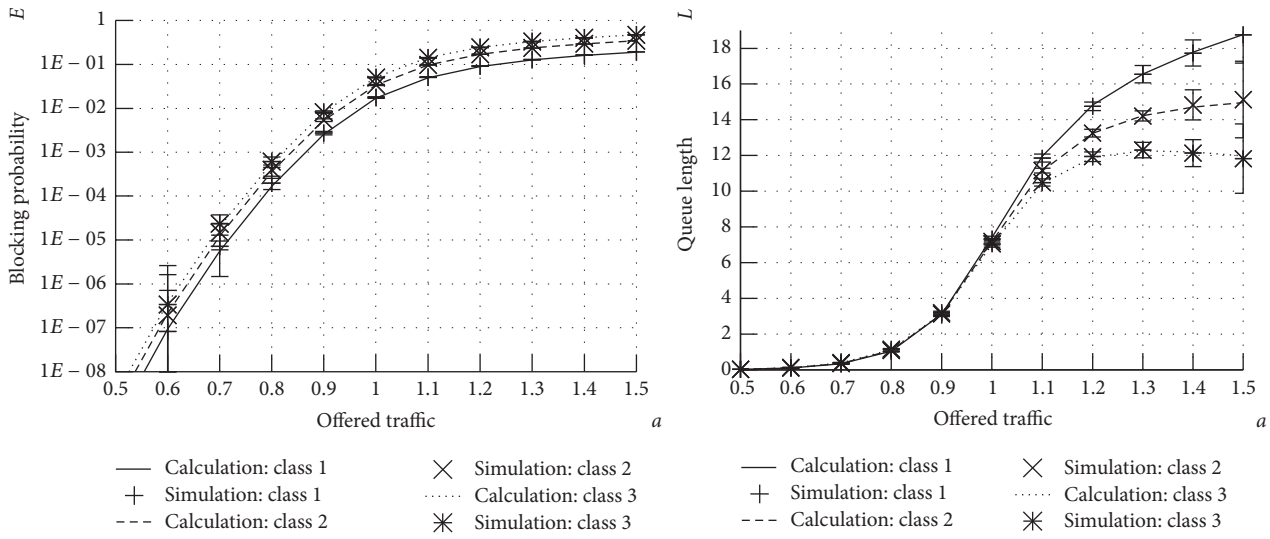


Figure 6: The blocking probability and the average queue length ($V = 20, U = 50, m = 3, t_{Er,1} = 1, t_{Er,2} = 2, t_{Er,3} = 3$).

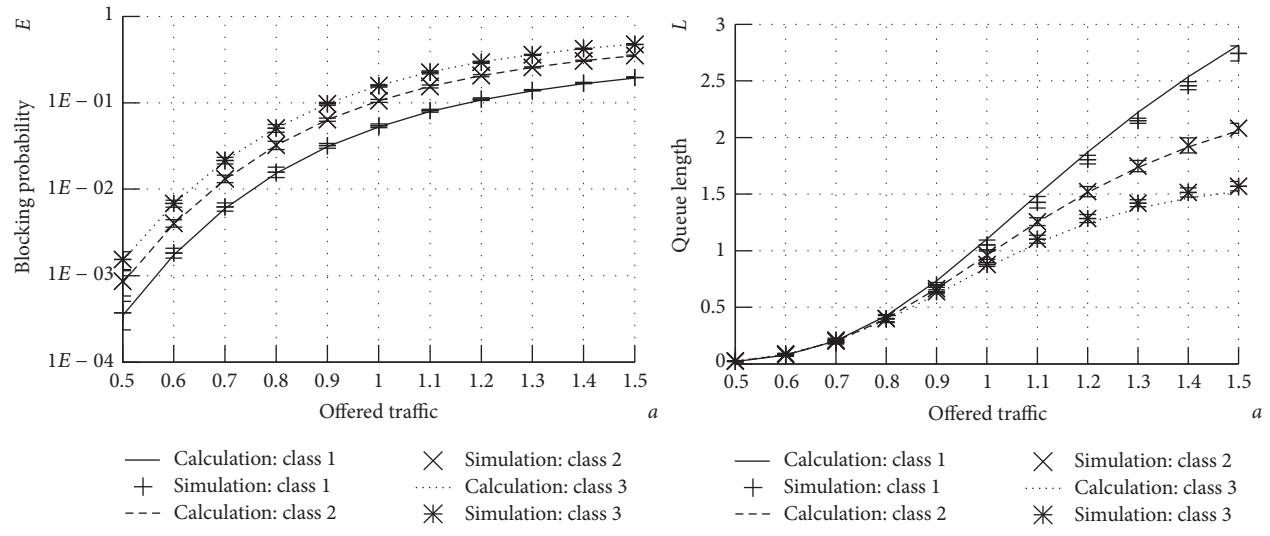


Figure 7: The blocking probability and the average queue length ($V = 20, U = 10, m = 3, t_{Er,1} = 1, t_{Er,2} = 2, t_{Er,3} = 3$).

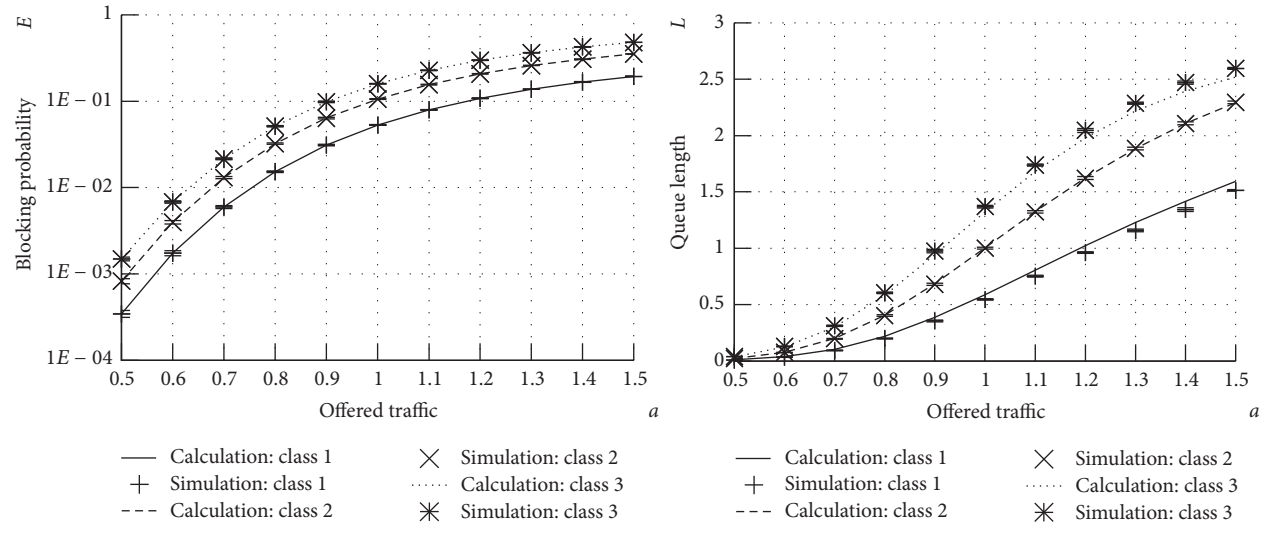


Figure 8: The blocking probability and the average queue length ($V = 20, U = 10, m = 3, t_{Er,1} = 1, t_{Er,2} = 2, t_{Er,3} = 3$).

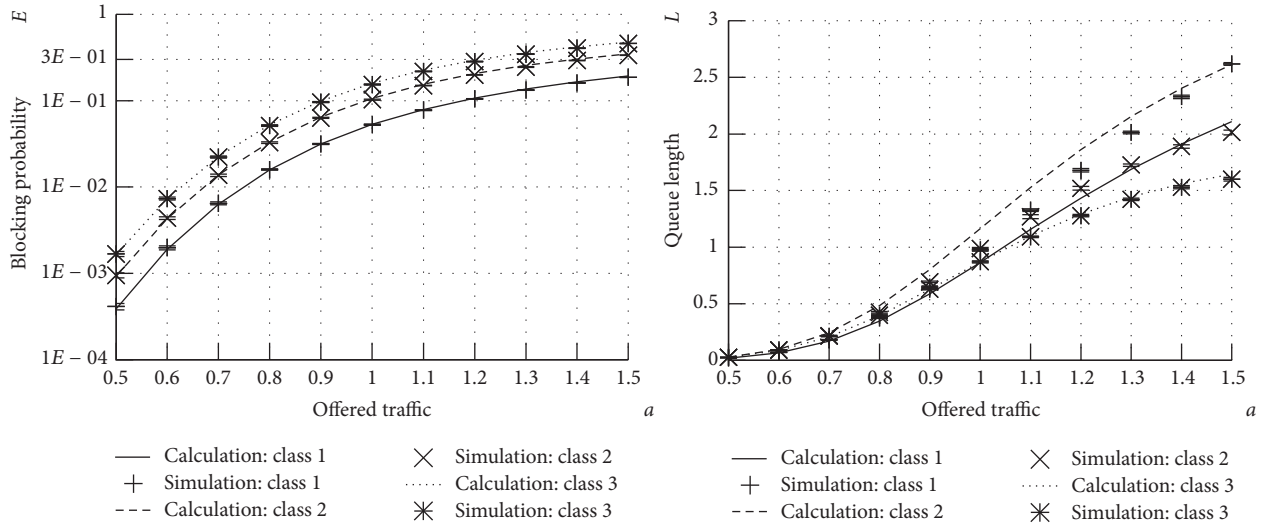


FIGURE 9: The blocking probability and the average queue length ($V = 20, U = 10, m = 3, t_{\mathrm{En},1} = 1, t_{\mathrm{Pa},2} = 2, t_{\mathrm{Er},3} = 3, \mu_{\mathrm{En},1} = 1, \mu_{\mathrm{Pa},2} = 1, \mu_{\mathrm{Er},3} = 1, S_{\mathrm{En},1} = 30, S_{\mathrm{Pa},2} = 20$).

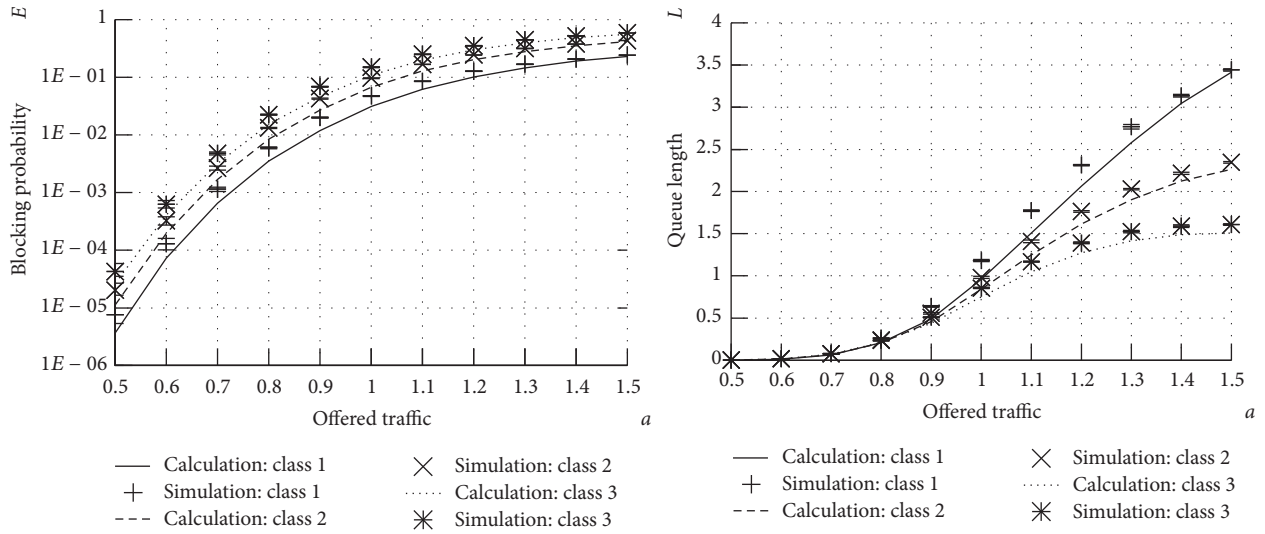


FIGURE 10: The blocking probability and the average queue length ($V = 20, U = 10, m = 3, t_{\mathrm{NE},1} = 1, t_{\mathrm{NE},2} = 2, t_{\mathrm{NE},3} = 3$).

service stream is given by the distribution $N(\mu_{\mathrm{NN},c}^{-1}, \mu_{\mathrm{NN},c}^{-2})$, where in the experiment was admitted $\mu_{\mathrm{NN},1} = \mu_{\mathrm{NN},2} = \mu_{\mathrm{NN},3} = 1$.

The results presented on Figure 12 were worked out for queueing system in which call streams are described by uniform distribution uni f(min, max) and service streams are described by exponential distribution. This type of traffic is denoted by symbol UE. It is assumed that for each traffic class the uniform distribution is given by parameters min = 0 and max = $2a_{\mathrm{UE},c}^{-1}$ (where $a_{\mathrm{UE},c}$ is the traffic intensity of class c). The intensities of service streams are equal to $\mu_{\mathrm{UE},1} = 1$, $\mu_{\mathrm{UE},2} = 1$, $\mu_{\mathrm{UE},3} = 1$.

The results of the study showed in Figures 6–12 confirm high accuracy of the proposed approximate analytical model of the queueing system with the cFIFO discipline. The diversified accuracy of the results stems from the dependence between the method for a determination of the average number of calls for individual classes in the system adopted in the study (independently for the server and queue) and the type of offered traffic. More accurate results were obtained for BPP streams, while those for non-Poisson streams were less accurate. This results from the adoption of the method of the approximation of the distribution in the system with the cFIFO discipline by the distribution in the SD-FIFO system that can be only precisely determined for Poisson streams. During the study no dependence between the accuracy of the model and the number of offered classes of calls and their demands was observed.

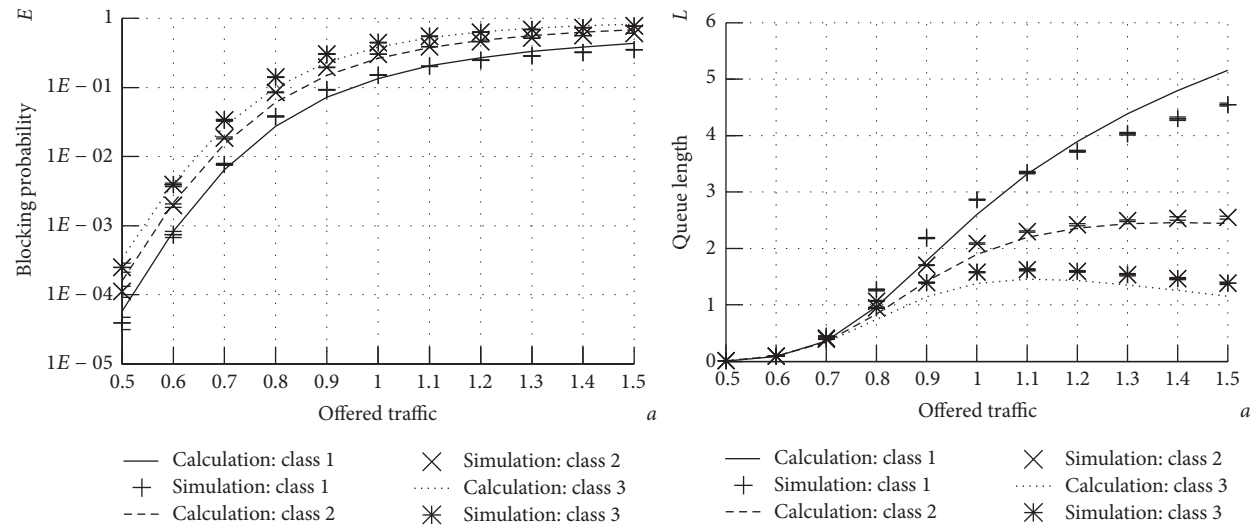


Figure 11: The blocking probability and the average queue length ($V = 20, U = 10, m = 3, t_{NN,1} = 1, t_{NN,2} = 2, t_{NN,3} = 3$).

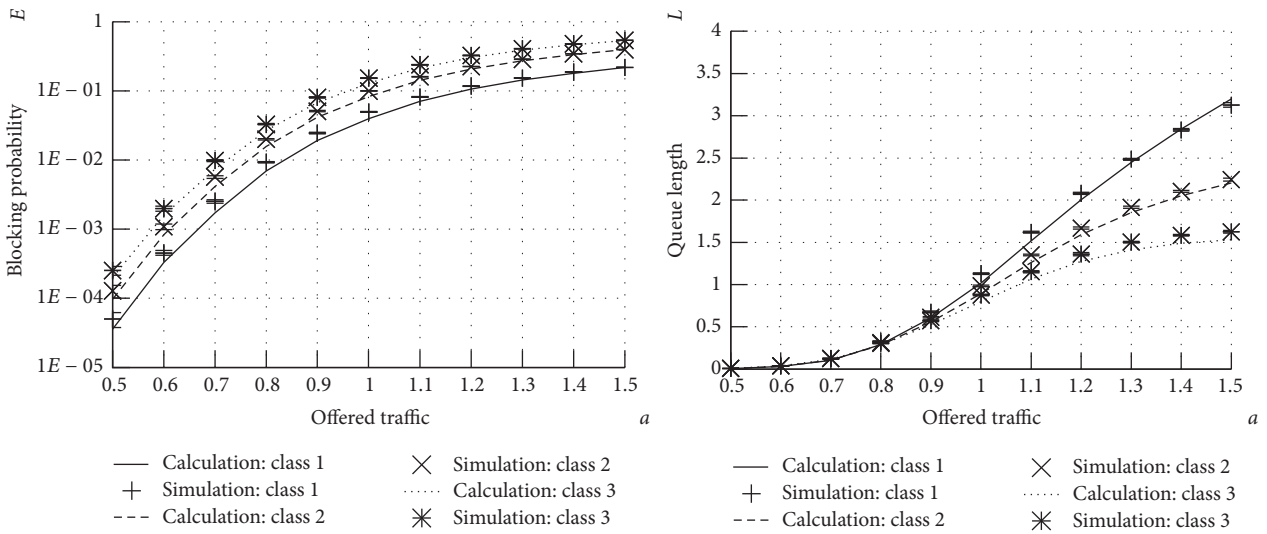


Figure 12: The blocking probability and the average queue length ($V = 20, U = 10, m = 3, t_{UE,1} = 1, t_{UE,2} = 2, t_{UE,3} = 3$).

6. Conclusions

This article proposes an approximate analytical model of a queueing system that services multiservice traffic with the cFIFO queueing discipline. The model is based on a convolution algorithm in which to determine certain characteristics, such as the average number of serviced calls, the formulas derived on the basis of accurate Markovian models for multiservice systems with zero and nonzero buffer (with the SD-FIFO service discipline for the queue) are used. The introduction of these assumptions makes it possible to construct a general model of a multiservice queueing system with the cFIFO discipline. The proposed model is characterized by high accuracy, which makes an analysis of cFIFO systems for call streams with any, mutually independent, probabilistic distributions possible. This high accuracy stems from the fact that any errors that may possibly result from the adopted approach will be minimized by the convolution operation.

Appendix

Resource Allocation Units

The application of multirate analytical models to analyze present-day broadband telecommunications systems is possible with a proper bandwidth discretization. Discretization allows the capacity of a system and bit rates demanded by calls to be expressed in the so-called allocation unit (AU).

The discretization process itself consists in replacing a variable bit rate (VBR) of call streams with a specific constant bit rate (CBR) called the equivalent bandwidth. There

are many methods and algorithms specific methods of determining the equivalent bandwidth for given network types and services in the literature of the subject, for example, [24, 26, 34–37]. The value of equivalent bandwidth depends on various parameters, including the admissible packet-loss ratio, admissible delay, link flow capacity (bit rate), average and peak bit rates, packet stream type (e.g., self-similar streams), and network type.

The procedure of discretization process for system services m call classes any type of T, and $T \in \{\mathrm{Er}, \mathrm{En}, \mathrm{Pa}, G\}$ can be written in the form of the following algorithm.

Algorithm 1.

Discretization

(1) Determination of the equivalent bandwidth for all offered call classes: $R_{T,1}, \ldots, R_{T,c}, \ldots, R_{T,m}$.

(2) Determination of the allocation unit (AU). The value of this parameter can be calculated as the greatest common divisor (GCD) of all equivalent bandwidths in a considered system:

$$R_{\mathrm{AU}} = \mathrm{GCD}\left(R_{T,1}, \ldots, R_{T,c}, \ldots, R_{T,m}\right). \tag{A.1}$$

(3) Determination of the number of allocation units demanded by calls of class c:

$$t_c = \left\lceil \frac{R_c}{R_{\mathrm{AU}}} \right\rceil. \tag{A.2}$$

(4) Determination of the capacity of the system (expressed in allocation units):

$$V = \left\lfloor \frac{C}{R_{\mathrm{AU}}} \right\rfloor, \tag{A.3}$$

where C is the capacity of the telecommunications system under consideration.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The present work is financed with the Ministry of Science and Higher Education resources for academic purposes in the year 2016 within the frame of own research project entitled "The Structure, Analysis and Designing of Modern Switching Systems and Telecommunications Networks."

References

[1] Ericsson, "Ericsson mobility report," Tech. Rep., Ericsson, 2016.

[2] M. Sobieraj, M. Stasiak, J. Weissenberg, and P. Zwierzykowski, "Analytical model of the single threshold mechanism with hysteresis for multi-service networks," *IEICE Transactions on Communications*, vol. 95, no. 1, pp. 120–132, 2012.

[3] J. S. Kaufman, "Blocking with retrials in a completely shared resource environment," *Performance Evaluation*, vol. 15, no. 2, pp. 99–116, 1992.

[4] I. D. Moscholios, M. D. Logothetis, J. S. Vardakas, and A. C. Boucouvalas, "Performance metrics of a multirate resource sharing teletraffic model with finite sources under the threshold and bandwidth reservation policies," *IET Networks*, vol. 4, no. 3, pp. 195–208, 2015.

[5] G. M. Stamatelos and V. N. Koukoulidis, "Reservation-based bandwidth allocation in a radio ATM network," *IEEE/ACM Transactions on Networking*, vol. 5, no. 3, pp. 420–428, 1997.

