CHALLENGES ON WIRELESS HETEROGENEOUS NETWORKS FOR MOBILE CLOUD COMPUTING

LEI LEI AND ZHANGDUI ZHONG, BEIJING JIAOTONG UNIVERSITY
KAN ZHENG, JIADI CHEN, AND HANLIN MENG,
BEIJING UNIVERSITY OF POSTS & TELECOMMUNICATIONS

ABSTRACT

Mobile cloud computing (MCC) is an appealing paradigm enabling users to enjoy the vast computation power and abundant network services ubiquitously with the support of remote cloud. However, the wireless networks and mobile devices have to face many challenges due to the limited radio resources, battery power and communications capabilities, which may significantly impede the improvement of service qualities. Heterogeneous Network (HetNet), which has multiple types of low power radio access nodes in addition to the traditional macrocell nodes in a wireless network, is widely accepted as a promising way to satisfy the unrelenting traffic demand. In this article, we first introduce the framework of HetNet for MCC, identifying the main functional blocks. Then, the current state of the art techniques for each functional block are briefly surveyed, and the challenges for supporting MCC applications in HetNet under our proposed framework are discussed. We also envision the future for MCC in HetNet before drawing the conclusion.

INTRODUCTION

In recent years, cloud computing has been widely recognized by both industry and academia as the next generation computing infrastructure. Compared with traditional IT infrastructure, it can offer many advantages such as scalability, agility, economic efficiency and so on. Meanwhile, with the rapid deployment of broadband wireless networks and increasing popularity of smart mobile devices (SMDs), more and more users are using SMDs to enjoy the Internet services. So, mobile cloud computing (MCC) is introduced as an integration of cloud computing into the mobile environment. MCC brings new types of services and facilities for mobile users to make full advantages of cloud computing. New mobile applications using MCC can be rapidly provisioned and released with the minimal efforts.

Existing research on cloud computing mostly focuses on problems such as parallelized processing on massive data volumes, flexible virtual machine (VM) management, large data storage and so on. Unlike cloud computing in wireline networks, the mobile-specific challenges of MCC arise due to the unique characteristics of mobile networks, which have severe resource constraints and frequent variations in network conditions [1]. The first major challenge of MCC comes from the limitation of bandwidth and communication latency. When there are large amount of data transferred in a wireless network, the network delay may be increased significantly and become intolerable. Therefore, efficient wireless resource management methods are required to provide Quality-of-Service (QoS) guarantee for the transmission of cloud services. Another serious concern is how to efficiently use the local resources of the mobile devices (MDs). A MD is frequently the entry point and interface of cloud online services. However, it has limited resources such as processing power, memory, and battery lifetime. Thus, it is desirable to offload some computation intensive tasks to the cloud for execution to extend the capabilities of MDs. Since the overhead in energy and response time involved in transmitting the migrated data via wireless networks may be greater than the offloading savings, a judicious decision must be made on whether and which computation tasks to offload.

The traditional way of deploying only macrocells in a wireless network has been proved to be effective only in providing the required coverage and capacity for voice and low data rate services. It is hard to meet the requirements of mobile cloud computing services, e.g., high data rate. Instead, Heterogeneous Network (HetNet), which has multiple types of radio access nodes in a 3GPP Long Term Evolution (LTE) network, e.g., the macro Evolved Node B (eNodeB), pico eNodeB, femto eNodeB, and relay, is widely accepted as a promising technique to meet the increasing traffic demand in broadband wireless networks [2, 3]. In HetNet, the macro eNodeBs ensure the coverage to meet the demands of low speed services, while the small cell eNodeBs such as pico eNodeBs and femto eNodeBs are deployed in the macro cells guaranteeing the