[6] T. Bonald and J. Virtamo, "A recursive formula for multirate systems with elastic traffic," *IEEE Communications Letters*, vol. 9, no. 8, pp. 753–755, 2005.

[7] J. W. Cohen, "The multiple phase service network with generalized processor sharing," *Acta Informatica*, vol. 12, no. 3, pp. 245–284, 1979.

[8] K. Lindberger, "Balancing quality of service, pricing and utilisation in multiservice networks with stream and elastic traffic," in *Proceedings of the 16th International Teletraffic Congress*, pp. 1127–1136, Edinburgh, UK, 1999.

[9] T. Bonald and J. W. Roberts, "Internet and the Erlang formula," *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 1, pp. 23–30, 2012.

[10] S. Hanczewski, M. Stasiak, and J. Weissenberg, "The queueing model of a multiservice system with dynamic resource sharing for each class of calls," in *Computer Networks*, A. Kwiecień, P. Gaj, and P. Stera, Eds., vol. 370 of *Communications in Computer and Information Science*, pp. 436–445, Springer, Berlin, Germany, 2013.

[11] S. Hanczewski, M. Stasiak, and J. Weissenberg, "A queueing model of a multi-service system with state-dependent distribution of resources for each class of calls," *IEICE Transactions on Communications*, vol. 97, no. 8, pp. 1592–1605, 2014.

[12] S. Hanczewski, D. Kmiecik, M. Stasiak, and J. Weissenberg, "Multiservice queuing system with elastic traffic," in *Proceedings of the IEICE General Conference*, pp. S-46–S-47, The Institute of Electronics, Information and Communication Engineers, March 2016.

[13] M. Stasiak, "Queuing systems for the internet," *IEICE Transactions on Communications*, vol. 99, no. 6, pp. 1234–1242, 2016.

[14] T. Bonald, A. Proutière, and J. Virtamo, "A queueing analysis of max-min fairness, proportional fairness and balanced fairness," *Queueing Systems*, vol. 53, no. 1, pp. 65–84, 2006.

[15] J.-P. Haddad and R. R. Mazumdar, "Congestion in large balanced multirate networks," *Queueing Systems*, vol. 74, no. 2-3, pp. 333–368, 2013.

[16] V. Iversen, "The exact evaluation of multi-service loss system with access control," in *Proceedings of the 17th Nordic Teletraffic Seminar*, pp. 56–61, Lund, Sweden, August 1987.

[17] K. Ross, *Multiservice Loss Models for Broadband Telecommunication Network*, Springer, London, UK, 1995.

[18] I. D. Moscholios, M. D. Logothetis, J. S. Vardakas, and A. C. Boucouvalas, "Congestion probabilities of elastic and adaptive calls in Erlang-Engset multirate loss models under the threshold and bandwidth reservation policies," *Computer Networks*, vol. 92, part 1, pp. 1–23, 2015.

[19] M. Glabowski, A. Kaliszan, and M. Stasiak, "Asymmetric convolution algorithm for full-availability group with bandwidth reservation," in *Proceedings of the Asia-Pacific Conference on Communications (APCC '06)*, IEEE, Busan, South Korea, September 2006.

[20] M. Głąbowski, A. Kaliszan, and M. Stasiak, "On the application of the asymmetric convolution algorithm in modeling of full-availability group with bandwidth reservation," in *Managing Traffic Performance in Converged Networks: 20th International Teletraffic Congress, ITC20 2007, Ottawa, Canada, June 17–21, 2007. Proceedings*, L. Mason, T. Drwiega, and J. Yan, Eds., vol. 4516 of *Lecture Notes in Computer Science*, pp. 878–889, Springer, Berlin, Germany, 2007.

[21] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Two-dimensional convolution algorithm for modelling multiservice networks with overflow traffic," *Mathematical Problems in Engineering*, vol. 2013, Article ID 852082, 18 pages, 2013.

[22] A. Kaliszan, M. Głabowski, and M. Stasiak, "Generalised convolution algorithm for modelling state-dependent systems," *IET Circuits, Devices & Systems*, vol. 8, no. 5, pp. 378–386, 2014.

[23] J. Roberts, Ed., *Performance Evaluation and Design of Multiservice Networks*, Final Report COST 224, Commission of the European Communities, Brussels, Belgium, 1992.

[24] J. Roberts, V. Mocci, and I. Virtamo, Eds., *Broadband Network Teletraffic, Final Report of Action COST 242*, Commission of the European Communities, Springer, Berlin, Germany, 1996.

[25] J. Y. Hui, "Resource allocation for broadband networks," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1598–1608, 1988.

[26] F. Kelly, "Notes on effective bandwidth," Tech. Rep., University of Cambridge, 1996.

[27] J. Karlsson, "Loss performance in trunk groups with different capacity demands," in *Proceedings of the 13th International Teletraffic Congress*, vol. Discussion, pp. 201–212, Copenhagen, Denmark, 1991.

[28] J. S. Kaufman, "Blocking in a shared resource environment," *IEEE Transactions on Communications*, vol. 29, no. 10, pp. 1474–1481, 1981.

[29] J. Roberts, "A service system with heterogeneous user requirements—application to multi-service telecommunications systems," in *Proceedings of Performance of Data Communications Systems and Their Applications*, G. Pujolle, Ed., pp. 423–431, North Holland, Amsterdam, The Netherlands, 1981.

[30] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Modeling product-form state-dependent systems with BPP traffic," *Performance Evaluation*, vol. 67, no. 3, pp. 174–197, 2010.

[31] R. Syski, *Introduction to Congestion Theory in Telephone Systems*, Studies in Telecommunication, North Holland, The Netherlands, 1986.

[32] M. Głabowski and A. Kaliszan, "Simulator of full-availability group with bandwidth reservation and multi-rate bernoulli-poisson-pascal traffic streams," in *Proceedings of the International Conference on "Computer as a Tool" (EUROCON '07)*, pp. 2271–2277, Warszawa, Poland, September 2007.

[33] S. Hanczewski and A. Kaliszan, "Simulation studies of queueing systems," in *Proceedings of the 10th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP '16)*, Prague, Czech Republic, July 2016.

[34] T. Neame, M. Zuckerman, and R. Addie, "A paractical approach for multimedia traffic modeling," in *Proceedings of the 5th International Conference on Broadband Communications*, pp. 73–82, Hong Kong, November 1999.

[35] N. L. S. Fonseca, G. S. Mayor, and C. A. V. Neto, "On the equivalent bandwidth of self-similar sources," *ACM Transactions on Modeling and Computer Simulation*, vol. 10, no. 2, pp. 104–124, 2000.

[36] S. Bodamer and J. Charzinski, "Evaluation of effective bandwidth schemes for self-similar traffic," in *Proceedings of the 13th ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management*, pp. 21.1–21.10, Monterey, Calif, USA, 2000.

[37] J. Charzinski, "Internet traffic measurment and characterisation results," in *Proceedings of the 13th International Symposium on Services and Local Access (ISSLS '00)*, Stockholm, Sweden, June 2000.

Hindawi Publishing Corporation
Mobile Information Systems
Volume 2016, Article ID 5187303, 16 pages
http://dx.doi.org/10.1155/2016/5187303

Research Article

Enhancing Radio Access Network Performance over LTE-A for Machine-to-Machine Communications under Massive Access

Fatemah Alsewaidi,[1] Angela Doufexi,[1] and Dritan Kaleshi[2]

[1]*Electrical and Electronic Engineering Department, University of Bristol, Bristol BS8 1UB, UK*
[2]*Digital Catapult, London NW1 2RA, UK*

Correspondence should be addressed to Fatemah Alsewaidi; eefaha@bristol.ac.uk, Angela Doufexi; a.doufexi@bristol.ac.uk, and Dritan Kaleshi; dritan.kaleshi@digicatapult.org.uk

Received 29 July 2016; Revised 7 October 2016; Accepted 12 October 2016

Academic Editor: Ioannis Moscholios

Copyright © 2016 Fatemah Alsewaidi et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The expected tremendous growth of machine-to-machine (M2M) devices will require solutions to improve random access channel (RACH) performance. Recent studies have shown that radio access network (RAN) performance is degraded under the high density of devices. In this paper, we propose three methods to enhance RAN performance for M2M communications over the LTE-A standard. The first method employs a different value for the physical RACH configuration index to increase random access opportunities. The second method addresses a heterogeneous network by using a number of picocells to increase resources and offload control traffic from the macro base station. The third method involves aggregation points and addresses their effect on RAN performance. Based on evaluation results, our methods improved RACH performance in terms of the access success probability and average access delay.

1. Introduction

Machine-to-machine (M2M) communication refers to data communication between entities (e.g., natural disaster alarms, smart meters, vehicle mobile global positioning systems (GPSs), and wearable health monitors) that do not necessarily need human interaction. Examples of M2M applications are shown in Figure 1. Different access-standardized technologies exist for M2M communications, such as wired networks (i.e., Ethernet), capillary (e.g., ZigBee and low-power WiFi), and cellular (e.g., General Packet Radio Service (GPRS) and Long Term Evolution-Advanced (LTE-A) standards). In this paper, we focus on the cellular M2M sector employing LTE-A technology. LTE-A provides benefits, such as ubiquitous coverage, large capacity, and interference management that enable it to cope with the needs of different M2M applications.

The general architecture of M2M communications over LTE networks and for M2M service requirements is described in [1–4]. Reference [5] introduces different network access methods for M2M devices (M2M-Ds). These methods are considered by the 3rd Generation Partnership Project (3GPP) in the description releases of the M2M work plan in [6]. M2M-Ds can directly establish a link with the evolved Node B (eNB) through an M2M gateway (M2M-GW) or with another M2M-D.

M2M communications will enable Internet of Things (IoT) connectivity. Advancements are swiftly moving from fourth-generation (4G) mobile communications toward ubiquitously connected devices. The increase of M2M-Ds is expected to reach 3.2 billion in 2019 [7]. 3GPP considered network enhancements for M2M communications in [4] and further optimization in LTE-A release 13 [8] for M2M communications that will enable LTE-A to play an essential role in fifth-generation (5G) systems.

Most M2M applications deal with infrequent small data transmissions. Nevertheless, this may cause network congestion, including radio access network (RAN) congestion, which affects network performance (such as by causing delays and reliability issues). This is especially the case if numerous devices access the network in a highly synchronized manner (e.g., after a power outage or violent windstorm). This leads to RAN congestion that causes an unacceptable delay, packet loss, or service unavailability [9]. The focus of this paper

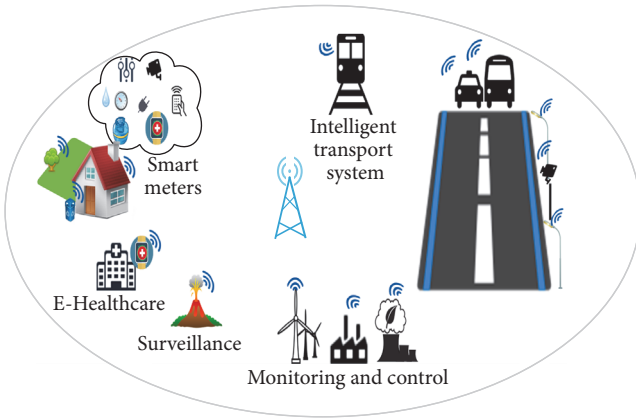


Figure 1: Examples of M2M applications.

is only on signalling congestion over RAN on account of the massive number of M2M-Ds simultaneously initiating a random access (RA) procedure.

A major research challenge in this context is development of an air interface to support the deployment of a massive number of M2M-Ds [10]. This paper addresses this challenge by investigating issues relating to RAN. These issues are highlighted below. In short, we

(i) simplify the complexity of the network to support the deployment of a massive number of M2M-Ds without influencing the LTE-A system architecture,

(ii) accommodate signalling overhead that is from a massive number of M2M-Ds,

(iii) achieve low latency, where some of the applications are nontolerant delay applications,

(iv) enhance the coverage for devices at the edge of the cell.

The major contributions of this paper are summarized below.

(i) We investigate the impact of the physical random access channel (PRACH) configuration index to increase random access opportunities (RAOs). The goal of this approach is to increase the number of RAOs and show how the increase affects RACH performance.

(ii) In addition, we examine the allocation of several picocells to increase the number of preambles and decrease the traffic in the macrocell.

(iii) We furthermore consider employing aggregation points or M2M-GWs on the access points of small networks. In reality, we can find small networks within the range of a macrocell, but with different RAN technologies.

(iv) The goal of this approach is to explore the effect of aggregation points or M2M-GWs (aggregation points and M2M-GWs are hereafter interchangeably used) on RACH performance. The role of the aggregation point is to collect device access requests from the small network and send them to the eNB and vice versa.

In this study, we aim to evaluate and enhance RACH performance over LTE-A under an extreme scenario (i.e., traffic model two in 3GPP [9]). In the extreme scenario, numerous M2M devices (up to 30,000) access the network over 10 s in a highly synchronized manner to enable implementation through beta distribution [9]. RACH performance is evaluated in terms of the preamble collision probability, average number of preamble transmissions, access success probability, and average access delay. The results are based on the unconditioned packet transmission [11, 12]. This study considers different density values according to [9] and statistics of available M2M-Ds in Bristol City Centre in the United Kingdom [13]. These values are used to analyse the RACH capacity. To validate the proposed approach, we built an RA procedure using MATLAB. The simulation results were validated in [13] with the 3GPP technical report [9].

The remainder of this paper is structured as follows. In Section 2, related work on existing RAN congestion control schemes is presented. The ways in which the proposed methods differ from those of previous works are also discussed. Section 3 overviews the contention-based RA procedure and RACH capacity evaluation metrics. RA improvement methods are outlined in Section 4. The system model and assumptions for the simulations are described in Section 5. An evaluation of RACH, including results and discussions, is presented in Section 6. Section 7 concludes this work.

2. Related Work

Various methods have been proposed to address the overload in RAN. General classification of these techniques based on [9, 10, 16–22] is shown in Figure 2. In [9], different solutions are proposed to control RAN congestion, including access class barring (ACB) schemes, separate RACH resources for M2M communications, dynamic allocation of RACH resources, slotted access, a specific backoff scheme, and a pull-based scheme. Those methods and others are likewise described in [17, 18]. In [23], the solutions proposed in [9] are evaluated for RACH overload (except slotted access and the pull-based scheme). Nonetheless, the mentioned methods are considered inefficient schemes if they are separately used [16].

In [24], the authors provided analysis that is applied to the RA procedure for M2M communications over LTE-A. The authors consider multiple classes with different qualities of service (QoS) for M2M-Ds in smart grids. The various classes are expressed by different ACB and backoff timers (BOs). They consider the on-off arrival process for M2M-Ds, which is a realistic approach for M2M communications in smart grid environments.

In [14], new mechanisms are proposed to solve RAN congestion considering only "delay tolerant" devices. The first method has a longer backoff value for preamble retransmission, which involves utilizing a longer backoff value in case of any collisions occurring to spread access reattempts from "delay tolerant" devices. The other method is the

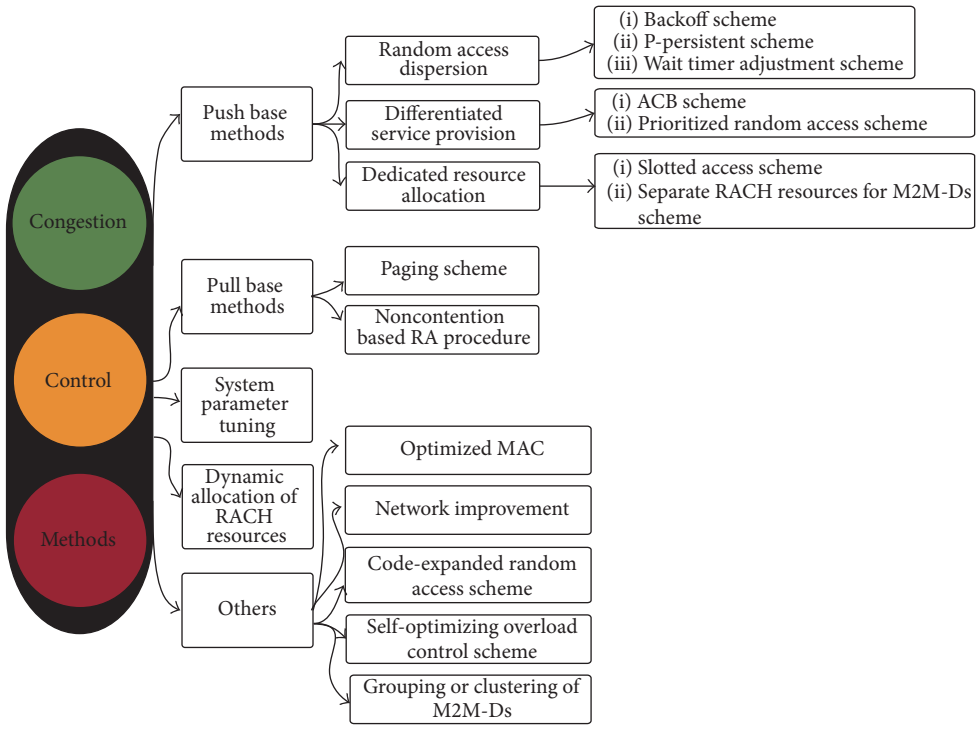


Figure 2: General classification of congestion control methods in RAN.

prebackoff approach undertaken before the first preamble transmissions, where the devices can read the random access response packet data unit (RAR PDU) of other devices to obtain the backoff information, even before performing the first attempt. This approach spreads the initial preamble transmission for a "delay-tolerant access" request over the timescale defined by the "delay tolerant access backoff value." The network can further prevent or spread the first preamble transmission with the prebackoff approach. Both methods have been evaluated under traffic model two with a maximum backoff value of 960 ms. The proposed schemes improve RACH performance in terms of access success, collision probability, and average preamble transmissions. However, no numerical results exist for the average access delay because they have only proposed solutions for "delay tolerant" devices. In [25], the authors proposed a dynamic backoff scheme to control the congestion in RAN. They evaluated the RACH performance, which considers extreme scenarios. This method enhances the access success probability; however, the drawback of this method is the increase of the access delay as a result of increasing the backoff timer. This increase is not accepted for nontolerant delay applications.

However, few studies consider the extreme scenario in which it generates synchronized traffic to evaluate RACH performance under the high density of devices within 10 s. The authors in [15] considered different traffic classes to address the RAN overload problem. They proposed the prioritized random access (PRA) architecture. The PRA architecture is comprised of two components: virtual resource allocation with class-dependent backoff procedures and dynamic access barring. They evaluated the RACH performance in terms of the access success probability and average access delay for each class. However, the average access delay for smart meters that arrived in a synchronized manner (i.e., the arrival rate follows the extreme scenario) is too high.

In [25], the authors proposed a dynamic backoff scheme to control the congestion in RAN. They evaluated the RACH performance with consideration of the extreme scenario. This method enhances the access success probability; nevertheless, it has the drawback of increasing the access delay as a result of increasing the BO. This increase is unacceptable for nontolerant delay applications. Meanwhile, the authors in [26] proposed a group-based optimization method with a resource coordination scheme in RACH. They classified the signalling messages into two types: diverse messages and redundant messages to avoid signalling congestion. Although this method enhances RACH performance in terms of access success probability, it provides no means to enhance RACH in terms of the access delay. Access delay is an important metric in the RACH performance evaluation and should therefore be considered.

In [27], a cooperative ACB scheme was proposed to enhance the performance of the ordinary ACB. This scheme is based on using the benefit of the heterogeneous multitier network in LTE-A. The authors deployed three picocells, each with 20% of the M2M-Ds among M devices, and seven macrocells with 6% of the M2M-Ds among M devices (i.e., except the centric macrocell with 4% of the M2M-Ds among M devices). Additionally, they jointly optimized the ACB parameters with all eNBs according to the level of congestion in each eNB. The scheme uses only one preamble to limit the random access resource to the time domain instead of

the preamble domain. The scheme improves the average access delay compared to the conventional ACB. However, the average access delay in the proposed scheme remains unacceptable because the average access delay for 30,000 M2M-Ds is approximately 4×10^4 ms. In addition, the authors did not indicate the type of traffic model used for the M2M-D arrival.

Recently, the authors in [28] provided a set of guidelines for the resource allocation task in RACH with an investigation on RACH performance in terms of the backoff timer and maximum preamble transmission attempt ($\text{Max}_{\text{PreamTrans}}$). However, traffic of the arrival devices follows a Poisson distribution, which is in contrast to the approach of this study. In [18], the effect of different settings of the PRACH configuration index (i.e., 0, 3, 6, and 9) was explored. Different values of $\text{Max}_{\text{PreamTrans}}$ (i.e., 3, 10, 15, and 50) increase the RA resources and chances for the devices to successfully access the network by increasing the attempts of, respectively, transmitting the preamble. The authors evaluated RACH performance under only simultaneous arrivals of more than 1,000 devices. The evaluation metrics used in [18] include the average access delay, average energy consumption, blocking probability, and average number of preamble transmissions. On the other hand, the focus of the present approach is traffic model two, whereby massive devices with different ranges (i.e., from 5,000 to 30,000 devices) synchronistically access the network within 10 s.

The authors in [13, 29] analysed RAN performance for 16,000 M2M-Ds for LTE-A in different frequency bands. The authors also considered tuning of different system parameters to enhance RACH performance, such as BO, the medium access control (MAC) contention resolution timer (*mac-Contention Resolution Timer*), and $\text{Max}_{\text{PreamTrans}}$. The results showed RACH enhancement in terms of the access success probability only for specific values of $\text{Max}_{\text{PreamTrans}}$. The BO was shown to improve RACH performance in terms of the access success probability; however, it increased the average access delay.

The motivation behind the approaches proposed in this paper is to address the congestion in RAN caused by signalling overhead using the existing LTE-A system architecture. In addition, this paper considers different densities of devices to evaluate RACH performance under extreme scenarios.

3. Random Access Channel

In LTE-A, M2M-Ds use the RA procedure to establish a radio link (i.e., creating a transition from the radio resource control (RRC) idle mode to the RRC connected mode) to complete an intrasystem handover for synchronizing the devices (in case they are in the RRC connected mode but not synchronized, and uplink or downlink data arrive). Alternatively, it synchronizes the devices to reestablish an RRC connection or to position or schedule a request. The RA procedure can be either contention-based or noncontention-based.

The contention-based RA procedure is used for connection establishment. The device randomly selects the access

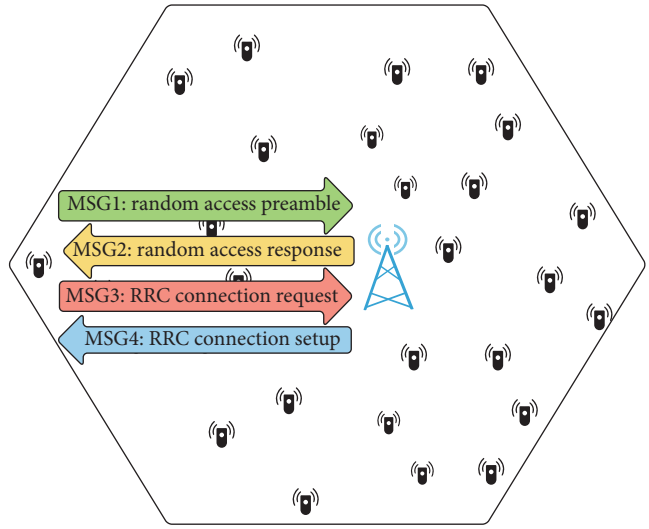


Figure 3: Contention-based RA procedure.

resources. On the other hand, the noncontention-based RA procedure is used for intrasystem handover and the arrival of the downlink data, where the access resources are assigned to the device by the eNB. In this study, our focus is on the contention-based approach, whereby the devices use the RA procedure to establish a radio link connection.

3.1. Contention-Based RA Procedure. As mentioned above, in our approach we use the contention-based RA procedure. It is a cross-layer procedure (i.e., MAC and physical layers) that deals with the logical, transport, and physical channels. The logical channels transfer data between the radio link control (RLC) and MAC sublayers (e.g., common control channel (CCCH)). The transport channels transfer data between the MAC and physical layers (e.g., RACH, downlink shared channel (DL-SCH), and uplink shared channel (UL-SCH)). However, the physical channels transfer data across the air interface (e.g., physical downlink control channel (PDCCH), PRACH, physical downlink shared channel (PDSCH), and physical uplink shared channel (PUSCH)).

The contention-based RA procedure messages pass through the mentioned channels. The contention RA procedure consists of four messages exchanged between the device and the eNB, as shown in Figure 3. RA procedure messages are described below.

(i) The first message (MSG1) is a random access preamble, whereby the device randomly selects a preamble out of 54 preambles, as assumed in [9], and sends the preamble to eNB. This message deals with RACH, which transfers the control information to PRACH. The device uses the transferred information to select a preamble and calculate the PRACH transmit power. It then transmits the preamble with a random access-radio network temporary identifier (RA-RNTI) on the PRACH to eNB in the next RAO. The RAOs are defined according to the PRACH configuration index, which is broadcasted within the system information

block two (SIB2). This step enables eNB to estimate the transmission time of the device to uplink the synchronization if there is no collision. A collision occurs if two or more devices send the same preamble to eNB in the same RAO, as defined in [9]. In this case, eNB will be unable to decode MSG1 from the collided devices; moreover, it will not respond to them with the random access response (RAR).

(ii) The second message (MSG2) is RAR, with which the eNB transmits the message to the device if there is no collision. This message includes a temporary cell-radio network temporary identifier (TC-RNTI) and a timing advance (TA) command (i.e., to adjust the device transmit timing). It assigns uplink resources to the device to be used in the third step. The device checks the PDCCH whose cyclic redundancy check (CRC) bits are scrambled by its RA-RNTI within the random access response window (RAR_{window}) to read the downlink control information (DCI) and obtain the downlink resource allocation information to identify the position of the RAR within the PDSCH. If the device does not find its PDCCH with its RA-RNTI, it means that either a collision occurred, as assumed in [9], or insufficient PDCCH resources are available.

(iii) The third message (MSG3) is an RRC connection request. Because we focus on the contention-based RA procedure for connection establishment, the device uses TC-RNTI to send the RRC connection request using signal radio bearer zero (SRB0) on CCCH. The data are then mapped onto UL-SCH, and uplink control information (UCI) is added to the outcome of the UL-SCH during physical layer processing for transfer to eNB using PUSCH. After sending MSG3, the device starts the contention resolution timer and awaits a response from eNB.

(iv) The fourth message (MSG4) is the RRC connection setup, wherein eNB sends MSG4 to the device using SRB0 on CCCH, which passes through DL-SCH using its TC-RNTI. The RRC connection setup message carries a cell-radio network temporary identifier (C-RNTI), which is used for further message exchange. The RA procedure is considered successful only if all steps are successfully completed. If the device does not receive a response within the *mac-Contention Resolution Timer*, then the device attempts to transmit a preamble again (but only if $Max_{PreamTrans}$ is not reached).

3.2. RACH Capacity Evaluation Metrics. The different measures that can be considered to evaluate the performance of RACH capacity for M2M communications are presented in a 3GPP report [9]. Here, we evaluate RACH by considering the collision probability under the unconditioned packet transmission. The knowledge of the collision probability is important for resource management.

The evaluation metrics used in this paper are the following [13]:

(i) Collision probability: the ratio of the number of occurrences in which a collision occurs to the overall number of opportunities (with or without access attempts) in the period.

(ii) Access success probability: the ratio between the number of devices successfully completing the RA procedure and the total number of devices.

(iii) Average access delay: the ratio between the total access delay time of the successful access devices and the time from the first RA procedure access to its successful completion when all devices successfully completed the RA procedure.

(iv) Average number of preamble transmissions: the ratio between the total number of preamble transmissions for all successful access devices and the total number of devices successfully completing the RA procedure within the maximum number of preamble transmissions.

4. Random Access Improvement Methods

As described above, we consider the contention-based RA procedure with a massive number of devices accessing the network within 10 s. This approach increases the contention on the RAOs and the PDCCH resources; moreover, it leads to a reduction of the access success probability. In [17], the authors indicated that RAN/core network (CN) resources are insufficient to meet the needs of all users and M2M-Ds. In this paper, we propose different methods to enhance RAN performance, as illustrated in Figure 4.

The first method increases the RAOs to increase the access resources by reconfiguring the PRACH configuration index. In the second method, we place several picocells in the macrocell range to increase PDCCH resources and reduce the traffic on eNB of the macrocell. In the last method, small networks are placed with aggregation points within the range of the macrocell. The evolution in 5G considers deploying aggregation points as one of the device access methods [30]. Those methods are presented in detail in the following subsections.

4.1. PRACH Configuration Index. The availability of RAOs relates to the PRACH configuration index. For example, if the configuration index is six, then there are two RAOs in each frame, as shown in Figure 5(a). By setting different values of this index, the availability of RAOs per frame changes. This fact has an intrinsic effect on the RACH performance. In Annex B of TR 37.868 by 3GPP, the RACH intensity is plotted against the required number of RAOs per second for a given collision probability of 1% [9]. They assumed that the arrival of RACH requests is uniformly distributed over time.

Meanwhile, the method in [18] uses 0, 3, 6, and 9 PRACH configuration index values to evaluate RACH performance of LTE with the assumption of fixing the initial number of simultaneous arrivals to a specific RA slot (i.e., RAO) without considering a traffic pattern for the simultaneous arrivals. The authors evaluate RACH with respect to the average access delay, blocking probability, average energy consumption, and

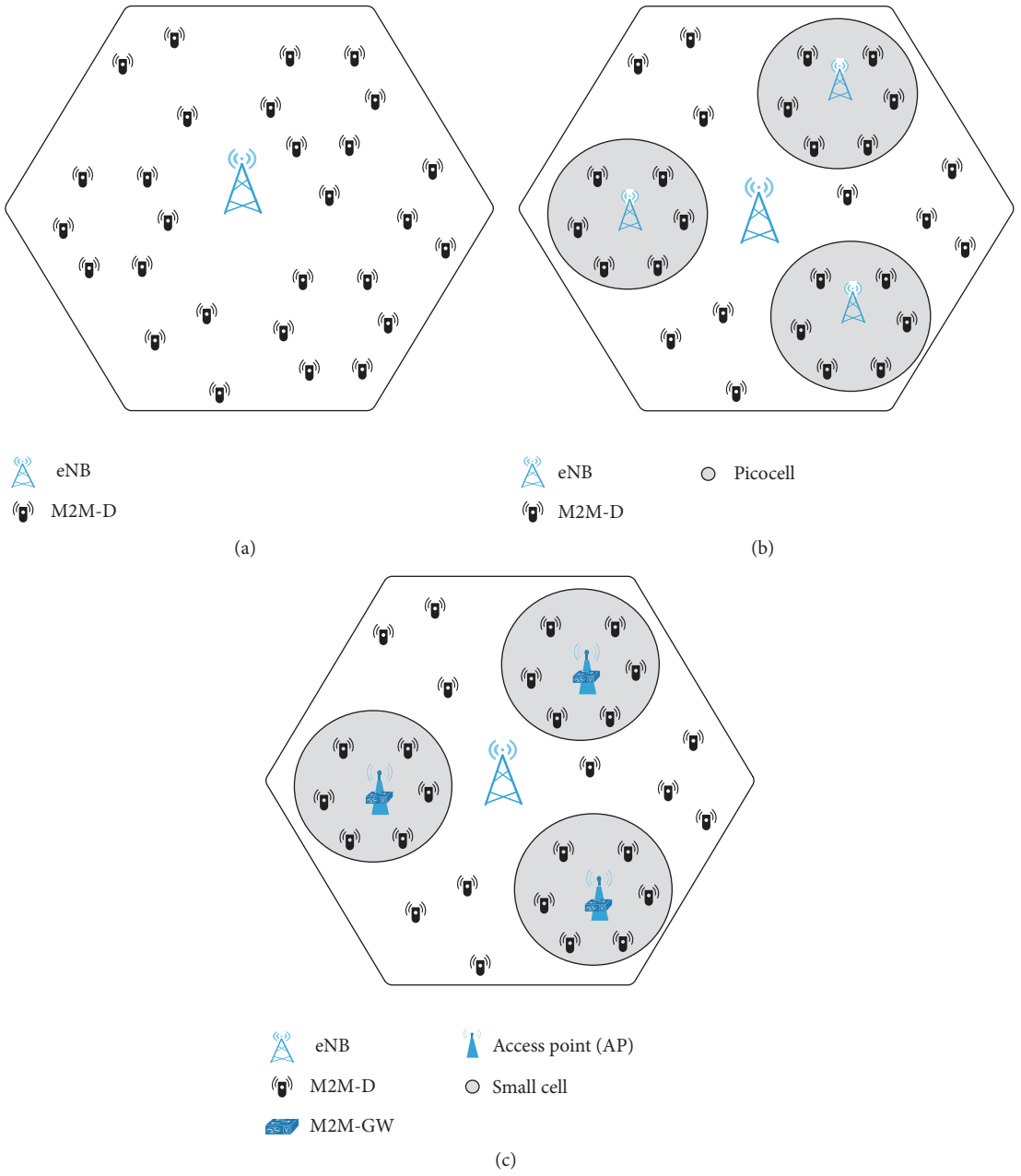


Figure 4: Proposed methods. (a) 3GPP scenario. (b) Picocells. (c) Aggregation points.

average number of preamble retransmissions. However, in this study, we investigate how the increase of RAOs affects the RACH capacity. Our study evaluates RACH performance under an extreme scenario (i.e., within 10 s), and the arrival of device access requests follow a beta distribution over time. To enhance RACH performance, we increase the RAOs per frame by setting the PRACH configuration index to 12. For this configuration, the availability of RAOs is five per frame, as shown in Figure 5(b).

4.2. Pico Cells. The primary role of heterogeneous networks is to provide more coverage and capacity (i.e., cover low-cost and low-power devices in coverage holes) [31]. For example, a large cell is covered by a macro base station, where femto access points (FAPs), pico base stations (PBSs), or relay stations (RSs) are used for coverage extension and capacity growth.

Because the given network elements improve network performance in terms of capacity and coverage, an enhancement of RAN performance is also expected. Therefore, we chose PBS on account of its advantages over the other networks elements. Moreover, PBS uses less power and costs less compared to MBS. In addition, it is accessible to all cellular devices because it is part of a network operator that deploys the public infrastructure and is controlled by the network operator, which aids in further management.

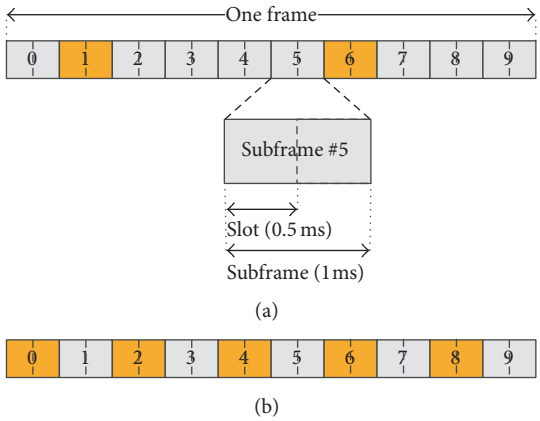


Figure 5: PRACH configuration index of frame structure type one. (a) PRACH configuration index six. (b) PRACH configuration index 12.

Furthermore, PBS transmissions are reliable and secure. In addition, placing PBSs in the area of MBS will increase access resources (i.e., preambles, PDCCH resources) that, in turn, will offload the traffic from the MBS to the PBSs, help to reduce MSG1 failures, and reduce the average access delay, especially in the case of many devices. Therefore, in our study, we place a different number of PBSs in the macrocell to improve RAN performance.

4.3. Aggregation Points. Involving aggregation points or M2M-GWs in an LTE-A system is being considered a solution to control RAN congestion in 5G systems [30]. It is also considered a radio access method for massive machine communications (MMC) [32]. The goal of using aggregation points is to provide interoperations with different wireless technologies [33]. In addition, deploying M2M-GW will help reduce device power consumption if it transmits through M2M-GW with low power [10].

In [5], an M2M-GW was introduced as an M2M-D access method to enable an efficient path for communication between devices. In [34], the authors proposed an architecture that supports the use of the M2M relay (M2M-R) as a data concentrator. The authors deployed an aggregation scheme in M2M-R, M2M-GW, and eNB. In addition, they proposed a possible design of M2M-GW in which the devices are linked to M2M-GW, which, in turn, is linked to eNB via an M2M-R. That study focused on data aggregation with a small number of devices (i.e., a maximum of 500 M2M-Ds). It showed a reduction in protocol overhead. In this paper, the aggregation point is used to gather the access requests of devices coming from the small network to which it belongs.

Two different scenarios for aggregation points are used for the access request under the extreme scenario (i.e., traffic model two in 3GPP [9]). In the first scenario, the aggregation point acts as a tunnel to pass device messages to and from eNB.

In the second scenario, we assume that the aggregation point is available for multipacket reception. The aggregation point has a behaviour that is similar to that of M2M-D in the RA procedure, whereby it shares the same access resources with M2M-Ds. The aggregation point collects device requests in each RAO and deals with the incoming requests as one request. The aggregator point refers to this request as a group request. Once the request of the group is granted, then the aggregation point grants the requests of the devices belonging to the same group. The devices in the granted group share the same uplink resources.

The aim of evaluating RAN performance in scenario one is to validate the implementation of aggregation points in our simulation for use in scenario two.

5. System Model and Assumptions

The system model accounts for the radio frame structure type one that is applicable to FDD. The M2M traffic arrival rate is assumed to follow a beta distribution (extreme scenario) with $\alpha = 3$ and $\beta = 4$. Under this scenario, numerous M2M-Ds attempt to access the network within 10 s in a highly synchronized manner [9]. A time domain random access structure of LTE is used. For statistically accurate results, an average of ten cells is deployed, each of which has a 1 km radius, which is taken as a typical size for a hexagon macrocell. The number of M2M-Ds in one macrocell is assumed to have the following values: 5,000, 10,000, 16,000, 20,000, 25,000, and 30,000.

The RA procedure was implemented using MATLAB. Our simulation results were validated in [13] with the 3GPP technical report [9]. The simulation parameters based on [9] are presented in Table 1.

We consider the limit of PDCCH resources that may cause an MSG2 failure. The RA configuration for the preamble format is zero, which will restrict the preamble length to 1 ms (T_{MSG1}). As mentioned in Section 3.1, the contention-based RA procedure has a total of 54 preambles (N_{preamble}). The use of PRACH configuration index six involves use of an RAO every half frame, as shown in Figure 5(a). Therefore, the total number of RAOs for the extreme scenario (over 10 s) is 2,000. Every activated device randomly sends a preamble (i.e., MSG1) within a maximum of ten preamble transmission attempts ($\text{Max}_{\text{PreamTrans}}$). Then, eNB processes MSG1 to check whether a preamble collision exists [9]. If there is no collision, eNB sends an RAR (i.e., MSG2) to the device within 3 ms (T_{MSG2}). Otherwise, the collided devices attempt access again after a period of time (i.e., T_{MSG2} + $\text{RAR}_{\text{window}}$ + the time uniformly selected by the device within BO) for a new RAO with a new preamble, as long as the number of preamble transmission attempts (i.e., $\text{Counter}_{\text{PreamTrans}}$) does not exceed $\text{Max}_{\text{PreamTrans}}$.

For simplicity, the ramping procedure, which is used to increase the power of the device after each retransmission, is implemented in this simulation as a function of $(1 - e^{-i})$ to describe the probability of a successful preamble transmission, where i represents the number of times the device transmits preambles ($\text{Counter}_{\text{PreamTrans}}$) [9]. The position of RAR for the granted devices is assigned through PDCCH within the $\text{RAR}_{\text{window}}$ [35]. It is assumed that in each RAR there are three uplink grants. The simulation assumes 16 common control elements (CCEs), where the aggregation level is four (i.e., the PDCCH format is two). Therefore, the

Table 1: Simulation parameters.

Symbol	Parameter	Value
B	Cell bandwidth	5 MHz
—	PRACH configuration index	6
N_{preamble}	Total number of preambles	54
$\text{Max}_{\text{PreamTrans}}$	Maximum number of preamble transmissions	10
—	Number of UL grants per RAR	3
—	Number of CCEs allocated for PDCCH	16
—	Number of CCEs per PDCCH	4
$\text{RAR}_{\text{window}}$	*RA-Response Window Size*	5 ms
—	*mac-Contention Resolution Timer*	48 ms
BO	Backoff timer	20 ms
—	Probability of successful delivery for both MSG3 and MSG4	90%
$\text{MaxRetrans}_{\text{HARQ}}$	Maximum number of HARQ transmissions for MSG3 and MSG4 (nonadaptive HARQ)	5
M	Number of MTC devices ($\times 10^3$)	5, 10, 16, 20, 25, 30
T_u	Number of available subframes over the distribution period	10,000
b	Periodicity of PRACH opportunities	5 ms
T_{MSG1}	MSG1 transmission time	1 ms
T_{MSG2}	Preamble detection at eNB and MSG2 transmission time	3 ms
T_{MSG3}	Device processing time before sending MSG3	5 ms
$T_{\text{TransMSG3}}$	MSG3 transmission time	1 ms
T_{MSG4}	Time of processing MSG3 and sending MSG4	5 ms

number of PDCCH candidates is four. As a result, 12 devices are granted per RAO, and the remainder will again attempt access after a period of time (i.e., T_{MSG2} + $\text{RAR}_{\text{window}}$ + the time uniformly selected by the device within BO) for a new RAO, unless $\text{Max}_{\text{PreamTrans}}$ is reached.

For the devices that successfully receive the RAR, they process their RRC connection request (i.e., MSG3) in 5 ms (T_{MSG3}). After that, the devices transmit MSG3 and wait for the RRC connection setup (i.e., MSG4) within 5 ms (T_{MSG4}). The probability of successful delivery is 90% for both MSG3 and MSG4 [9]. The device that fails to deliver MSG3 or receive MSG4 attempts to resend the failure message (i.e., MSG3 or MSG4) with a maximum of five retransmissions ($\text{MaxRetrans}_{\text{HARQ}}$).

It is assumed that the retransmission of MSG3 and MSG4 is a nonadaptive hybrid automatic repeat request (nonadaptive HARQ). This model was validated against the 3GPP technical report [9] and insignificant differences were found between the two in [13]. Modifications to the system model for adaptation to the proposed approaches are presented in the next subsections.

5.1. PRACH Configuration Index. To consider the PRACH configuration index approach in our system model, we must set the PRACH configuration index to 12. This is accomplished by configuring the RAOs in subframes—0, 2, 4, 6, and 8—where the RAOs increase to reach five RAOs in each frame, as shown in Figure 5(b). As a result, the total number of RAOs for the extreme scenario (over 10 s) is 5,000.

5.2. Pico Cells. To deploy picocells in a macrocell, we must consider different issues. It is important to know where to locate PBS to achieve good coverage extension, the required number of picocells to enhance RACH performance, and the strategy of devices to join PBS. In [36], the authors refer to the importance of increasing the distance between MBS and PBS to improve system performance. Therefore, in our system model, we consider a picocell of a 100 m radius that is placed 750 m away from MBS to achieve good coverage for edge cell devices. Please note that this is a simple assumption to evaluate RACH performance and not an optimum PBS placement, which is beyond the scope of this paper.

In our simulation, we evaluated the RAN performance with 3 and 15 picocells, as shown in Figure 6. Each picocell has its own set of preamble sequences to help reduce collisions (i.e., reducing MSG1 failures). Additionally, each PBS has its own PDCCH resources that increase the number of granted devices. The devices located in the range of the picocell connect through its PBS.

5.3. Aggregation Points. The same assumptions for the picocells are assumed for the aggregation points to enable a consistent comparison between them. Therefore, we follow the picocells scenario by assuming that small networks exist in the same location of the picocells. For those small networks, regardless of their used technology, an aggregation point is placed on the access point. This is used to aggregate device access requests for the devices that are located within the area of the small network. The only condition of the technology used in the small networks is that the coverage of the small cell must support M2M-GW with a good signal quality on the M2M-GW to MBS link. The only difference between them placing PBSs or M2M-GWs is that the M2M-GWs will share the preambles and PDCCH resources with the MBS.