1536-1284/13/$25.00 © 2013 IEEE

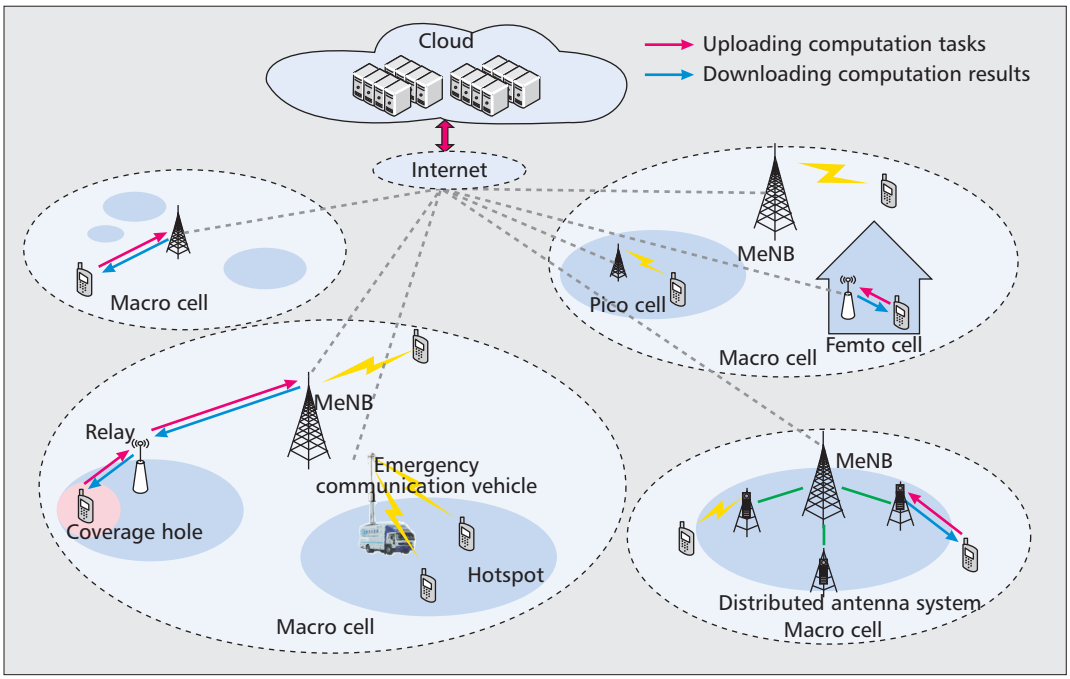

Figure 1. *Access nodes in HetNet for MCC.*

With SMDs becoming increasingly ubiquitous and mobile application economy continuing to show impressive growth, future wireless networks have to support exploding mobile data traffic and should be optimized for mobile broadband traffic. The introduction of MCC puts further strain on wireless networks

hotspot coverage for capacity enhancement. The reduced coverage area of small cells means that the number of MDs sharing the same cell is lower compared to macro cells, giving MDs more freedom of using the bandwidth with lower transmission power.

In this article, we consider smart mobile devices with handheld size and limited computing power, which are efficiently connected to the Internet by HetNet. The full potential of mobile cloud computing applications can be explored only when computation and storage are offloaded into the cloud with acceptable latency and overhead, and doesn't disturb user interactivity with the mobile applications. As the wireless environments may change, the application has to shift its computation workload between MD and cloud without operation interruptions, considering the time-varying wireless connections in HetNet. To deliver cloud services in HetNet environment, we have to face several problems. For example, wireless networks are not always reliable to guarantee cloud service delivery due to

- User mobility
- Propagation effects of wireless channels
- Traffic load variations affecting both the multi-user resource sharing within a serving cell and interference across neighboring cells

So, MDs have to collaborate with the eNodeBs in making offloading decisions. Also, the eNodeBs have to guarantee the QoS of cloud service transmission with proper radio resource management (RRM) decisions such as admission control, cell association, power control and resource allocation. To the best of the authors' knowledge, there have been quite few work in the existing literatures on these problems in HetNet for MCC applications.

The scope of this article is hence to examine how SMDs can work so as to enjoy MCC applications in HetNet reliably. We first introduce a framework of HetNet needed to fulfill MCC service requirements. Then, several challenges are discussed for HetNet in supporting MCC applications. We also envision the future for MCC in HetNet before drawing the conclusion.

Framework of HetNet for MCC

In MCC, MDs connect to the Internet and then to the cloud via the wireless connections. Therefore, the wireless network is an important component of the MCC architecture. With SMDs becoming increasingly ubiquitous and mobile application economy continuing to show impressive growth, future wireless networks have to support exploding mobile data traffic and should be optimized for mobile broadband traffic. The introduction of MCC puts further strain on wireless networks since additional data associated with cloud services between the MDs and the cloud have to be transported via the wireless networks with potential QoS and/or power constraints.

HetNets offer promising solutions for these challenges. In this section, we discuss the framework of HetNet for MCC, identifying the main functional blocks. We especially focus on the traditional centralized cloud, of which the computing resource pool is placed in the remote cloud and MDs can access the resources by connecting to the existing wireless network. As illustrated by Fig. 1, the HetNet mainly consists of two components, i.e., macrocells and small cells, where the former provide mobility while the latter boost coverage and capacity.