6. RACH Evaluation

The RACH evaluation was conducted with different density values: 5,000, 10,000, 16,000, 20,000, 25,000, and 30,000 [9, 13]. According to the different device density values, we assumed that the devices were uniformly distributed in the range of 50 to 1,000 m from the centre of the macrocell.

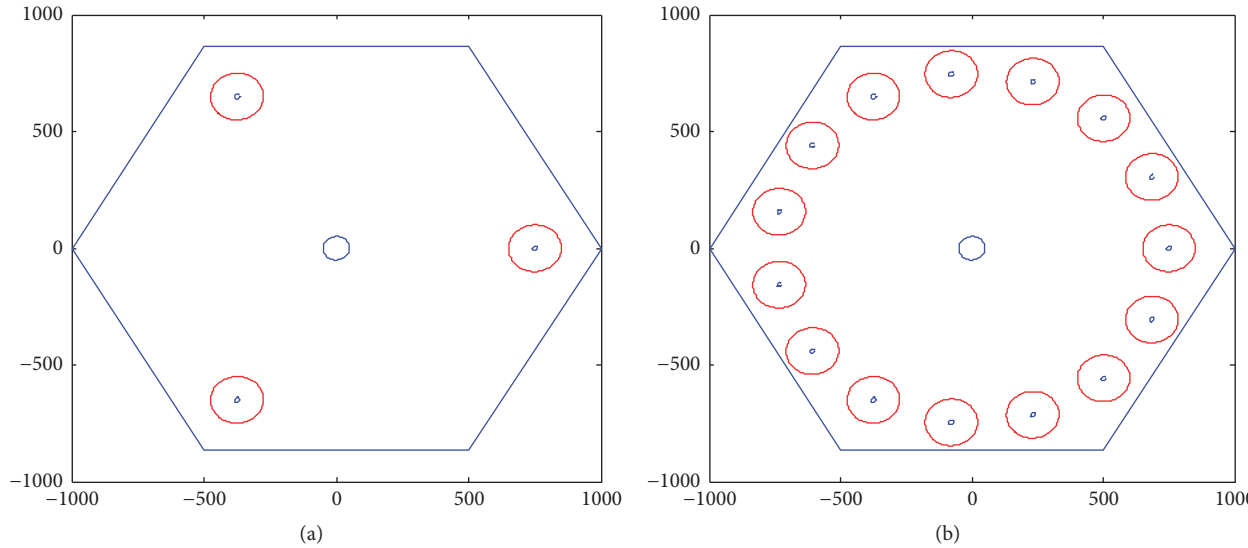


Figure 6: Small cell distribution. (a) Three deployed PBSs. (b) Fifteen deployed PBSs.

Owing to the different case studies considered herein, we refer to these cases as follows:

(i) *3GPP-compl. sim.*: 3GPP-compliant simulation that has only one macrocell.

(ii) *PRACH config. index*: 3GPP-compliant simulation with PRACH configuration index 12.

(iii) *3 picocells*: 3GPP-compliant simulation with 3 picocells.

(iv) *15 picocells*: 3GPP-compliant simulation with 15 picocells.

(v) *3 agg. points (1:1)*: 3GPP-compliant simulation with 3 aggregation points (1:1), where the aggregation points act as a tunnel.

(vi) *15 agg. points (1:1)*: 3GPP-compliant simulation with 15 aggregation points (1:1), where the aggregation points act as a tunnel.

(vii) *3 agg. points (1:k)*: 3GPP-compliant simulation with 3 aggregation points (1:k), where the aggregation points aggregate device requests in each RAO.

(viii) *15 agg. points (1:k)*: 3GPP-compliant simulation with 15 aggregation points (1:k), where the aggregation points aggregate device requests in each RAO.

(ix) *3 picocells + PRACH config. index*: 3 picocells combined with PRACH configuration index 12.

(x) *15 picocells + PRACH config. index*: 15 picocells combined with PRACH configuration index 12.

(xi) *3 agg. points (1:1) + PRACH config. index*: 3 aggregation points combined (1:1) with PRACH configuration index 12.

(xii) *15 agg. points (1:1) + PRACH config. index*: 15 aggregation points (1:1) combined with PRACH configuration index 12.

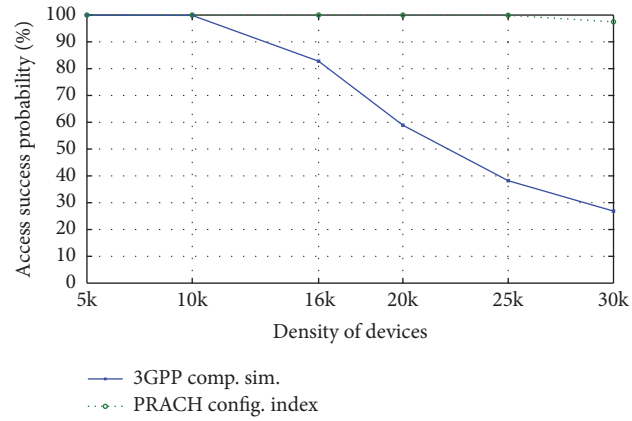


Figure 7: Access success probability for 3GPP-compliant simulation versus PRACH configuration index scenario.

(xiii) *3 agg. points (1:k) + PRACH config. index*: 3 aggregation points (1:k) combined with PRACH configuration index 12.

(xiv) *15 agg. points (1:k) + PRACH config. index*: 15 aggregation points (1:k) combined with PRACH configuration index 12.

6.1. RACH Analysis Results. As shown in Figures 7 and 8, it is clear that RACH performance in the PRACH configuration index scenario outperforms RACH performance in the 3GPP-compliant simulation scenario. The increase of RAOs in the PRACH configuration index scenario has a significant effect on the evaluation metrics. The access success probability for most of the density values approaches 100% with at most 48 ms of an average access delay.

In addition, as shown in Figure 9, the average number of preamble transmissions for all density values does not exceed 2.6, which explains the reason behind the reduction

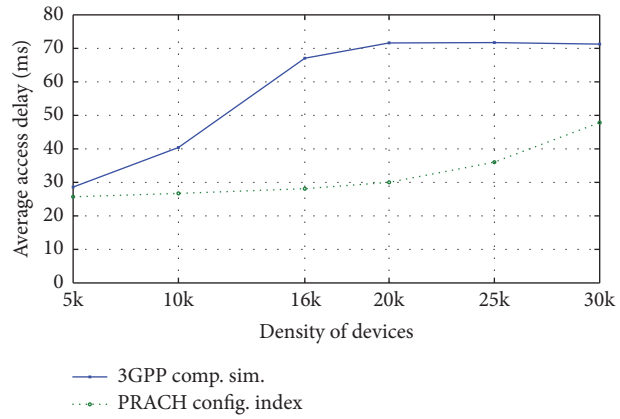


Figure 8: Average access delay for 3GPP-compliant simulation versus PRACH configuration index scenario.

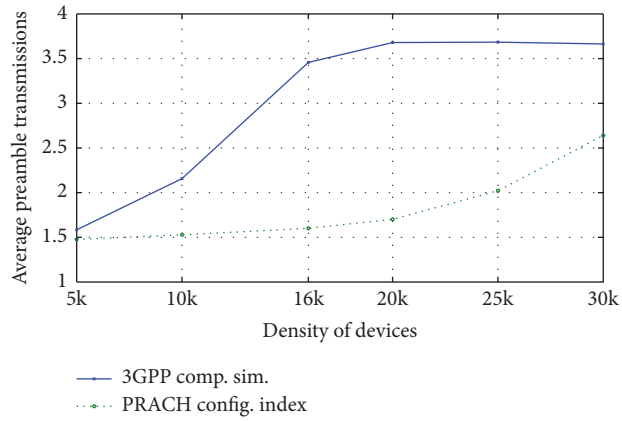


Figure 9: Average preamble transmissions for 3GPP-compliant simulation versus PRACH configuration index scenario.

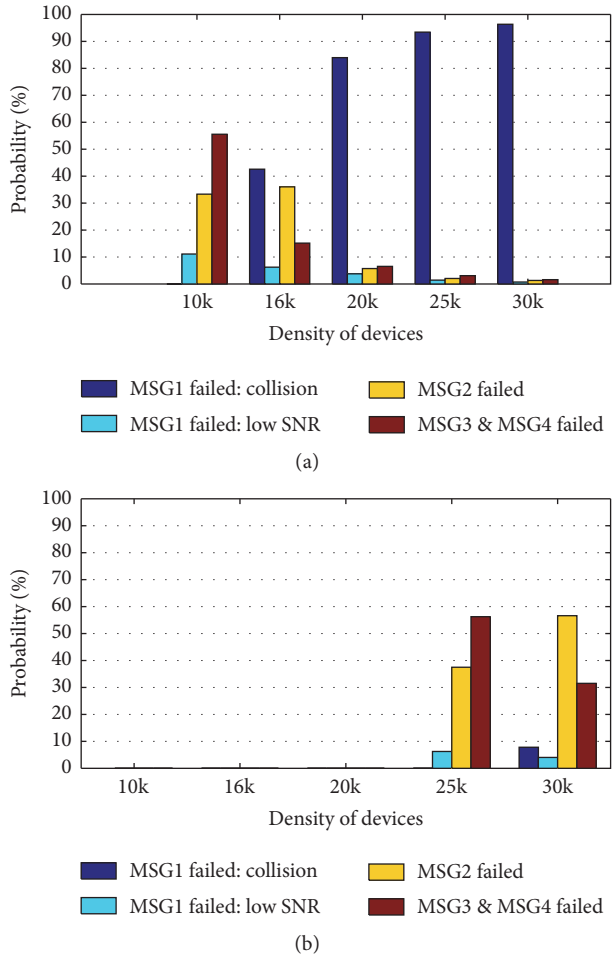


Figure 10: Access failure analysis describing the different reasons of access failure. (a) Scenario of 3GPP-compliant simulation. (b) PRACH configuration index scenario.

in the average access delay. Figure 10 illustrates the analysis of the access failure for both scenarios showing the percentage for each reason of the access failure probability. For 5,000 devices, the access success probability in all scenarios is 100%. Therefore, this density value is excluded from the analysis.

In a 3GPP-compliant simulation, for high-density values, the main reason for the RACH failure is the failure of MSG1 because of a preamble collision, as shown in Figure 10(a). For example, in the case of 30,000 devices, the access failure probability is 73.14%. Out of this access failure probability, 96.36% of the devices failed on account of an MSG1 preamble collision, 0.72% are due to MSG1 having a low signal to noise ratio (SNR) because those devices are located on the cell edge, 1.31% are due to MSG2 lacking PDCCH resources, and 1.61% are due to MSG3 and MSG4 failures. MSG3 and MSG4 failed on account of the system model assumption, where the probability of an unsuccessful delivery for both MSG3 and MSG4 is 10%.

For the PRACH configuration index scenario, the only density values with a RACH failure were 25,000 and 30,000 devices, as shown in Figure 10(b). The main reason for the RACH failure in this scenario for the case of 25,000 was the failures of both MSG3 and MSG4. In the case of 30,000 devices, the main reason for the RACH failure was that MSG2 failed on account of the shortage of PDCCH resources. It is furthermore evident in Figure 11 that the collision probability is reduced because there are more RAOs in the PRACH configuration index scenario. We conclude that, by employing the PRACH configuration index scenario, we can achieve a high access success probability with a low average access delay.

In comparing the results of the 3GPP-compliant simulation with the results of the picocell approach in Figure 12, it is obvious that the RACH performance in the picocell approach outperforms the 3GPP-compliant simulation scenario in terms of the access success probability. Referring to Figure 10, the main reason that the access fails in the 3GPP-compliant simulation is the failure of MSG1 on account of the preamble collision. The approaching picocells increase the number of preambles and PDCCH resources. This has an important effect on improving RACH performance. In addition, the role of the picocells to offload the traffic from the macrocell has a significant effect. However, because of the limited coverage

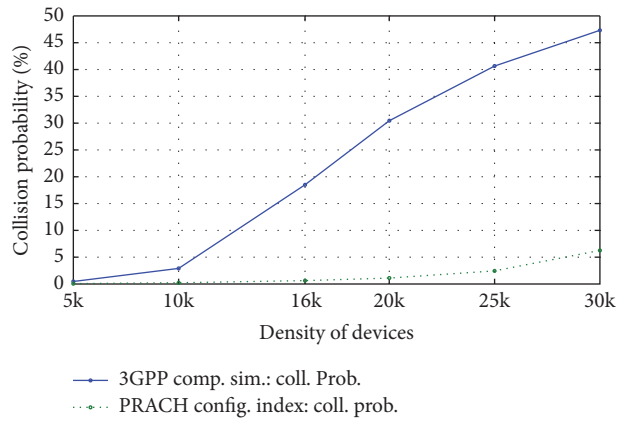


Figure 11: Collision probability with respect to RAOs for 3GPP-compliant simulation versus PRACH configuration index scenario.

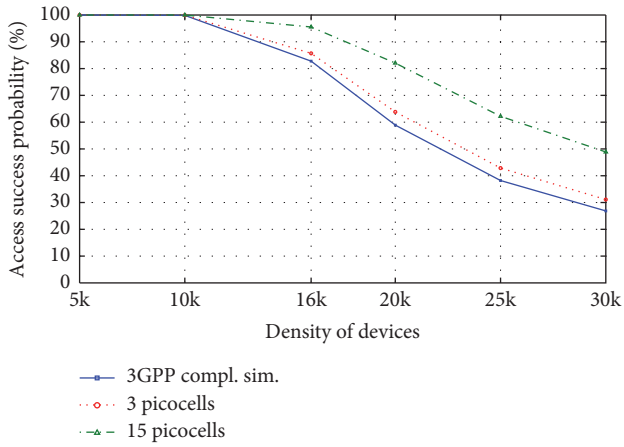


Figure 12: Access success probability for 3GPP-compliant simulation versus 3 and 15-picocell scenarios.

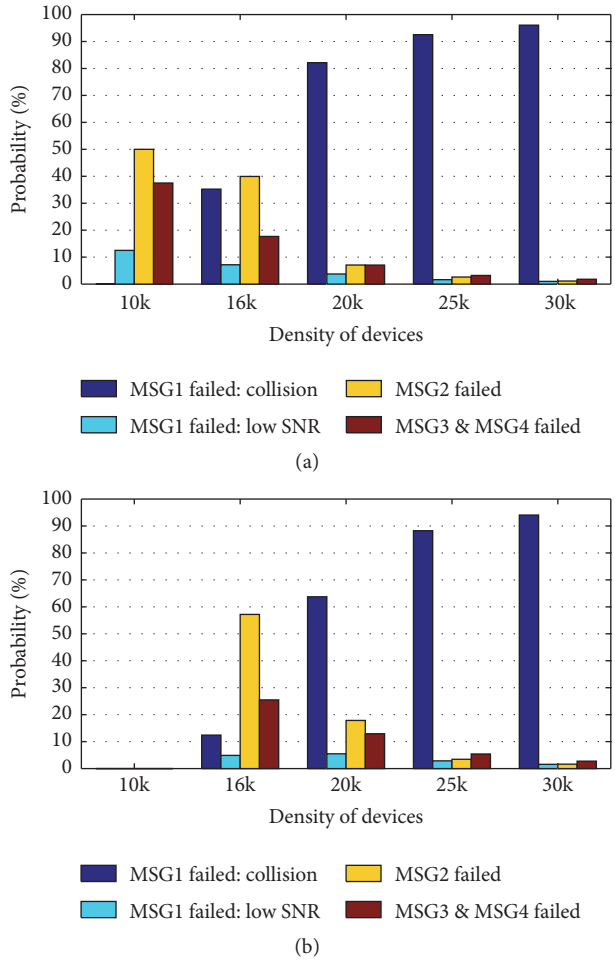


Figure 13: Access failure analysis describing the different causes of access failure. (a) Three-picocell scenario. (b) Fifteen-picocell scenario.

of the picocells (i.e., limited in its ability to host a large number of devices because of the assumed picocoverage), not all devices will obtain the benefits of the deployed picocells.

However, the picocell approach improves the RACH performance and increases the access success probability for all ranges of density values in both cases (i.e., 3 and 15 picocells), as shown in Figure 12.

In the three picocells scenario, the increase of the access success probability is small. However, in the 15-picocell scenario, the access success probability is substantially increased compared to the 3GPP-compliant simulation. The analysis of the failure of access for both scenarios is shown in Figure 13. In the three picocells scenario, if the density of devices is less than or equal to 16,000, the access success probability is high. The main reason there is a RACH failure is the lack of PDCCH resources, which can cause an MSG2 failure (three picocells are not adequate).

For the higher density values, the main causes of the access failure are the collisions in MSG1 on account of the high density of devices attempting access in a short period of time. In the 15-picocell scenario, the access success probability is 100% for the 5,000 and 10,000 cases. For the 16,000 devices, the access failure probability is 4%. The main cause of the RACH failure is again the MSG2 failure on account of the lack of PDCCH resources. However, as the density of devices increases to 30,000 devices, the prime cause of failure is again the collision in MSG1.

In this study, we additionally investigated how the picocell approach affected the average access delay. In this approach, the average access delay was reduced compared to the performance in the 3GPP-compliant simulation, as shown in Figure 14. The same observation was made in terms of the average preamble transmissions, as depicted in Figure 15.

For the scenario of 3 and 15 aggregation points (1:1), the RACH access success probability was similar to that of the 3GPP-compliant simulation, as shown in Figure 16 (the same was observed for the other metrics as well). This result was expected because the role of aggregation points in these scenarios is to pass access devices request to eNB without accumulating requests. We used the mentioned scenarios to verify the implementation of aggregation points in scenarios of 3 and 15 aggregation points (1:k).

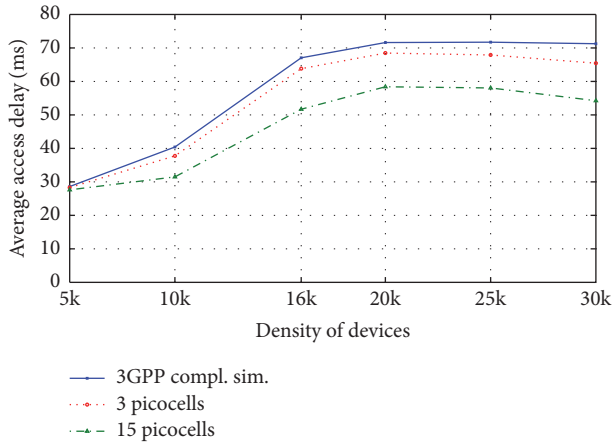


Figure 14: Average access delay for the 3GPP-compliant simulation versus 3- and 15-picocell scenarios.

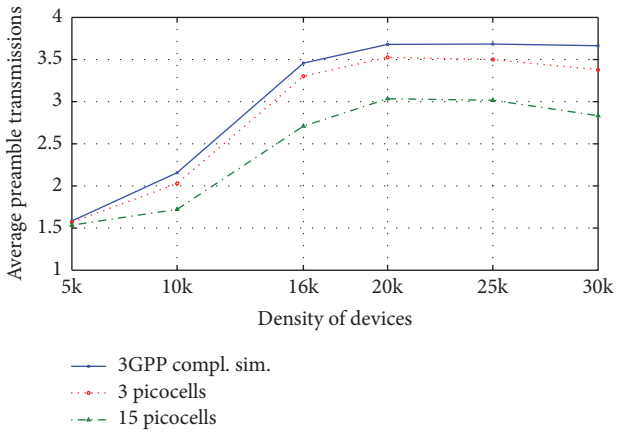


Figure 15: Average preamble transmissions for the 3GPP-compliant simulation versus 3- and 15-picocell scenarios.

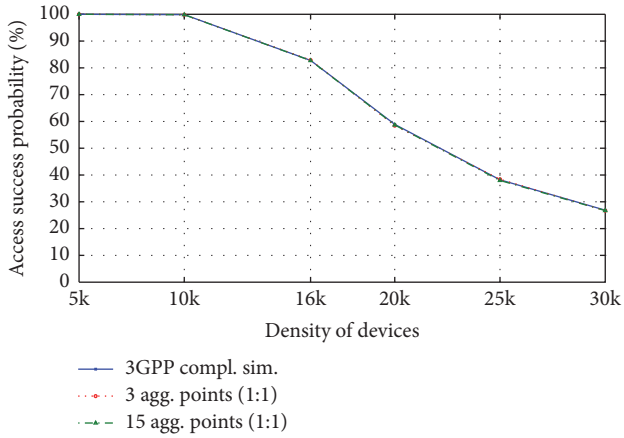


Figure 16: Access success probability for the 3GPP-compliant simulation versus 3 and 15 aggregation points (1:1).

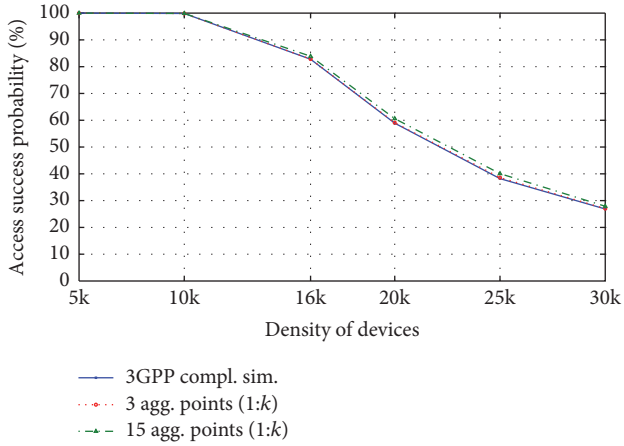


Figure 17: Access success probability for the 3GPP-compliant simulation versus 3 and 15 aggregation points (1:k).

In the aggregation point approach, the aggregation point collected or aggregated the request of the devices in each RAO (i.e., 3- and 15-aggregation-point (1:k) scenarios). The results showed a slight improvement in terms of access success probability only in the scenario of 15 aggregation points (1:k) (Figure 17). This was the case because of the very small reduction of the collision probability since the role of the aggregation point is to group the device requests in the same RAO and send them as one request. This causes a slight reduction in the average number of preamble transmissions.

However, it is important to note that the aggregation points did not perform as well because of the small amount of aggregated requests and owing to the traffic pattern, where the arrival of devices followed a beta distribution (i.e., a maximum of four requests). The remaining results of the 3- and 15-aggregation point (1:k) scenarios were similar to those of the 3GPP-compliant simulation. Thus, the figures are not included.

The analysis of RACH failure for scenarios of 3 and 15 aggregation points (1:k) was slightly different compared to the analysis in the 3GPP-compliant simulation (Figure 18). For 10,000 devices, the probability of access failure was very low. For the case of 16,000 devices, the cause of the RACH failure was the failure of MSG2. This was different from the analysis of the failure in the 3GPP-compliant simulation scenario, where the high percentage of RACH failures was because of collisions. For the high-density values, the main causes of the RACH failure were the collisions, as in the 3GPP-compliant simulation. In the scenarios of 3 and 15 aggregation points (1:k), the failure due to collisions decreased compared to that of the 3GPP-compliant simulation. However, this resulted in an increase in the contention of the devices on the eNB requesting PDCCH resources, which led to the MSG2 failure.

In this study, we additionally evaluated a combination of the proposed methods (the PRACH configuration index scenario combined with picocell and aggregation point approaches) (Table 2). The table includes the numerical results of the 3GPP-compliant simulation and the PRACH configuration index scenarios for a comparison with the previously discussed results for the latter scenarios.

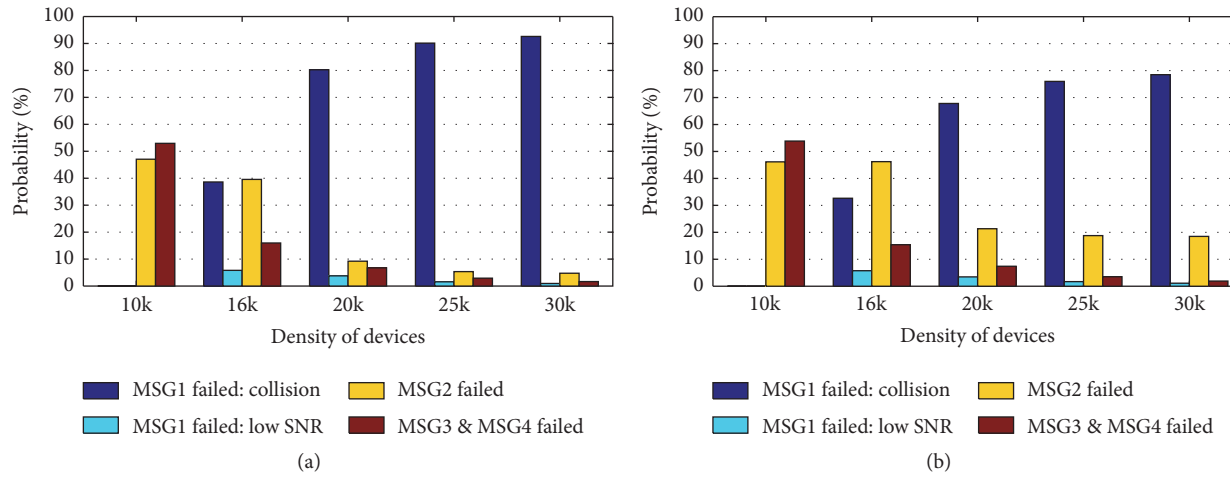


Figure 18: Access failure analysis describing the different causes of access failure. (a) Three-aggregation-point (1:k) scenario. (b) Fifteen-aggregation-point (1:k) scenario.

Table 2: RACH performance for 3GPP-compliant simulation and the proposed solutions.

Evaluation metrics	M	3GPP compl. sim.	PRACH config. index	3 picocells + PRACH config. index	15 picocells + PRACH config. index	3 agg. points (1:1) + PRACH config. index	15 agg. points (1:1) + PRACH config. index	3 agg. points (1:k) + PRACH config. index	15 agg. points (1:k) + PRACH config. index
Access success probability (%)	5k	100	100	100	100	100	100	100	100
	10k	99.86	100	100	100	100	100	100	100
	16k	82.78	100	100	100	100	100	100	100
	20k	58.90	100	100	100	100	100	100	100
	25k	38.21	99.91	99.95	100	99.92	99.91	99.91	99.92
	30k	26.86	97.44	98.53	99.94	97.39	97.41	97.51	97.69
Average access delay (ms)	5k	28.61	25.70	25.49	25.2	25.55	25.49	25.36	25.73
	10k	40.41	26.71	26.44	25.87	26.62	26.61	26.77	26.56
	16k	67.05	28.11	27.85	26.79	27.9	28	27.85	27.89
	20k	71.62	30.01	29.29	27.61	30.18	30.07	29.9	29.91
	25k	71.71	36.05	34.36	29.41	35.75	35.92	36.05	35.64
	30k	71.26	47.86	44.39	33.47	47.76	47.68	47.55	47.27
Average number of preamble transmissions	5k	1.58	1.48	1.46	1.45	1.47	1.47	1.46	1.48
	10k	2.16	1.53	1.51	1.48	1.52	1.52	1.53	1.52
	16k	3.46	1.60	1.59	1.53	1.59	1.6	1.59	1.58
	20k	3.68	1.70	1.67	1.58	1.71	1.71	1.7	1.69
	25k	3.68	2.02	1.93	1.67	2	2.01	2.02	1.98
	30k	3.66	2.64	2.46	1.89	2.64	2.63	2.62	2.59

In the combined approach of the picocell and PRACH configuration index, the advantage of the PRACH configuration index approach (i.e., the increase of the RAOs) supplements the advantages of the picocell approach (i.e., additional preambles and the offloading feature). As evident in Table 2, the results for this scenario outperform the results of all the previous scenarios for access success probability, average access delay, and average preamble transmissions. Note that, by following the combined approaches of the PRACH configuration index and picocells, the access success probability is approximately 100% for all density values, except for the 30,000 devices, in the case of using three picocells in the combined picocell and PRACH configuration index approach.

For the remaining scenarios, in which the PRACH configuration index is combined with the aggregation point approach, the RACH performance shown in Table 2 is similar to the performance of scenario two, as expected from the previous results.

A comparison of the best results of RACH performance for the proposed schemes in [9, 14, 15, 23] that considers traffic model two with 30,000 devices and select methods of the

Table 3: Comparison results of proposed and existing methods.

Scheme	Access success prob. (%)	Avg. access delay (ms)	Avg. preamble trans.
Longer backoff scheme (max. 960) [14]	100	—	1.86
Prebackoff scheme (max. 960) [14]	100	—	1.51
ACB [14]:			
M2M ACB factor = 0.5	68.98	—	—
ACB time = 16			
Separate RACH resources [14]:	11.46	—	—
M2M/UE is 53/1			
Dynamic allocation of RACH resources [14]:			
100% additional subframes for M2M dedicated access	21.26	—	—
PRA [15]	93.9	2937	—
Some methods proposed in this work			
PRACH config. index	97.44	47.86	2.64
15 picocells	48.88	54.25	2.83
15 agg. points (1:k)	27.89	71.10	3.60
15 picocells + PRACH config. index	99.94	33.47	1.89
15 agg. points (1:k) + PRACH config. index	97.69	47.27	2.59

proposed schemes of this work are presented in Table 3. As shown in the table, the longer backoff scheme and prebackoff scheme proposed in [14] have the highest access success probability and low average preamble transmissions compared to the other schemes. However, there are no numerical results for the access delay because those schemes are only proposed to solve RAN congestion for delay-tolerant devices. In this study, our approaches (i.e., PRACH configuration index, 15 picocells with the PRACH configuration index, and 15 aggregation points with the PRACH configuration index) can serve different M2M applications with an acceptable average access delay, high access success probability, and low average preamble transmissions.

6.2. Discussion. Our results show that the PRACH configuration index approach substantially improves RACH performance of access success probability, average access delay, average preamble transmissions, and collision probability on account of the increase of the RAOs. This approach is suitable for nondelay tolerant M2M applications because of its advantages. On the other hand, because the RA procedure uses six resource blocks in each subframe, 12.5% of the uplink resources in a 5 MHz bandwidth are consumed once the PRACH configuration index is set to 12 [9]. Nevertheless, we believe this approach can be used without sacrificing the service quality of the upload transmission, especially if we switch to a higher bandwidth (e.g., 20 MHz, where the number of resource blocks is 100) because most of the M2M applications consider small data transmissions. This approach is applicable to general M2M service requirements, such as subscription management, adding or removing M2M characteristics, or controlling traffic. In this approach, the network operator controls MBS, which, in turn, manages the cellular M2M-Ds.

The picocell approach performs well, particularly in terms of access success probability, average access delay, and average preamble transmissions for all density values if the number of deployed picocells is increased. This result is on account of the increased number of preambles, availability of PDCCH resources, and reduced traffic on eNB, which can effectively improve congestion and enhance RACH performance. However, there is an associated cost with introducing additional picocells. This approach is applicable to general M2M service requirements because PBS is likewise controlled by the network operator.

Deploying a large number of aggregator points to collect device requests in M2M architecture does not considerably enhance RACH performance in this scenario.

As expected in the combination approaches, RACH performance is improved. In our analysis, the most promising solution that achieves high access success probability, low average access delay, and low average preamble transmissions is the case of 15 picocells combined with the PRACH configuration index.

7. Conclusions

This paper provides an analysis of the RA procedure for M2M communications over LTE-A. The focus of this study was an extreme scenario with a heavy density of devices attempting to access the network in a short period of time and in a synchronized manner. In this paper, we proposed three methods to improve RACH capacity performance. The PRACH configuration index approach achieved a significant improvement in RACH performance for all cases including a massive number of devices in terms of access success probability, average access delay, and average preamble transmissions.

A significant reduction in the collision probability compared to the 3GPP-compliant simulation was additionally determined. The picocell approach with 15 picocells enhanced RACH performance in terms of access success probability, average access delay, and average number of preamble transmissions. For the case of aggregation points, only a very slight enhancement was observed for the number of aggregation points investigated. The method that combined the PRACH configuration index with picocells performed better than all methods. In short, deploying any of the mentioned approaches depends on different issues, such as the type of M2M application and deployment costs.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

Fatemah Alsewaidi would like to thank the Public Authority for Applied Education and Training (PAAET), Kuwait, for sponsoring her Ph.D. studies.

References

[1] ETSI TS 102 690 V2.1.1, "Machine-to-Machine Communications (M2M): Functional Architecture," October 2013.

[2] ETSI TS 102 689 V1.1.1, "Machine-to-Machine Communications (M2M): M2M Service Requirements," August 2010.

[3] 3GPP TR 23.888 V11.0.0, "System Improvements for Machine-Type Communications (MTC)," September 2012.

[4] 3GPP TS 22.368 V13.1.0, "Service Requirements for Machine-Type Communications (MTC); Stage 1," December 2014.

[5] K. Zheng, F. L. Hu, W. B. Wang, W. Xiang, and M. Dohler, "Radio resource allocation in LTE-advanced cellular networks with M2M communications," *IEEE Communications Magazine*, vol. 50, no. 7, pp. 184–192, 2012.

[6] "Standardization of Machine-type Communications, V0.2.4," June 2014.

[7] Cisco Visual Networking Index, "Cisco visual networking index: global mobile data traffic forecast update, 2014–2019," White Paper, 2015, http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html.

[8] 3GPP TSG RP-141865, "Revised WI: Further LTE Physical Layer Enhancements for MTC," RAN Meeting no. 66, Ericsson, Edinburgh, Scotland, September 2014.

[9] 3GPP TR 37.868 V11.0.0, "Study on RAN Improvements for Machine-type Communications," September 2011.

[10] H. Shariatmadari, R. Ratasuk, S. Iraji et al., "Machine-type communications: current status and future perspectives toward 5G systems," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 10–17, 2015.

[11] 3GPP TSG R2-112198, "Clarification on the Discussion of RACH Collision Probability," ITRI 3GPP TSG-RAN WG2 Meeting no. 73b, Shanghai, China, April 2011.

[12] R.-G. Cheng, C.-H. Wei, S.-L. Tsao, and F.-C. Ren, "RACH collision probability for machine-type communications," in *Proceedings of the IEEE 75th Vehicular Technology Conference (VTC '12)*, pp. 1–5, IEEE, Yokohama, Japan, June 2012.

[13] F. Alsewaidi, D. Kaleshi, and A. Doufexi, "Analysis of radio access network performance for M2M communications in LTE-A at 800 MHz," in *Proceedings of the IEEE Wireless Communications and Networking Conference Workshops (WCNCW '14)*, pp. 110–115, Istanbul, Turkey, April 2014.

[14] 3GPP TSG R2-112863, "Backoff Enhancements for RAN Overload Control," 3GPP TSG RAN WG2 no. 73bis, Barcelona, Spain, May 2011.

[15] J.-P. Cheng, C.-H. Lee, and T.-M. Lin, "Prioritized Random Access with dynamic access barring for RAN overload in 3GPP LTE-a networks," in *Proceedings of the IEEE GLOBECOM Workshops (GC Wkshps '11)*, pp. 368–372, December 2011.

[16] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 66–74, 2011.

[17] M.-Y. Cheng, G.-Y. Lin, H.-Y. Wei, and A. C.-C. Hsu, "Overload control for machine-type-communications in LTE-advanced system," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 38–45, 2012.

[18] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 4–16, 2014.

[19] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: issues and approaches," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 86–93, 2013.

[20] M.-Y. Cheng, G.-Y. Lin, H.-Y. Wei, and C.-C. Hsu, "Performance evaluation of radio access network overloading from machine type communications in LTE—a networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference Workshops (WCNCW '12)*, pp. 248–252, April 2012.

[21] C.-Y. Oh, D. Hwang, and T.-J. Lee, "Joint access control and resource allocation for concurrent and massive access of M2M devices," *IEEE Transactions on Wireless Communications*, vol. 14, no. 8, pp. 4182–4192, 2015.

[22] F. Ghavimi and H.-H. Chen, "M2M communications in 3GPP LTE/LTE-A networks: architectures, service requirements, challenges, and applications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 525–549, 2015.

[23] 3GPP TSG R2-104662, "MTC Simulation Results with Specific Solutions," RAN WG2 no. 71, Madrid, Spain, August 2010.

[24] J. S. Vardakas, N. Zorba, C. Skianis, and C. V. Verikoukis, "Performance analysis of M2M communication networks for QoS-differentiated smart grid applications," in *Proceedings of the IEEE Globecom Workshops (GC Wkshps '15)*, pp. 1–6, IEEE, San Diego, Calif, USA, December 2015.

[25] G.-Y. Lin, S.-R. Chang, and H.-Y. Wei, "Estimation and adaptation for bursty LTE random access," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2560–2577, 2016.

[26] Y. Chang, C. Zhou, and O. Bulakci, "Coordinated random access management for network overload avoidance in cellular machine-to-machine communications," in *Proceedings of the European Wireless, 20th European Wireless Conference*, pp. 1–6, Barcelona, Spain, 2014.

[27] S.-Y. Lien, T.-H. Liau, C.-Y. Kao, and K.-C. Chen, "Cooperative access class barring for machine-to-machine communications," *IEEE Transactions on Wireless Communications*, vol. 11, no. 1, pp. 27–32, 2012.

[28] G. Foddis, R. G. Garroppo, S. Giordano, G. Procissi, S. Roma, and S. Topazzi, "On RACH preambles separation between human and machine type communication," in *Proceedings of the IEEE International Conference on Communications (ICC '16)*, Kuala Lumpur, Malaysia, 2016.

[29] F. Alsewaidi, D. Kaleshi, and A. Doufexi, "Performance comparison of LTE-A RAN operating in 800MHz and 2.4GHz bands for M2M communications," in *Proceedings of the 11th International Symposium on Wireless Communications Systems (ISWCS '14)*, pp. 824–829, Barcelona, Spain, August 2014.

[30] R. Ratasuk, A. Prasad, Z. Li, A. Ghosh, and M. Uusitalo, "Recent advancements in M2M communications in 4G networks and evolution towards 5G," in *Proceedings of the 18th International Conference on Intelligence in Next Generation Networks (ICIN '15)*, pp. 52–57, IEEE, Paris, France, February 2015.

[31] S.-P. Yeh, S. Talwar, G. Wu, N. Himayat, and K. Johnsson, "Capacity and coverage enhancement in heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 32–38, 2011.

[32] P. Popovski, G. Mange, A. Roos et al., "Deliverable D6. 3 intermediate system evaluation results," ICT 317669, METIS, 2014.

[33] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. D. Johnson, "M2M: from mobile to embedded internet," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 36–43, 2011.

[34] A. Lo, Y. W. Law, and M. Jacobsson, "A cellular-centric service architecture for machine-to-machine (M2M) communications," *IEEE Wireless Communications*, vol. 20, no. 5, pp. 143–151, 2013.

[35] C. W. Johnson, *Long Term Evolution in Bullets*, Northampton, England, UK, 2012.

[36] P. Tian, H. Tian, L. Gao, J. Wang, X. She, and L. Chen, "Deployment analysis and optimization of Macro-Pico heterogeneous networks in LTE-A system," in *Proceedings of the 15th International Symposium on Wireless Personal Multimedia Communications (WPMC '12)*, pp. 246–250, September 2012.

Hindawi Publishing Corporation
Mobile Information Systems
Volume 2016, Article ID 4816325, 10 pages
http://dx.doi.org/10.1155/2016/4816325

Research Article

Maintaining Mobile Network Coverage Availability in Disturbance Scenarios

Joonas Säe and Jukka Lempiäinen

Department of Electronics and Communications Engineering, Tampere University of Technology, Tampere, Finland

Correspondence should be addressed to Joonas Säe; joonas.sae@tut.fi

Received 11 July 2016; Accepted 22 September 2016

Academic Editor: Ioannis Moscholios

Copyright © 2016 J. Säe and J. Lempiäinen. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Disturbance and disaster scenarios prevent the normal utilization of mobile networks. The aim of this study is to maintain the availability of cellular networks in disturbance scenarios. In order to extend the disaster time functionality, energy usage optimization is needed to maintain reasonable coverage and capacity. Simulations performed with different network layouts show the effects of choosing only a portion of evolved node B (eNB) macrosites to operate at a time. Different sets of three to nine three-sectored eNB sites are selected to study how the network would perform with a limited number of eNB sites. Simulation results show how the mobile network availability duration can be sustained by selecting a set of eNB sites to operate at a time and still maintain a reasonable service level and availability in disturbance scenarios. An increase of 100% to 500% can be achieved in the duration of "backup coverage" in cellular networks with backup batteries when the percentage of active eNB sites is reduced down to 20%.

1. Introduction

Disaster and disturbance scenarios usually occur without a warning. Whether they are natural weather-based storms or disasters caused by human, such as accidents or sabotage, the effects can be devastating and usually prevent the normal utilization of mobile networks. Typically, storms are the cause of blackouts in electrical grids [1]; furthermore, they have an impact on public safety and commercial mobile networks, thus yielding service and communication outages in urban and rural areas. This can eventually prevent citizens from requesting emergency help in these outage areas. In addition, maintenance and rescue teams can not communicate through commercial mobile networks and have to have a separate communication system.

Service outages in mobile networks are mainly caused by (storm-related) power outages. In order to enable some service in these cases, evolved node B (eNB) macrosites are typically supplied with backup batteries. These reserve energy resources provide power to run eNBs, but only for a limited time period. In Finland, this corresponds to 2–4 hours, as required by the Finnish Communications Regulatory Authority [2]. After strong weather phenomenon, in the worst case, the repair-work may take several days resulting in the unavailability of commercial cellular networks that may also endanger rescue operations. An alternative is to have aggregates over the network or at certain critical eNB sites, to guarantee their electricity supply for a longer period of time. Because aggregates are slightly costly to be supplied and used at every site, some optimization is needed in a similar way as in the case of battery backups; that is, how many aggregates should be enough to enable sufficient cellular network coverage? Moreover, in case of longer term of electrical cut-offs (i.e., over one day), aggregates are eventually required to guarantee (cost-efficient) mobile network communications in disturbance scenarios.

Another type of critical discontinuity in mobile networks may happen due to major malfunctions in the core transmission network, major damage in the core network elements, such as controllers, or in the switching functions. These malfunctions can also be caused by sabotage or a cyberattack, which may cause very wide discontinuity in the whole mobile

communications network. However, these are out of the scope of this study as this paper is concentrating more on optimizing the backup energy utilization.

Some solutions for mobile communications have been proposed to manage these disturbance and disaster scenarios to improve the resiliency of mobile networks. In Japan, the so called critical sites have been implemented in urban areas to durable high-rise buildings [3] with at least 24 h backup batteries to give continuous service, for example, in case of disasters. High altitude platforms (HAPs) [4] have been discussed earlier as a possible solution to provide service coverage in the case where hurricanes and tornadoes have destroyed the existing infrastructure of a cellular network in the disaster area. This idea has been even taken further with Google's project Loon, which targets to provide worldwide Internet access through HAPs, implemented with balloon platforms [5]. Also Facebook has their HAPs approach in their internet.org project [6]. Instead of utilizing balloons, Facebook relies on solar powered drones.

There are also solutions to utilize satellite communication services for disaster areas although these systems are rather expensive and might not work inside buildings. Moreover, temporary movable networks, like mobile ad hoc networks (MANETs), are a popular research area for disaster scenarios as seen from publications on the topic in [7–9]. Many of these solutions are replacing the existing infrastructure with a new one instead of trying to improve it.

In this paper, the aim is to utilize the traditional macrocellular network infrastructure and extend the availability of the network several times from the current backup time when the electrical grid is down and the mobile network relies on backup power. This is possible with an approach, where only a portion of the eNB sites are utilized at a time and switched to another set of sites when the first set runs out of energy. This differs from a traditional sleep mode technique where the nonactive eNBs are not utilized at all during the "sleep time" to save as much energy as possible, which is important in disturbance situations. This will obviously degrade the coverage availability and capacity, but disturbance and disaster scenarios require exceptional methods to enable even some service as long as possible. Only macrosites are considered in this study because small cells are usually available only within cities and they are not equipped with backup power. Maintaining reasonable coverage for disturbance scenarios in both urban and rural areas is emphasized; that is, the focus is on having more resilient mobile networks. Long term evolution (LTE) is utilized in the simulations with evolved UMTS terrestrial radio access (E-UTRA) operating bands 20 and 3 (800 MHz and 1800 MHz) for rural environment. E-UTRA bands 3 and 7 (1800 MHz and 2600 MHz) are utilized for urban environment. All of the considered frequency bands are frequency division duplex (FDD) channels.