Access Nodes in HetNet

Macro/Micro Cells — The inter-site distance (ISD) between two macro or micro eNBs (MeNBs) is usually no less than 500 meters, which can pro-

In order to enable applications and systems to continue to operate in dynamic wireless environments, mobile cloud applications must dynamically adjust the computing functionalities between the MD and cloud depending on the changes in mobile environments.

vide a ubiquitous coverage for MDs in a wide range of area, and support high mobility MDs by minimizing the handover frequency. However, due to factors such as channel fading and traffic congestion, the connectivity between MDs and MeNB has relatively low data rate and is unstable.

Small Cells — Since SMDs and cloud applications are hungry for a high-speed and stable connectivity to the cloud, the small cells become a better choice.

Pico Cells — The low-powered radio access nodes, which have a coverage range of about 200 meters or less, are deployed in Pico cells. Their access is open to all cellular MDs. Usually, the pico cells are used to provide hotspot coverage in malls, airports or stadiums.

Femto Cells — A femto cell has a small, low-power base station, typically designed for MDs in a home or small business, whose access node is referred to as Home eNB in LTE. Similar to WiFi, the coverage range for a femto cell is less than 100 meters. The access to a femto cell can be either restrained to a limited set of MDs in the femto cell's access control list (closed access mode), or open to all cellular MDs (open access mode).

Distributed Antenna System — A distributed-antenna system is a network of spatially separated antenna nodes connected to a common source via a transport medium that provides wireless service within a geographic area or structure. It can create small virtual cells by distributing antennas of macro eNBs across entire cell. The antennas are connected to a common processing unit via fiber.

Relay Nodes — The relay nodes are low power base stations that can provide coverage and capacity enhancement to macro cells at the cell edge. A relay node is connected to its Donor eNB (DeNB) via a radio interface. Due to the instability of DeNB coverage, MDs in some locations of the macro cell may have a failure in access to the DeNB. The deployment of relay nodes can solve this problem effectively.

Main Functional Blocks to Support MCC

Offloading Decision (OD) — With the continuous enhancement of MDs' capabilities, they gradually become the primary tools for accessing service clouds to maximize their functionalities. The two most widely studied cloud services are cloud storage services and cloud processing services. Computation offloading is an enabling technology for the resource limited SMDs to access the latter services, where processing capabilities of SMDs are augmented by outsourcing computation intensive components of the mobile applications to the resourceful servers in clouds. The main purposes of computation offloading are to save the SMDs' power consumption and speed up the application processing. However, the above two goals may not always be achieved depending on the particulars of the computation task, the server load, and the connectivity to the network. For example, a task with high computation and low communication requirements is more likely to benefit from offloading than a task with low computation and high communication requirements. Therefore, a judicious decision has to be made on whether to offload a computation task or not.

In order to enable applications and systems to continue to operate in dynamic wireless environments, mobile cloud applications must dynamically adjust the computing functionalities between the MD and cloud depending on the changes in mobile environments. However, it is hard to predict the data transmission delay of wireless network and the power consumption of MD for computation offloading when an offloading decision has to be made. Existing dynamic offloading approaches mostly use the measured metrics before and at the decision instant as a reference and assume that the network condition will remain constant during the execution interval, which is apparently inaccurate [6, 7]. Another method is to model the wireless channel as a Markov chain to capture its fading effects [8, 9]. However, since the wireless resources are usually shared by multiple MDs in HetNet, the data transmission delay and power consumption of a MCC application are also dependent on the traffic load and interference conditions of the associated cell, which are neglected in the above studies.

In this article, we propose a framework of HetNet for MCC as shown in Fig. 2, where the HetNet offers to the offloading decision function a range of wireless transmission services with different serivce classes and monetary prices, and the offloading decision is made considering both the offloading gain and the cost of using the HetNet when a Service Level Agreement (SLA) is established with it. Specifically, the SLAs are based on the estimated average values or worst case values with a given violation probability of two attributes, i.e., power consumption and wireless transmission delay. The former is the total power consumed in the MD and the latter is the total delay via the HetNet for transmitting the computation tasks to the cloud and receiving the computation results from the cloud. Although there are some existing research on the QoS framework for MCC [4, 5], we focus on the interactions between the offloading decision function of MCC and the radio resource management function of HetNet, which haven't been addressed in prior work to the best of our knowledge.