2. Related Work

This work is related to at least two different energy-saving concepts: cellular networks with sleep modes and green cellular networks. However, the idea of these concepts is not utilized in the traditional way. This paper utilizes the idea of saving energy only in disturbance situations, that is, in disaster scenarios in the following manner: saving energy by utilizing only a portion of the eNB sites to provide "backup coverage" to areas where the electricity supply from the grid has been (temporarily) cut off, for example, due to strong weather phenomenon. Thus, this study goes beyond the state of the art for utilizing mobile network backup energy as this kind of approach has not been proposed for disturbance situations.

The idea of green cellular networks is to have sustainable development in cellular networks towards utilizing sustainable energy sources and taking into account environmental aspects. This usually relates to favouring renewable energy resources to reduce the carbon footprint of cellular network infrastructure or saving energy by reducing the energy consumption of base stations in some manner. Green cellular networks are nowadays studied widely and several publications are available in this field. The authors in [10], for example, present a survey on the state of green cellular networks and the possible challenges they currently have. The largest power consuming part of cellular networks is base stations, as presented in [11]. Furthermore, the largest part of base station power consumption comes from power amplifiers (50–80%) [12].

The traditional sleep mode concept has been studied, for example, in [13], where the effect of one or more base stations turned to "sleep" is studied. A case study of dynamically switching off base station sites is presented in [14], where the authors show great energy saving in the cellular network during low-traffic periods. One of the most recent wide surveys on these energy-efficient base stations, that is, base station sites with sleep modes as well as green cellular networks, is presented in [15]. The authors present an extensive list of recent publications on these topics and discuss the assumptions and simplifications utilized in these papers to show the great effect they can have in the achievable benefits in actual networks.

Most of the traditional sleep mode techniques consider only a "normal" network utilization. This is because the idea behind most sleep mode techniques is based on saving energy in the low-traffic time periods during the day and usually the sleep modes are only considering these circumstances. This does not necessarily mean that these techniques could not be utilized during disturbance scenarios. However, it should be highlighted that traffic load increases during disturbance scenarios as people try to contact other people from these areas (and vice versa) and utilize the mobile network to search for information regarding the cause of the disturbance situation. Thus, algorithms based on low-traffic network utilization would not work during these events.

A modified sleep mode concept in cellular networks is the main approach utilized in this paper for the functionality of eNB sites operating only with backup power (i.e., during a disturbance scenario). However, during this sleep mode, the eNB sites just wait for their turn to power on, not broadcasting anything during this time. This way, for example, power amplifiers and air conditioning are not needed and signal processing power requirements are very low as there is nothing to transmit or receive during the sleep time. Thus, the power consumption of these "sleeping eNB sites" is very marginal.

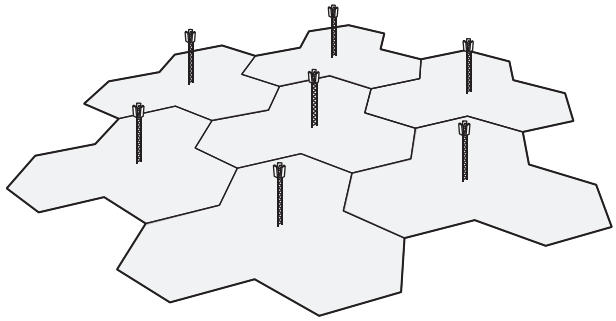


Figure 1: Cloverleaf tessellation with seven three-sectored eNB sites.

This is more effective in saving energy than having traditional sleep mode techniques, where the eNBs "wake up" to check the situation every now and then. The approach utilized in the paper does not take into account traffic loads, instead the target is to find out what would be the effect of selecting only a portion of the available eNB sites to the network performance. Moreover, the study does not include algorithms in how to actually select the usable eNB sites (although the selection of eNB sites is based on choosing some eNB sites evenly within a target area).

3. Radio Network Planning

Radio network planning is a process to design the best possible wireless communication network such that coverage, capacity, and the quality of service meet (and exceed) the requirements set by the amount of subscribers in a target area. In order to maximize these aspects, a radio network planner needs to pay careful attention to some very important factors. These include the environment, which defines the propagation slope, the model utilized to calculate the maximum path loss, and the network layout in terms of base station locations with respect to other base stations.

3.1. Network Layout. The performance of a cellular network depends on many factors and one of the important factors is the used network layout or more precisely the used tessellation. Ideally, tessellations form a continuous network that has raster-like properties. This means that the layout has symmetric shapes regarding the positions of the eNB sites and antenna directions. The half-power beamwidth (HPBW) of the used antennas also affects the choice of tessellation, but for a 65 degree HPBW antenna the cloverleaf tessellation is the most optimal choice [16]. It is the most commonly used tessellation in Europe and an example of it with seven eNB sites is presented in Figure 1.

Figure 1 shows that every eNB site antenna is pointed directly towards one of the closest neighboring eNB sites. However, this is done in such a way that the antenna beams are pointed in between two other antenna beams thus reducing interference as much as possible.

3.2. Propagation Model. Coverage of a mobile network is defined as a geographical area, where the eNB and user equipment (UE) can still communicate with each other. Coverage estimations can be calculated with reasonable accuracy, for example, by using a well-known Okumura-Hata prediction model equation (or its extended version, the COST-Hata model) [17]:

$$L = A + B \cdot \log_{10}(f_{\mathrm{MHz}}) - 13.82 \cdot \log_{10}(h_{\mathrm{eNB}}) - a(h_{\mathrm{UE}}) + (C - 6.55 \cdot \log_{10}(h_{\mathrm{eNB}})) \cdot \log_{10}(d_{\mathrm{km}}) + C_m, \quad (1)$$

where L is path loss in [dB], A, B, and C are constants, f_{MHz} is the used frequency in [MHz], h_{eNB} is eNB antenna height in [m], h_{UE} is UE antenna height in [m], $a(h_{\mathrm{UE}})$ is a city size dependent function, d_{km} is distance betweene NB and UE in [km], and C_m is an area correction factor.

Before the Okumura-Hata model can be used to get the maximum cell distance, some link budget calculations are needed to define the maximum path loss between the eNB and UE.

It should be noted that the Okumura-Hata model (1) has to be tuned for every environment separately to get more accurate results. In addition, the area correction factor, C_m, has to be set regarding the propagation environment morphological values. Correspondingly, the propagation slope, which is defined by $(C-6.55\cdot\log_{10}(h_{\mathrm{eNB}}))$, has to be set regarding the eNB antenna height and the wanted propagation path loss exponent by tuning the C constant.

3.3. Planning Thresholds. Planning thresholds are used together with propagation prediction models to calculate the maximum path loss in the link budget. The main planning threshold value is the slow fading margin [18]. It is calculated from the standard deviation of the slow fading, the propagation slope, and the coverage probability for the service.

It should also be addressed whether the radio network coverage is designed for outdoor or for indoor usage. When designing a radio network with outdoor macrocells, and there should also be some indoor coverage, some additional planning thresholds are needed. First, the value of slow fading margin should be increased and then the building penetration loss should be added to the link budget calculation in order to expect more realistic predictions.

One example of a typical average distance between neighboring eNB sites, that is, intersite distance (ISD), is 750 m in urban and 7500 m in rural environment [19]. While maintaining the antenna height and moving eNB sites closer to each other, more coverage overlapping occurs. This means that areas with coverage from one eNB site have coverage also from other neighbor eNB sites. In rural areas, the coverage overlapping is minimized because of large ISDs and, respectively, in urban areas there is more coverage overlapping because of shorter ISDs; that is, the eNB sites are close to each other. This means that the possibilities to decrease the number of eNB sites and still have available coverage should be greater in urban areas than in rural areas, but this depends heavily on the configuration of the network.

3.4. Practical Capacity Requirements. Mobile network capacity requirements depend on population density and the

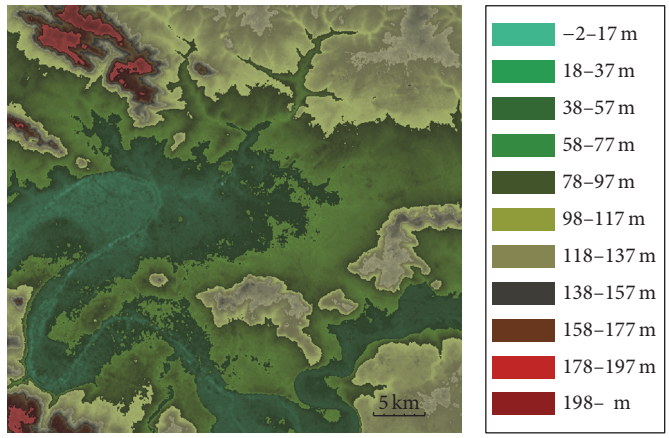


Figure 2: Digital elevation map of the simulation area.

requested services. Since mobile network cells are shared mediums, the available capacity per UE depends on the total number of mobile network subscribers within the serving cell and the data throughput each UE requests. To expand this limited capacity, more channels, that is, additional frequency bands, can be utilized together with different technologies. Thus, recent implementations include several frequency layers (bands) and technologies at one eNB site [20]. An example of this would be a base station site that has a GSM base station, a UMTS base station, and a LTE base station (eNB) and each of these technologies would operate on at least two different frequency bands. These could be, for example, GSM operating at 900 MHz and 1800 MHz, UMTS operating at 900 MHz and 2100 MHz, and LTE operating at 800 MHz and 2600 MHz. To maximize availability and thus to minimize the required electrical power, several technologies, for example, GSM and UMTS, and several frequency layers should be switched off (e.g., only the lowest frequency band of the remaining technology is utilized for service, such as LTE 800 MHz) and still certain services such as short messages, speech, or limited data services should be maintained (depending on the available technology and its capabilities). This results in conserving backup power in crisis situations but still enables services for subscribers for a longer period of time compared with a situation where eNB sites would continue to operate normally and quickly consume the available reserve energy.

4. Simulations

Link-level simulations were implemented using commercially available radio network planning software called ICS Designer. The area was a randomly picked and relatively flat area in France, which had a 25 m resolution digital elevation map as shown in Figure 2. After the simulation results were obtained, they were exported into Matlab software for further analysis and visualization.

4.1. Simulation Setup and Environment. All simulations were based on a macrocellular cloverleaf tessellation with eNB sites that had three sectors. A conventional eNB grid of 19 sites was used as a reference case, which has one eNB site in the middle and two tiers, first tier with six sites and second tier with twelve eNB sites, around it forming a hexagonal grid. All eNB sites in the original layout had an equal ISD of 500 m for urban environment and 4000 m for rural environment. Circular areas with 1.25 Km and 10 Km radii from the centre eNB site were the base "area of interest" for the analysis. In total 57 cells were the maximum amount of cells in one area. In order to take into account sites outside the target area, three more tiers of eNB sites were placed around the central grid of eNB sites. Thus, continuous coverage was formed such that the selection of eNB sites was repeated around the target area to avoid errors in the edge areas of the calculation area.

The antenna height values of 20 m and 40 m with five-degree antenna tilting were used in the simulations. The horizontal antenna HPBW was 65 degrees with 17.22 dBi gain for eNB sites while UE antennas were omnidirectional antennas with 0 dBi gain.

Okumura-Hata model was used for calculating the path loss values in the simulations. Fading was taken into account with log-normal distribution having zero mean and 9 dB standard deviation. All key parameters regarding the simulations are presented in Table 1. The distribution of transmit power to different LTE downlink physical channels is visible in Table 2.

4.2. Simulation Scenarios. Simulation scenarios had a large number of different types of network layouts. The idea was to find suitable configurations to achieve good enough coverage for the simulation area and at the same time minimize the service outage. The cellular network can function for an extended period of time when only a part of the network is working on backup energy at a time. In order to reduce the number of different possible combinations which would be possible for a set of eNB sites, the number of different kinds of sets out of all possible sets was reduced. If all possible combinations would be considered, the maximum number of different sets (in any order), for example, with nine eNB sites out of 19 eNB sites would be $\binom{19}{9} = 92\,378$, and even three eNB sites out of 19 eNB sites would have $\binom{19}{3} = 969$ different sets. Thus, the number of different sets was set at a small amount,

Table 1: Key simulation parameters.

Parameter	Value	Unit
Operating frequency band 20 (FDD)	800	MHz
Operating frequency band 3 (FDD)	1800	MHz
Operating frequency band 7 (FDD)	2600	MHz
Bandwidth	10	MHz
Number of usable resource blocks	50	pcs
Loading	75	%
Calculation resolution	25	m
Intersite distance, urban	500	m
Intersite distance, rural	4000	m
Antenna height, urban	20	m
Antenna height, rural	40	m
Building height	8	m
eNB antenna HPBW	65	∘
eNB antenna gain	17.22	dBi
Additional losses	3	dB
UE antenna height	1.5	m
UE antenna gain	0	dBi
Max. eNB TX power	20	W

Table 2: Distribution of downlink transmit power to different LTE physical channels.

LTE physical channel	Share	Unit
Reference signal	4.76	%
Physical downlink shared channel (PDSCH)	74.33	%
Physical downlink control channel (PDCCH)	20.24	%
Physical broadcast channel (PBCH)	0.33	%
Primary synchronization signal (P-SS)	0.17	%
Secondary synchronization signal (S-SS)	0.17	%
Total	100	%

six in this paper, to study how the performance of the network behaves as a function of available eNB sites. The focus is especially to study having less than half of the eNB sites per studied area. The eNB sites chosen for a set were based on spreading the available sites as evenly as possible to the studied area, since, for example, choosing only the three closest eNB sites from one "corner" of the grid would results in the uneven distribution of backup coverage in the target area of interest.

Figure 3 shows how eNB sites were chosen for three to nine eNB site cases. As can be seen, the eNB sites were chosen in a way to be able to serve the target area as evenly as possibly. An example of service probability with two different sets of six eNB sites is visualized in Figure 4.

5. Results

Figure 5 shows the mean reference signal received power (RSRP) level, Figure 6 the mean service probability, and Figure 7 the average signal-to-interference-and-noise ratio (SINR) with respect to the number of eNB sites. Figures 8 and 9 present the average area throughput to study the available capacity with respect to the amount of eNB sites. The corresponding exact values are presented in Tables 3 and 4.

The mean RSRP values in Figure 5 show that as the number of eNB sites grow the average RSRP increases, which is expected. Rural environment with E-UTRA band 3 (1800 MHz) has the lowest average RSRP values and correspondingly rural environment with E-UTRA band 20 (800 MHz) has the highest RSRP values. The average values of RSRP are clearly different for all cases. First of all, the environment type with different antenna heights for the rural and urban areas has a huge impact on the average RSRP values, since the antenna height for the rural area is two times the height of the urban area. Next, another important factor is the utilized frequency; that is, for rural area, the higher frequency band is more than twice the utilized lower frequency band. In the urban cases, the higher frequency band is less than twice the utilized lower frequency band; thus, the differences between the average RSRP values for the two rural cases is more dramatic than the corresponding differences for the two urban cases. Finally, the urban case is also affected by the buildings, thus lowering the average RSRP values. In this study, the lowest value was −82.68 dBm with 15.79% of eNB sites available.

Figure 6 presents the average probability of having backup coverage in the different cases. The lowest probability of having service coverage was with E-UTRA band 7 (2600 MHz) in the urban environment. The lowest value was 55.44% with 15.79% of the eNB sites available. Thus, although the urban area had denser eNB site placement than the rural area, the coverage areas with a higher frequency band are quite limited as well as overlapping areas. This is mostly because the attenuation is higher in the urban environment (with buildings) as well as in a higher frequency band. When the same frequency band was utilized in both the urban and rural environment, the urban E-UTRA band 3 had slightly higher probability of having a service than its rural counterpart (rural E-UTRA band 3). The highest service availability was achieved with the lowest frequency, E-UTRA band 20 in the rural case, where the probability of having a service was over 80% even with the lowest number of eNB sites (15.79%). In order to have over 80% probability of (backup) coverage in other cases, E-UTRA band 3 cases (1800 MHz) would need around 26.32% in urban and 31.58% in rural scenarios and E-UTRA band 7 (2600 MHz) around 50% in the urban environment.

When the exact values are observed from Table 3 as a function of available eNB sites, the differences between different E-UTRA bands and environments are easy to compare as well as the differences between each parameter. Intuitively thinking, each parameter (RSRP, service probability, and SINR) should have a better value as the percentage of eNB sites per area is increased and in general this is observed in Table 3. The only difference is the utilized lower urban frequency band (1800 MHz), where the average SINR value is slightly increased when more than 31.58% of eNB sites are utilized. However, this increase stops after more than 42.11% of eNB sites are utilized.

In Figure 7, the mean SINR values are presented. It can be noticed that, in rural environment, the achievable values are higher than in urban cases, mostly because of low interference

Table 3: Simulation results.

Percentage of eNB sites per area	Mean RSRP (dBm)				Mean service probability (%)				Mean SINR (dB)			
	Rural 800 MHz	Rural 1800 MHz	Urban 1800 MHz	Urban 2600 MHz	Rural 800 MHz	Rural 1800 MHz	Urban 1800 MHz	Urban 2600 MHz	Rural 800 MHz	Rural 1800 MHz	Urban 1800 MHz	Urban 2600 MHz
15.79%	−76.37	−82.68	−78.83	−80.83	80.35	68.52	72.50	55.44	5.22	5.62	1.53	3.51
21.05%	−75.35	−81.75	−77.91	−80.38	83.35	72.44	75.81	59.81	4.53	3.89	1.13	2.55
26.32%	−74.29	−80.88	−76.77	−79.79	86.28	76.41	80.50	66.32	4.11	3.53	1.17	1.99
31.58%	−73.56	−80.29	−76.07	−79.25	89.37	80.66	82.62	69.23	3.98	3.30	1.04	1.64
36.84%	−72.91	−79.67	−75.33	−78.79	90.66	82.58	85.06	73.05	3.96	3.27	1.27	1.60
42.11%	−72.39	−79.20	−75.03	−78.51	91.61	84.24	86.36	75.29	3.69	3.02	1.32	1.57
47.37%	−71.91	−78.72	−74.29	−78.23	92.38	85.47	87.33	77.49	3.45	2.79	1.32	1.57

Table 4: Mean area throughput.

Percentage of eNB sites per area	Mean area throughput (Mbps/km^2)			
	Rural 800 MHz	Rural 1800 MHz	Urban 1800 MHz	Urban 2600 MHz
15.79%	0.23	0.24	9.07	11.79
21.05%	0.28	0.26	11.44	13.88
26.32%	0.33	0.31	14.38	16.11
31.58%	0.39	0.36	16.94	18.43
36.84%	0.46	0.42	20.40	21.38
42.11%	0.50	0.46	23.49	24.32
47.37%	0.55	0.50	26.43	27.37

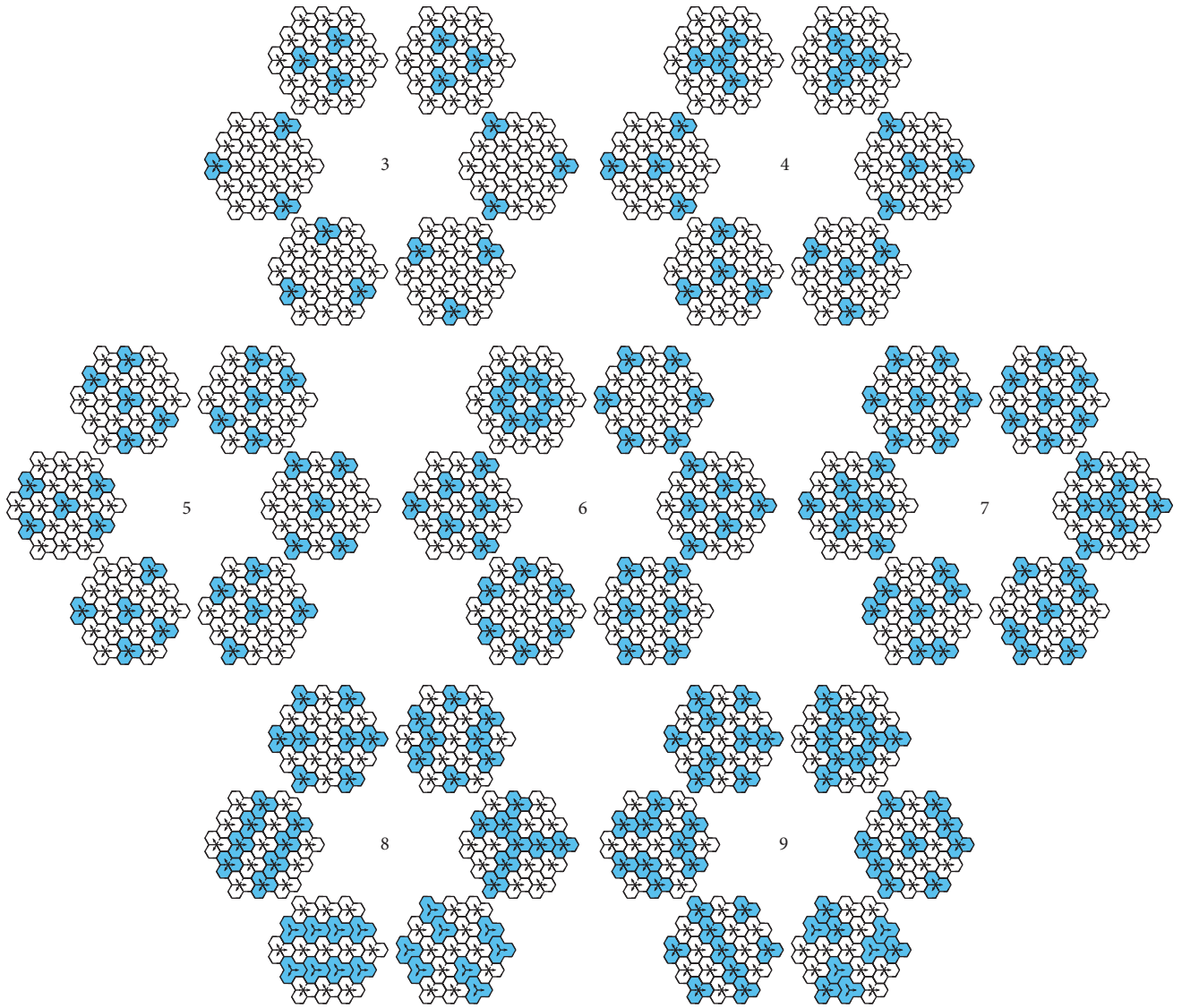


Figure 3: Selection of different three to nine eNB site configurations that were utilized in the simulations for a target area (and repeated in a similar way around this area to form a continuous coverage). The selected eNB sites are highlighted with blue color.