Our general model for the SLA between the OD function and the HetNet consists of a tuple $\langle L_i, L_r, P, T, C, x_p, x_t \rangle$, where L_i and L_r are the maximum amount of allowed data for computation tasks and results to be transmitted via the wireless uplink and downlink, respectively. P and T are the average or worst case values of power consumption and wireless transmission delay, respectively. C is the service cost for data transmission in HetNet, while x_p and x_t are the penalty rates that HetNet will refund its users with for possible violations of the service class power consumption and wireless transmission delay, respectively. The OD function decides whether to offload a computation task and with what service class. If the OD function makes a positive

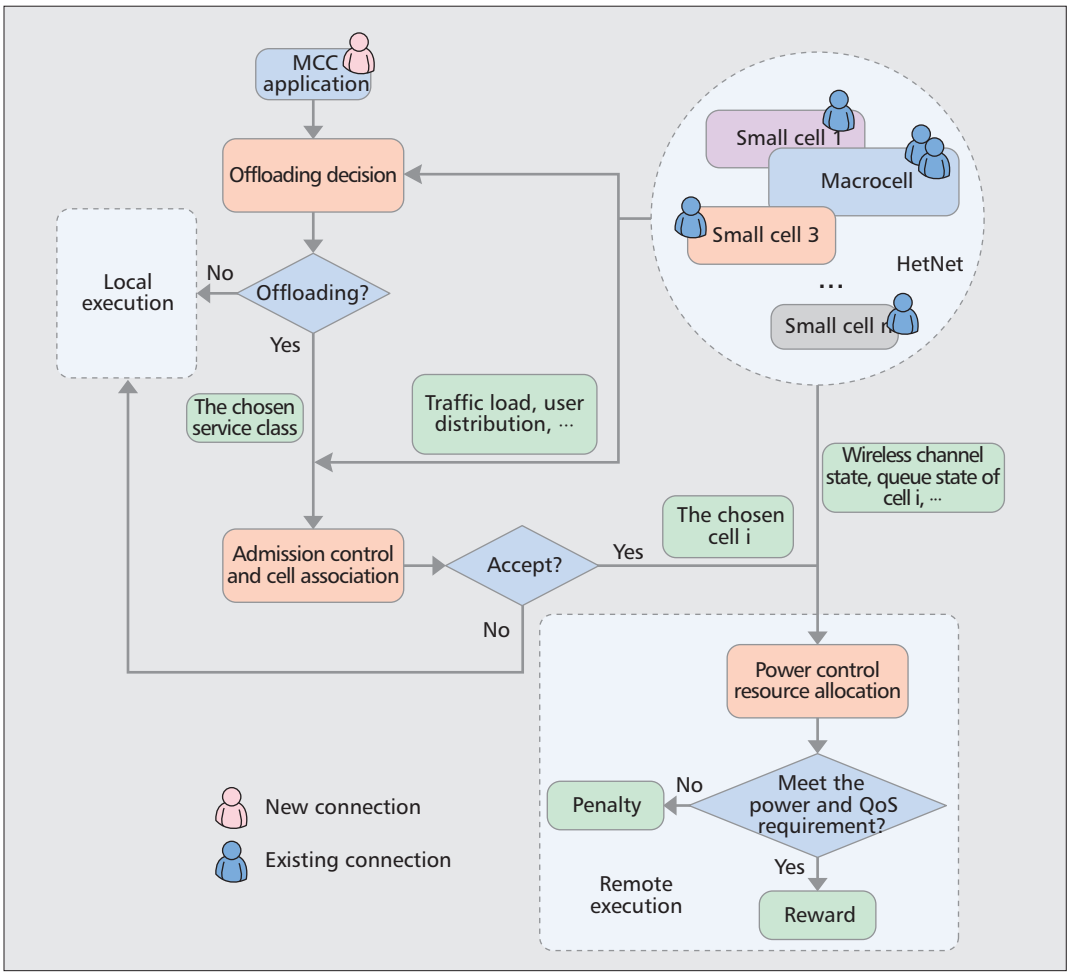

Figure 2. *Framework of HetNet for MCC.*

offloading decision, it sends the required service class to the HetNet, which decides whether to accept the offloading requirement and associate the SMD with a specific cell for data transmission by the admission control and cell association function. Once the HetNet accepts the offloading requirement and establishes the SLA with the OD function, it provides QoS guarantee by the power control and resource allocation function.