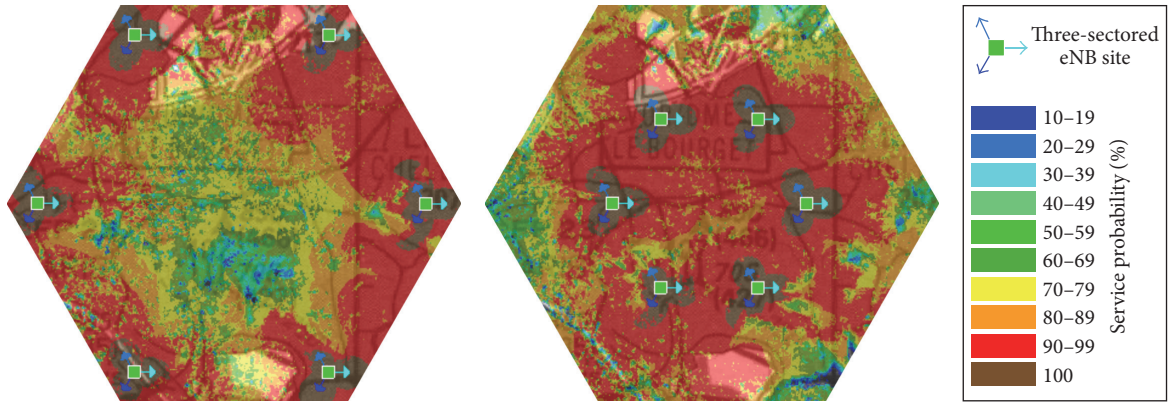


Figure 4: An example of service availability with two different sets of six eNB sites.

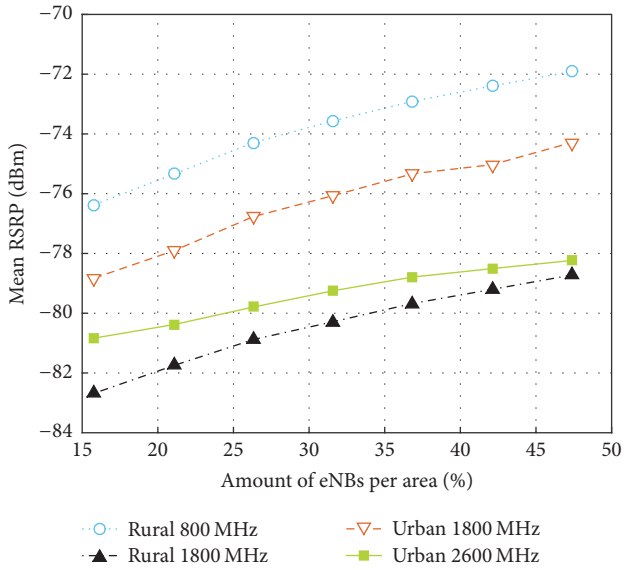


Figure 5: Average RSRP level with respect to the amount of eNB sites per area.

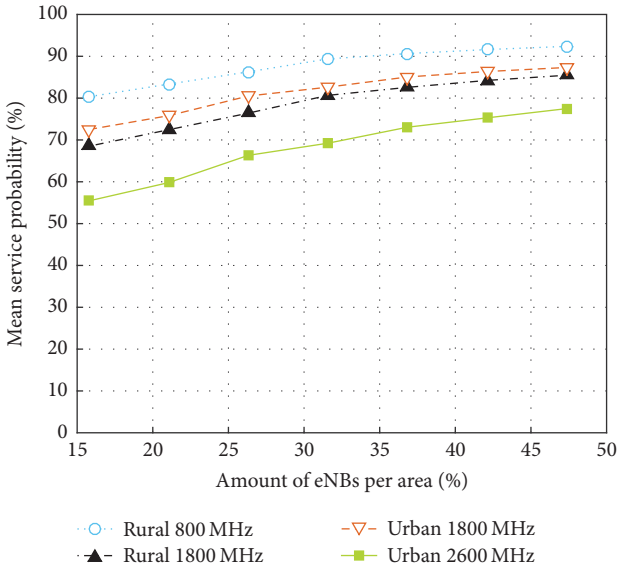


Figure 6: Average service probability with respect to the amount of eNB sites per area.

from less dense eNB site placements; that is, the minimum ISD is 4000 m. When more eNB sites are available, the average SINR degrades because of the increased interference levels.

Figures 8 and 9 show the average throughput per area. As can be noted, the available average data rates are much higher in urban cases since the ISD is much lower; that is, the network is much denser. It should be remembered that the loading of the network was set to 75%; thus only 37 resource blocks were available out of 50 resource blocks available in the 10 MHz bandwidth for LTE.

6. Conclusions and Discussion

In this paper, the availability of mobile networks in disturbance scenarios was studied by simulating different network layouts for macro-eNB sites. Thus, a solution for improving the cellular network functionality in disturbance situations was presented, which is a very important topic in the resilience of mobile networks. The target of the simulations was to find out how a limited network configuration would perform in terms of coverage and capacity with mobile network availability (without adding additional backup power sources). This minimum configuration can be utilized with battery backups to extend the network availability called backup coverage. As the name suggests, this exceptional state of the network cannot have "normal" service availability as the goal is to provide "acceptable" or at least some coverage that would last much longer than initially. This would mean that some areas would not have coverage at every time instant during the backup coverage time window, but since the operating sets of eNB sites are cycled over time, all areas would have a certain time window when the service is still available. Thus, only 20% to 42% of eNB sites are needed in order to achieve approximately 75% availability in urban areas. Correspondingly, 15% to 25% of eNB sites are needed in rural areas for the same availability. This means that

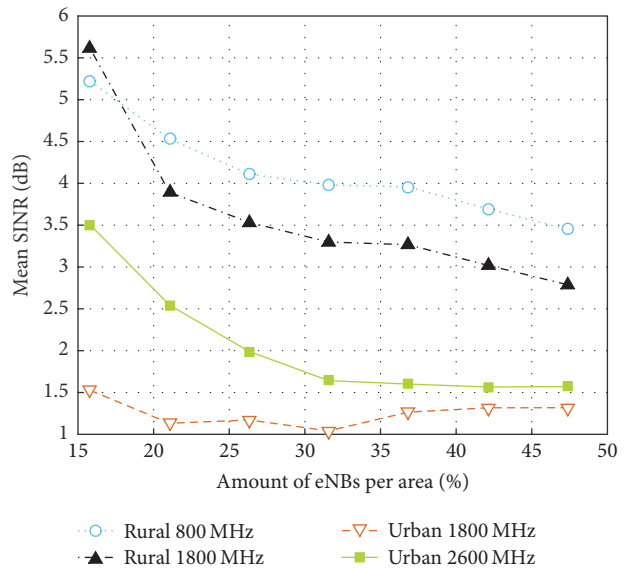


Figure 7: Average SINR with respect to the amount of eNB sites per area.

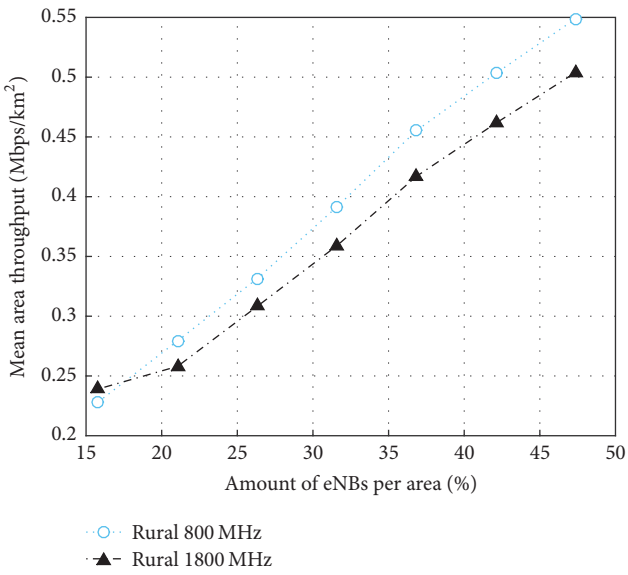


Figure 8: Average area throughput with respect to the amount of eNB sites per area in rural environment.

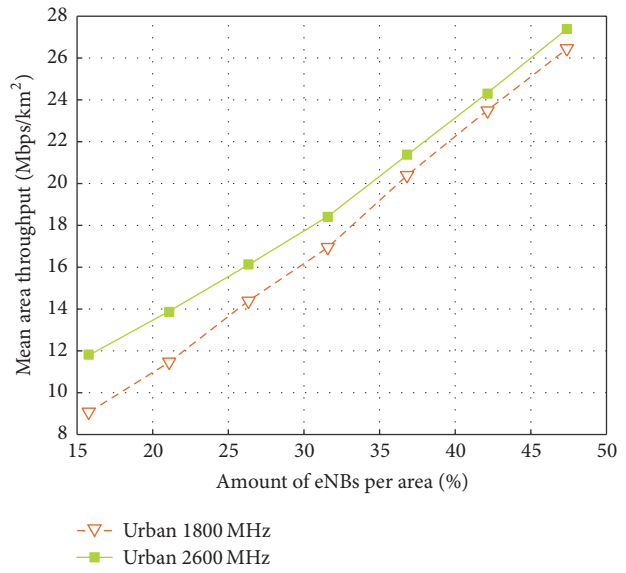


Figure 9: Average area throughput with respect to the amount of eNB sites per area in urban environment.

an increase of 100% to 500% could be achieved on the operational time of backup coverage, correspondingly. This also translates to having mediocre coverage for a longer period of time if approximately 20% of eNB sites would be equipped with aggregates or any other means of powering up eNB sites, which would be cheaper than having, for example, 24 h backup batteries deployed to every possible eNB site.

The comparison of the achieved results with other sleep mode techniques is not that straightforward. First of all, the traditional sleep mode techniques are designed to save energy during low-traffic time periods and this is not the case during disturbance situations. Second, many studies such as [13–15] consider different sleep mode techniques with different kinds of eNB site layouts with heterogeneous networks (besides macro eNB sites, there can also be micro eNB sites and small cell eNB sites). In order to get some idea about the effectiveness of the approach utilized within this paper, the authors in [21] present a comparable sleep mode technique, where part of the eNB sites remains active all the time and part of the eNB sites can switch to sleep mode. The results in the study suggest that on average the power savings from an actual network implementation would be from 7.38% to 27.72% on a 24 h time period. Thus, although still not directly comparable, the energy savings are clearly a lot less than in the approach utilized in this study. The difference is mostly explained with the fact that the sleep mode technique is designed to keep high quality of service (QoS), that is, high service availability. In this paper the requirements for QoS are clearly lower since the idea is lower QoS to enable even some sufficient service level for a clearly longer period of time by aggressively saving energy.

Although it seems that there would be more overlapping in the rural areas, it should be noticed that the service availability is higher in urban areas when the same E-UTRA frequency bands are utilized (band 3). This indicates that the operational time of the backup coverage in urban areas would last somewhat longer. Thus, the achievable gain is strongly related to the overlapping rate of neighboring cells which is also defined by the utilized frequency band as well as the configuration of the eNB sites.

The studied availability can be reached in existing networks by utilizing existing configurations and resources by adding only controlling units. Moreover, the achieved results assume that transmission lines and core networks are properly functioning (which is usually the case, at least in Finland, as the backhaul element backup power regulation is stricter than for eNB sites by a factor of 2–6 [2]), and thus more specific solutions need to be proposed to maintain their

availabilities in case of disconnections. It should be noted that, in real operational networks, the selection of eNB sites is not so straightforward. The geographical distribution of eNB sites is not symmetrical and only roughly follows the cloverleaf tessellation. Moreover, this study only showed how the network would perform if each eNB site was independent of another eNB site; that is, the study did not take into account the possible backhaul connection limitations of linked eNB sites. This would have more serious effects within rural environments where wireless microwave links are utilized more frequently compared with urban areas, where wireless links between eNB sites rarely exist. The results are valid only for the given configuration; however, they were chosen so they represent possible real-life implementations and provide insight on the level of service availability with a limited cellular network configuration.

The future work on this topic will consider more practical eNB site layouts; that is, the distribution of sites will be closer to real-life implementations with more clustered locations of eNB sites. Moreover, the effect of wireless microwave links, that is, the backhaul connection, will also be taken into account based on an operational cellular network operator.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

Author Joonas Säe would like to thank Tuula and Yrjö Neuvo Fund and Finnish Foundation for Technology Promotion for supporting the research work. The authors would also like to thank European Communications Engineering (ECE) Ltd. and the Finnish Funding Agency for Innovation (TEKES) for funding the research work.

References

[1] J. Strandén, H. Krohns, P. Verho, and J. Sarsama, "Major disturbances—development of preparedness in finland during the last decade," in *Proceedings of the 21st International Conference on Electricity Distribution*, p. 4, Frankfurt, Germany, June 2011.

[2] FICORA 54 B/2014 M, "Määräys viestintäverkkojen ja—palvelujen varmistamisesta sekä viestintäverkkojen synkronoinnista," December 2014 (Finnish), https://www.viestintavirasto.fi/attachments/maaraykset/Viestintavirasto54B2014M.pdf.

[3] NTT DoCoMo, "Deployment status of the new disaster preparedness measures," 2012, https://www.nttdocomo.co.jp/english/info/media_center/pr/2012/pdf/20120223_attachment02.pdf.

[4] P. Lähdekorpi, T. Isotalo, K. Kylä-Liuhala, and J. Lempiäinen, "Replacing terrestrial UMTS coverage by HAP in disaster scenarios," in *Proceedings of the European Wireless Conference (EW '10)*, pp. 14–19, Lucca, Italy, April 2010.

[5] Google, "Loon for all," http://www.google.com/loon/.

[6] Internet.org by Facebook, https://info.internet.org/en/story/connectivity-lab/.

[7] M. Conti and S. Giordano, "Mobile ad hoc networking: milestones, challenges, and new research directions," *IEEE Communications Magazine*, vol. 52, no. 1, pp. 85–96, 2014.

[8] L. Sassatelli, A. Ali, M. Panda, T. Chahed, and E. Altman, "Reliable transport in delay-tolerant networks with opportunistic routing," *IEEE Transactions on Wireless Communications*, vol. 13, no. 10, pp. 5546–5557, 2014.

[9] A. Polydoros, N. Dimitriou, G. Baldini, I. N. Fovino, M. Taddeo, and A. M. Cipriano, "Public protection and disaster relief communication system integrity: a radio-flexibility and identity-based cryptography approach," *IEEE Vehicular Technology Magazine*, vol. 9, no. 4, pp. 51–60, 2014.

[10] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green cellular networks: a survey, some research issues and challenges," *IEEE Communications Surveys and Tutorials*, vol. 13, no. 4, pp. 524–540, 2011.

[11] C. Han, T. Harrold, S. Armour et al., "Green radio: radio techniques to enable energy-efficient wireless networks," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 46–54, 2011.

[12] L. M. Correia, D. Zeller, O. Blume et al., "Challenges and enabling technologies for energy aware mobile radio networks," *IEEE Communications Magazine*, vol. 48, no. 11, pp. 66–72, 2010.

[13] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "On the effectiveness of single and multiple base station sleep modes in cellular networks," *Computer Networks*, vol. 57, no. 17, pp. 3276–3290, 2013.

[14] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 56–61, 2011.

[15] J. Wu, Y. Zhang, M. Zukerman, and E. K.-N. Yung, "Energy-efficient base-stations sleep-mode techniques in green cellular networks: a survey," *IEEE Communications Surveys and Tutorials*, vol. 17, no. 2, pp. 803–826, 2015.

[16] J. Itkonen, B. Tuzson, and J. Lempiainen, "Assessment of network layouts for CDMA radio access," *EURASIP Journal on Wireless Communications and Networking*, vol. 2008, Article ID 259310, 11 pages, 2008.

[17] D. Cichon and T. Kümer, "Propagation prediction models," in *Digital Mobile Radio Towards Future Generation Systems—Final Report*, COST Action 231, Ed., chapter 4, pp. 115–208, European Commission, Tech. Rep, 1999.

[18] J. Lempiäinen and M. Manninen, *UMTS Radio Network Planning, Optimization and Qos Management: For Practical Engineering Tasks*, Kluwer Academic, Norwell, Mass, USA, 2004.

[19] 3GPP, "Technical specification group radio access network; UMTS 900 MHz work item technical report (Release 8)," 3rd Generation Partnership Project (3GPP), Tech. Rep. TR 25.816, 2009.

[20] Nokia, "Flexi multiradio 10 base station," http://networks.nokia.com/sites/default/files/document/nokia_flexi_multiradio_10_base_station_brochure_0.pdf.

[21] A. Ali and S. E. Elavoubi, "Design and performance evaluation of site sleep mode in LTE mobile networks," in *Proceedings of the 26th International Teletraffic Congress (ITC '14)*, pp. 1–6, September 2014.

Hindawi Publishing Corporation
Mobile Information Systems
Volume 2016, Article ID 7048482, 6 pages
http://dx.doi.org/10.1155/2016/7048482

Research Article

LDPC Decoding on GPU for Mobile Device

Yiqin Lu, Weiyue Su, and Jiancheng Qin

School of Electronic and Information Engineering, South China University of Technology, Tianhe District, Guangzhou, China

Correspondence should be addressed to Weiyue Su; weiyue.su@gmail.com

Received 17 May 2016; Revised 17 August 2016; Accepted 25 August 2016

Academic Editor: Mariusz Głąbowski

Copyright © 2016 Yiqin Lu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A flexible software LDPC decoder that exploits data parallelism for simultaneous multicode words decoding on the mobile device is proposed in this paper, supported by multithreading on OpenCL based graphics processing units. By dividing the check matrix into several parts to make full use of both the local memory and private memory on GPU and properly modify the code capacity each time, our implementation on a mobile phone shows throughputs above 100 Mbps and delay is less than 1.6 millisecond in decoding, which make high-speed communication like video calling possible. To realize efficient software LDPC decoding on the mobile device, the LDPC decoding feature on communication baseband chip should be replaced to save the cost and make it easier to upgrade decoder to be compatible with a variety of channel access schemes.

1. Introduction

Low Density Parity Check (LDPC) error correcting code is a kind of linear block codes, proposed by Gallager in 1962 [1] and rediscovered by Mackay and Neal in 1996 [2]. It takes its name from its sparse check matrix. LDPC codes are capacity-approaching codes, which means that it allows the noise threshold to be set very close to the Shannon limit for a symmetric memoryless channel; thus, the practical constructions of LDPC code exists.

Good performance of the LDPC code is at the cost of a very large amount of calculation. DCP decoding computation has very high parallel computation. The current commercial LDPC decoder is based on the hardware implementation, which only allows several kinds of specific LDPC codes at the same time and is difficult to upgrade. There are a large number of studies using FPGA to realize the efficient LDCP decoder [3, 4]. With the rapid development of the graphics processing units (GPU) on the desktop, there are a lot of researches using CUDA framework for LDPC decoding [5, 6]. The LDPC code is widely used in the fourth generation of mobile telecommunications technology, which makes it significant to develop efficient software LDPC decoding on the mobile device. At the same time, software LDPC code can dynamically change the parameters, including code length, code rate, and the number of iterations to quickly deal with all kinds of network environment.

Open Computing Language (OpenCL) [7] is a framework for writing programs that execute across heterogeneous platforms consisting of CPU, GPU, DSP, FPGA, and other processors or hardware accelerators. This technical specification was reviewed by the Khronos members and approved for public release on 2008. Compute Unified Device Architecture (CUDA) [8] also enables developers to develop parallel computing program on GPU at the desktop. OpenCL appears later, but it supports more scenarios. With the rapid development of mobile devices, many mobile devices especially mobile phone began to have their own high-performance GPU chips. Some vendors such as Qualcomm, Imagination PowerVR, ARM, and Vivante are beginning to support the OpenCL on their mobile GPU [9], which make developing parallel computing program on mobile devices based on GPU easier. In this article, we tried to develop a LDPC decoder on the mobile GPU based on the OpenCL. Nevertheless, the global memory is limited on a mobile GPU; therefore, the performance is not as good as on the desktop GPU. We improve the decoding through making full use of the local memory of each computing unit and the private memory of each processing unit. At the same time, we properly reduce the number of threads per code word and add code-words

in decoding process, and better performance is obtained. In our experiments, as the best result in the decoder, the throughput reached 160 Mbps, which can satisfy the current mobile wireless communication in many cases, and delay time is less than 2 milliseconds (ms), which can satisfy many real-time applications like video calling.

2. MSA for LDPC Decoding

Belief propagation (BP) algorithm is a kind of important message passing algorithm, often used in the field of artificial intelligence [9]. Algorithm between each node transfers the belief information. For example, the belief information from bit node BN_n to check node CN_m depends on the observation of BN_n and all the check nodes BN_n connected with, except CN_m. Similarly, the belief information from check node CN_m to bit node BN_n depends on the observation of CN_m and all the bit nodes CN_m connected with, except BN_n. As a BP algorithm, the Min Sum Algorithm (MSA) is a very efficient LDPC decoding algorithm [10]. It is based on the belief propagation between nodes connected as indicated by the Tanner graph [11] edges. Figure 1 shows the Tanner graph of a particular 4×8 $\mathbf{H}$ matrix. MSA, proposed by Gallager, operates in the logarithmic probabilistic domain.

LDPC code is a special form of linear (N, K) block code, defined by sparse binary parity check $\mathbf{H}$ matrices of dimension $M \times N$, while $M = N - K$. We assume that the channel is an additive white Gaussian noise (AWGN) channel with the mean 0 and the variance σ^2. BPSK modulation maps a code-word $\mathbf{c} = (c_1, c_2, \ldots, c_N)$ onto the sequence $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, according to $x_i = (-1)^{c_i}$. The received sequence is $\mathbf{y} = (y_1, y_2, \ldots, y_N)$, with $y_i = x_i + n_i$. In the case of receiving y_n, the logarithmic a priori probability of x_n is Lp_n^0. MSA is as shown in Figure 2.

Before entering the loop iteration, we use the received sequence $\mathbf{y}$ to initialize the prior probabilities of BN_n as follows:

$$Lp_n^0 = \ln \frac{p_n^0(0)}{p_n^0(1)} = 2\frac{y_n}{\sigma^2}, \qquad Lq_{nm}^0 = Lp_n^0. \tag{1}$$

In this algorithm, we do not compute the posterior probabilities of BN_n and CN_m directly; instead, we compute the message transferring between the bit nodes and check nodes as well as the posterior probabilities before hard decoding.

In the step of updating message CN_m to BN_n, for ith iteration, accessing $\mathbf{H}$ in row-major order, Lr_{mn}^i as the message sent from CN_m to BN_n is updated according to any bit nodes connected to CN_m in Tanner graph, except the BN_n. The update process, called minimum step, is as follows:

$$Lr_{mn}^i = \prod_{n' \in N(m) \backslash n} \operatorname{sign}\left(Lq_{n'm}^{i-1}\right) \min_{n' \in N(m) \backslash n} \left|Lq_{n'm}^{i-1}\right|. \tag{2}$$

Using the $\mathbf{H}$ matrix and Tanner graph in Figure 1, for instance, $Lr_{0,0}^i$ is updated by BN_1 and BN_2, as in Figure 3, $r_{0,0}^i = f(Lq_{1,0}^{i-1}, Lq_{2,0}^{i-1})$.

The posterior probabilities of BN_n is updated by the prior probabilities of BN_n and all the check nodes connected to BN_n:

$$Lp_n^i = Lp_n^0 + \sum_{m \in M(n)} Lr_{mn}^i. \tag{3}$$

Similarly, in the step of updating message BN_n to CN_m, for ith iteration, Lq_{nm}^i, as the message sent from BN_n to CN_m is updated according to any check nodes connected to BN_n in Tanner graph, except the CN_m. The update process is called sum step.

$$Lq_{nm}^i = Lp_n^i - Lr_{mn}^i. \tag{4}$$

Using the Tanner graph in Figure 1, for instance, $Lq_{0,0}^i$ is updated by CN_2, as in Figure 4, $Lq_{0,0}^i = f(Lr_{2,0}^i)$.

Actually, the steps of updating Lp_n^i and Lq_{nm}^i can be exchanged. If we update Lp_n^i first, the result of Lp_n^i can be used to update Lq_{nm}^i, which reduces the repeated computation.

The final hard decoding is performed at the end of an iteration.

$$c_n^i = \begin{cases} 1 & \text{if } Lp_n^i < 0 \\ 0 & \text{if } Lp_n^i > 0. \end{cases} \tag{5}$$

The iteration procedure is stopped if the decoded word $\mathbf{c}$ verifies all parity check equations $\mathbf{cH}^T = 0$, or the maximum iteration is reached.

The implementation of decoder is achieved by a flood scheduling algorithm [12]. It guarantees that the bit nodes would not interfere with each other in the update step and when updating check nodes, check nodes will not interfere with each other too. Using this principle allows the true parallel execution of MSA for LDPC decoding based on the stream-based computing method.

3. OpenCL for Mobile GPU

Modern GPU is based on ultra high parallel computing ability and programmable pipeline. Stream processor of GPU is able to do general-purpose computation [13]. GPU is more efficient than CPU floating point performance especially when we deal with the single instruction multiple data (SIMD) and the completion of compute-intensive tasks, in which data processing operation needs far more time than the data scheduling and data transmission [14].

Unlike the dedicated GPU for desktop computers, a mobile GPU is typically integrated into an application processor, which also includes a multicore CPU, an image processing engine, DSPs, and other accelerators [15]. Recently, modern mobile GPUs such as the Qualcomm Adreno GPU [16], the Imagination PowerVR GPU, ARM Mali, and GPGPU on Vivante tend to integrate more compute units in a chip. Mobile GPUs have gained general-purpose parallel computing capability thanks to the multicore architecture and emerging frameworks such as OpenCL, and they are likely to offer flexibility similar to vendor specific solutions designed for desktop computers, such as CUDA of Nvidia.

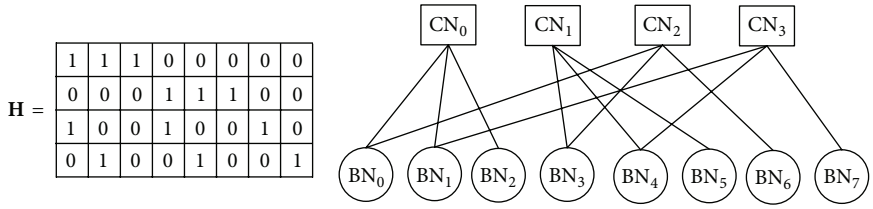


Figure 1: A 4×8 **H** matrix and its Tanner graph representation.

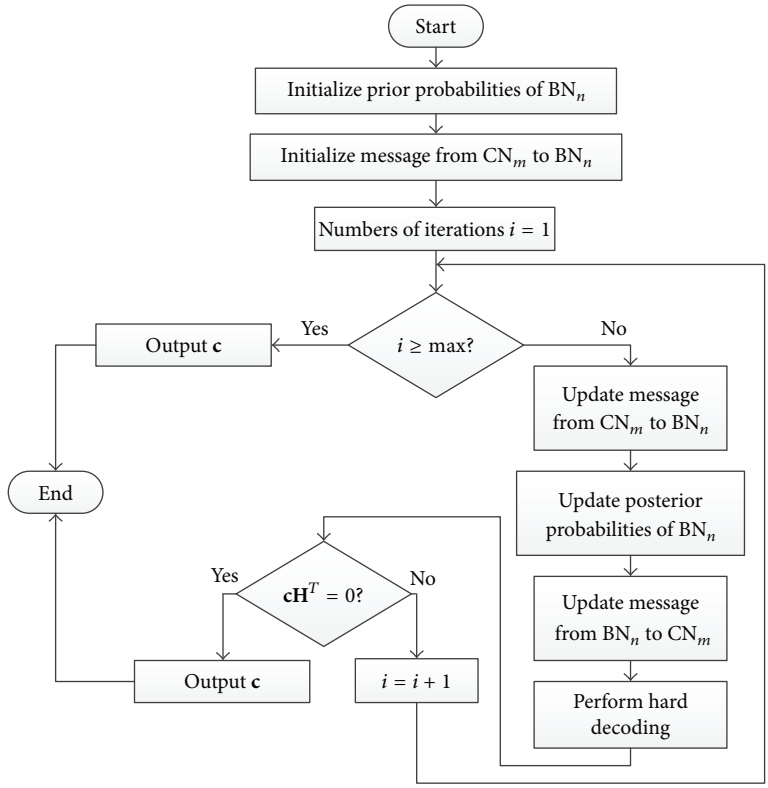


Figure 2: Process of Min Sum Algorithm.

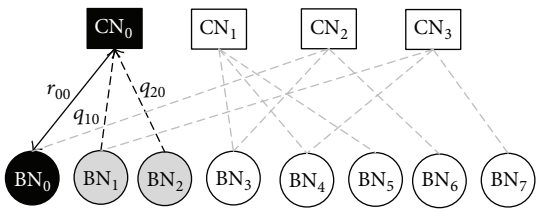


Figure 3: Example for updating $Lr^i_{0,0}$, the message from CN_0 to BN_0.

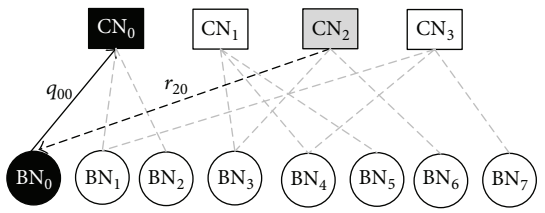


Figure 4: Example for updating $Lq^i_{0,0}$, the message from BN_0 to CN_0.

OpenCL is a programming framework designed for heterogeneous computing across various platforms [17]. In OpenCL, a host processor (typically a CPU) manages the OpenCL context and is able to offload parallel tasks to several compute devices (for instance, GPU).

The parallel jobs can be divided into work-groups, and each of them consists of many work-items which are the basic processing units to execute a kernel in parallel.

OpenCL defines a hierarchical memory model containing a large global memory but with long latency and a small but fast local memory which can be shared by work-items in the same work-group; what is more, each work-item has its own memory, which is not shared with other items and is fastest accessing.

To efficiently and fully utilize the limited computation resources on a mobile processor for better performance, we partition the tasks between CPU and GPU and explore the algorithmic parallelism, and memory access optimization needs to be carefully considered.

On embedded platform, to handle various tasks is becoming a trend. OpenCL specification describes a subset of the OpenCL specification for handheld and embedded platforms.

The OpenCL embedded profile has some restrictions; for instance, there are optional support for 3D images and no support for 64-bit integers and no support for 64-bit integers. The details of the OpenCL embedded profile can be found in Khronos's website [17].

Despite these specification restrictions, it is possible to use OpenCL to accelerate the program on the mobile devices. The compute-intensive computation on the mobile device is transferred to the GPU or other devices supporting OpenCL; not only these tasks can perform even more efficiently, but also CPU can handle more tasks that it is good at. Actually, LDPC decoding is a kind of traditional compute-intensive computation.

```
(1) as the work-item k in a work-group: do
(2)     m = k
(3)     for all BN_n (columns in H_mn): do
(4)         if H_mn == 1 then
(5)             H_BN[idx++] = n
```

Algorithm 1: Generating compact H_{BN} from matrix **H**.

```
(1) as the work-item k in a work-group: do
(2)     for offset = from 0 to N/M − 1: do
(3)     n = k × N/M + offset
(4)     for all BN_n (columns in H_mn): do
(5)         if H_mn == 1 then
(6)             H_CN[idx++] = n
```

Algorithm 2: Generating compact H_{CN} from matrix **H**.

4. Parallel MSA LDPC Decoding on Mobile GPU

MSA is an intensive processing, which should be processed in a high-performance specific computing engine, or in a highly parallel programmable device. On the mobile device, the GPU is a good choice. This general model, supported by GPU using OpenCL, executes kernels in parallel on several multiprocessors. Each processor is composed by several cores that dispatch multiple threads. In this section, a parallel processing to save the information of matrix **H** into work-items is showed. In order to save the private memory, each work-item only keeps the compressed information that related to its own computation. After that, the specific parallel algorithm in OpenCL kernel is introduced. Given an (N, K) LDPC code, it is important to manage the computation to reduce the expenditure in parallel programming. Instead of using $M \times w_c$ work-items (w_c is the maximum column weight of matrix **H**), the model uses $M \times 1$ work-items in each work-group, and each work-item updates the message about one check node, which means M work-items work for M check nodes, respectively.

4.1. Compact Representation of the Tanner Graph. The Tanner graph of a LPDC code is defined as **H**. We propose it in two separate data structures, namely, H_{BN} and H_{CN}. This is because one iteration of the LDPC decoder can be decomposed into horizontal and vertical processing, which means we update message from CN_n to BN_m and message from BN_m to CN_n, respectively.

The data structure used in the horizontal step is defined as H_{BN}. It is generated by scanning the matrix **H** in a row-major order and mapping only the bit nodes' edges associated with nonnull elements in **H** used by a single check node equation in the same row. Algorithm 1 describes this procedure in detail for a matrix having M rows and N columns. H_{BN} is saved in the private memory. Because each work-item updates the message in the whole row, H_{BN} is not necessary to be accessed by any other work-items.

The H_{CN} data structure is used in the vertical processing step. It can be defined as a sequential representation of the edges associated with nonnull value in **H**. It is generated by scanning the **H** matrix in a column-major order. H_{CN} is also saved in the private memory. Because each work-item updates the message in the neighbor N/M rows, H_{CN} is not necessary to be accessed by other work-items too.

4.2. Programming the MSA on the OpenCL Grid. Each work-group contains M work-items that represent threads. Instead of the whole matrix **H** or H_{BN}, each work-item can save the necessary part of information of H_{BN} in the private memory, which make access to perform the update faster. Again, the same principle applies to the update of Lq_{nm}^{i} messages.

According to LDPC code length, the CPU on mobile do allocate memory in GPU, including the global memory for storing the check matrix **H**, input data, output data, and the local memory for saving the message data sent from bit nodes to check nodes, marked as Lr_{mn}^{i} and from check nodes to the bit nodes, marked as Lq_{nm}^{i} (Algorithm 3).

In step (2), the compact H_{BN} and H_{CN} are generated in private memory by Algorithms 1 and 2.

The same as the normal MSA algorithm, the loop execution from step (3) will end until the output code word is current or it reaches the maximum loop times.

It executes a horizontal processing, a vertical processing, and a synchronization for all threads in steps (5)–(9). Generally, all threads should be synchronized after the horizontal and vertical processing, but in this algorithm, every work-item takes charge of its own check node, and Lr_{mn}^{i} data is not shared with other work-items, so it is able to cancel the synchronization after horizontal processing to improve performance. Lq_{nm}^{i} data is still shared with all work-items, so the synchronization after vertical processing is retained.

(1) Initialize the work-group size (or number of work-item per work-group).
(2) Generating compact H_{BN}, H_{CN} from matrix **H**
(3) **while** ($\mathbf{cH}^T \neq 0 \cap i < I$)
(4) **as** the work-item k on an M work-group: **do**
(5) **for** all H_{BN}: **do**
(6) $m = k$
(7) update the message sent from BN_n to CN_m
(8) update the message sent from CN_m to BN_n
(9) Synchronize all threads
(10) **for** offset = 0 to $N/M - 1$: **do**
(11) $n = k \times N/M + \text{offset}$
(12) **for** all H_{CN}[offset]: **do**
(13) update the posterior probabilities of BN_n
(14) Synchronize all threads
(15) perform hard decoding

Algorithm 3: MSA kernel executing on the GPU grid.

After the synchronization, it calculates the posterior probabilities of BN_n and every work-item deals with N/M bit nodes as in steps (10)–(13). After the second synchronization, it performs the hard decoding by posterior probabilities, according to the method described in Section 2.

True parallel execution is conducted and the overall processing time required to decode a code word can be significantly reduced as a result, as it will be seen in the next section. More data parallelism can be exploited by decoding several code words simultaneously, but it was not considered in this work.

5. Implementation and Experimental Results

The experimental setup to evaluate the performance of the proposed parallel LDPC decoder on the GPU consists of a PowerVR G6200 with 256 MB global memory and 4 KB local memory and was programmed using the C language and the OpenCL programming interface (version 1.1). In this algorithm, each code word is decoded in a work-group. Because of the limited local memory, only small LDPC code can be used in this test mobile phone. However, the work-group number can be large due to the relatively large global memory on the GPU.

To decode a batch of code words, whose original size is 1 Mbit, the variation in performance is minimal and in Figure 5 we show only the best results achieved. As a 144 × 576 matrix, the work-items per work-group are equal to their row number, which means we use 144 work-items per work-group and 1000 work-groups in this experiment.

The decoding times reported in Figure 5 define global processing times, including data transmission time and decoding time. The decoding time increases along with the increase of iterations. They have a linear relation. The computation capacity of GPU is fully used. The throughput decreases as iterations increase when iterations increase when the size of data for decoding remains the same.

On the mobile device we attach as much importance to the delay as the throughput. Figure 6 shows the decoding delay when the speed is from 10 Kbps to 100 Mbps. With the speed exponential increase, the delay increases but slowly. Actually, the size of data for decoding on the GPU in a decoding cycle is too small that some capacity of GPU is waste and the parallel effect is not obvious with low speed.

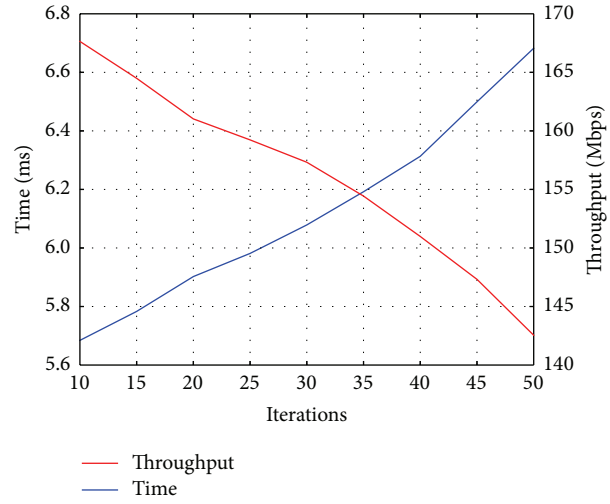


Figure 5: LDPC decoding times on the GPU and corresponding throughput using MSA.

It is obvious that the delay increases when code words for GPU decoding increase. However, the time of a decoding cycle, which is the most important part of the delay, increases but slowly thanks to the more fully use of the computation capacity. The mean time for a code word decreases in higher speed. Thus, it can be applied for some high-speed mobile services, like large file transmission, and delay-sensitive services like video calling.

6. Conclusion

This paper proposes a multicode word parallel LDPC decoder using a GPU on the mobile device running OpenCL. LDPC is widely used in the fourth generation of mobile telecommunications technology, so it is significant to realize high-speed LDPC decoding on the mobile devices.

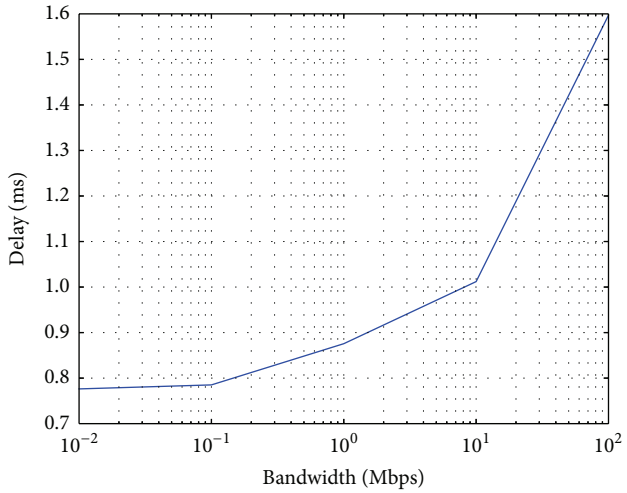


Figure 6: Decoding delay time in when the speed is from 10 Kbps to 100 Mbps.

For an instance, popular video calling software, Skype, has its bandwidth requirements noticed on its official website [18]. The bandwidth required by Skype depends on the type of calls. The minimum speeds required for normal screen sharing video calling, high-quality video calling, and HD video calling are 0.3 Mbps, 0.5 Mbps, and 1.5 Mbps. In the experiment above the decoding delay is 0.84 ms, 0.86 ms, and 0.98 ms in Figure 6. The HD video calling has less than 1 ms delay. It can meet its requirements apparently.

With the software realization of LDPC decoding on mobile devices, LDPC can dynamically change the parameters, including code length, code rate, and the number of iterations. All of them can be fast dynamic switched on OpenCL device, which can quickly deal with all kinds of network environment. With the bad network the code rate can be reduced to improve the ability of error correction, while the code rate can be improved when the network is fine. Compared with the traditional way of hardware decoding, our proposed decoding algorithm based on the software implementation of decoding on the mobile GPU is more efficient, for it can switch at any time according to actual environment.

Competing Interests

The authors declare that they have no competing interests.

References

[1] R. G. Gallager, "Low-density parity-check codes," *IRE Transactions on Information Theory*, vol. 8, no. 1, pp. 21–28, 1962.

[2] D. J. C. MacKay and R. M. Neal, "Near Shannon limit performance of low density parity check codes," *Electronics Letters*, vol. 32, no. 18, pp. 1645–1646, 1996.

[3] T. Zhang and K. K. Parhi, "A 54 Mbps (3,6)-regular FPGA LDPC decoder," in *Proceedings of the IEEE Workshop on Signal Processing Systems (SIPS '02)*, October 2002.

[4] D. Chang, F. Yu, Z. Xiao et al., "FPGA verification of a single QC-LDPC code for 100 Gb/s optical systems without error floor down to BER of 10-15," in *Proceedings of the 2011 Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC '11)*, Los Angeles, Calif, USA, March 2011.

[5] G. Falcão, V. Silva, and L. Sousa, "How GPUs can outperform ASICs for fast LDPC decoding," in *Proceedings of the 23rd International Conference on Supercomputing*, Yorktown Heights, NY, USA, June 2009.

[6] G. Falcão, L. Sousa, and V. Silva, "Massive parallel LDPC decoding on GPU," in *Proceedings of the 13th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP '08)*, pp. 83–90, Salt Lake City, UT, USA, February 2008.

[7] https://www.khronos.org/opencl/.

[8] http://www.nvidia.com/object/cuda_home_new.html.

[9] R. J. McEliece, D. J. C. MacKay, and J.-F. Cheng, "Turbo decoding as an instance of Pearl's 'belief propagation' algorithm," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 2, pp. 140–152, 1998.

[10] J. Zhao, F. Zarkeshvari, and A. H. Banihashemi, "On implementation of min-sum algorithm and its modifications for decoding low-density parity-check (LDPC) codes," *IEEE Transactions on Communications*, vol. 53, no. 4, pp. 549–554, 2005.

[11] R. M. Tanner, "A recursive approach to low complexity codes," *IEEE Transactions on Information Theory*, vol. 27, no. 5, pp. 533–547, 1981.

[12] S. Lin and D. J. Costello, *Error Control Coding*, Pearson Education India, 2004.

[13] D. G. Merrill and A. S. Grimshaw, "Revisiting sorting for GPGPU stream architectures," in *Proceedings of the 19th International Conference on Parallel Architectures and Compilation Techniques (PACT '10)*, Vienna, Austria, September 2010.

[14] V. W. Lee, C. Kim, J. Chhugani et al., "Debunking the 100X GPU vs. CPU myth: an evaluation of throughput computing on CPU and GPU," *ACM SIGARCH Computer Architecture News*, vol. 38, no. 3, 2010.

[15] G. Wang, Y. Xiong, J. Yun, and J. R. Cavallaro, "Accelerating computer vision algorithms using OpenCL framework on the mobile GPU—a case study," in *Proceedings of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '13)*, IEEE, Vancouver, Canada, May 2013.

[16] https://developer.qualcomm.com/category/tags/opencl.

[17] A. Munshi, *The OpenCL Specification*, Khronos OpenCL Working Group 1, 2009.

[18] http://skype.gmw.cn/help/content_69_579.html.