Admission Control (AC) and Cell Association (CA) — The OD function in MCC requires that the HetNet provides data transmission service with a power and QoS (e.g., delay) guarantee. However, as the network traffic load varies with time, admission control is needed in HetNet to balance the goals of maximizing bandwidth utilization and ensuring sufficient resources for the offloading requirements with power and QoS constraints. The HetNet uses AC policies to determine the admissibility of an offloading requirement from the OD function once it requests to establish a SLA for transferring the offloading data. The basic principle of the AC policy is to admit an offloading requirement only if the power and QoS constraints of the new connection can be met while the power and QoS guarantees for the existing connections will not be jeopardized. Moreover, since admitting the new offloading requirement may introduce interference to the neighboring cells which reuse the same frequency with the serving cell, the existing connections in both the serving cell and its neighboring cells should be taken into account. Finally, the reward, cost and SLA violation penalties should all be considered in an optimized AC policy.

Since there are multiple types of access nodes in HetNet, e.g., macro eNBs and pico eNBs, cell association (CA) has to decide which access node should be associated with the SMD for the offloading data transmission. The simplest policy is to associate the SMD with the access node whose signal is received with the largest average strength. Other more complex policies consider the load balancing and interference between different cells.

In our framework of HetNet for MCC, the AC and CA functions have to be performed jointly. When selecting the associated cell for a SMD with offloading requirement, the CA function needs to consider whether this requirement can be admitted by the AC function of a candidate cell, and an optimized cell is selected only among those cells considered eligible by the AC function.

Power Control (PC) and Resource Allocation (RA) — Once an offloading requirement is admitted into the HetNet, power control and resource alloca-

When selecting the associated cell for a SMD with offloading requirement, the CA function needs to consider whether this requirement can be admitted by the AC function of a candidate cell, and an optimized cell is selected only among those cells considered eligible by the AC function.

Functional Blocks	State-of-Art	Challenges
Offload Decision	Improved versions of a simple model [10] from the following aspects: • Optimization objectives, • Energy and speed tradeoffs, • Application partition, • Dynamic variation of environment.	• The impact of wireless network traffic load on the power consumption and execution time for remote execution is ignored; • The cost charged and the penalty rates refunded by the wireless network for providing and violating the QoS guarantee are not considered.
Admission Control	Criteria for accepting/rejecting a connection: • Transmit power, • Network load, • Achievable throughput, • QoS. Design objectives: • Minimization of both false rejections and false admissions.	• The power consumption and wireless transmission delay should be considered as the criteria for acceptance/rejection/association decision; • Ensure that the required QoS be met when a MD roam to a neighboring cell, especially during the download phase.
Cell Association	Design objectives: • Best SNR, • Load balance, • Throughput maximization under static interference, • Performance optimization under dynamic interference.	• The interference introduced to the neighboring cells should be considered.
Power Control	PC in LTE: • Open-loop; • Closed-loop. PC in HetNet: • Outgoing interference based, • Incoming interference based.	• An uplink scheduling algorithm with delay and power guarantee has to be designed; • The scheduling algorithm should consider the mix traffic scenario.
Resource Allocation	Optimization objectives of downlink scheduling: • Maximum throughput and fairness tradeoff, • Queue stability, • Minimum delay or queue overflow probability, • Energy efficiency and delay tradeoff, • A mixture of optimization objectives for different traffic. Uplink scheduling • The resources assigned to the same MD must be contiguous in the frequency domain. Interference coordination: • Semi-static vs. dynamic.	• The scheduling algorithm should be optimized jointly with the power control and interference coordination functions in HetNet.

Table 1. *State-of-art and challenges of HetNet for MC.*

tion functions are used to ensure its QoS guarantee whenever feasible. Resource allocation decides which MD is going to transmit on each time-frequency resource in a cell, and power control determines the amount of power allocated to each scheduled MD on the allocated resources. The wireless transmission is performed on a slot-by-slot basis. In each time slot t, power control and resource allocation algorithms normally make use of the current channel state $S(t)$ and queue state $Q(t)$ information when making a decision. Given $S(t)$, $Q(t)$, and a power control and resource allocation action $A(t)$ for uplink or downlink transmission, the corresponding uplink or downlink instantaneous transmission rate $R(t)$ and transmit power $P(t)$ of each MD are determined, where $R(t)$ can be seen as its service rate in a queuing system. The wireless transmission delay of an offloading requirement T corresponds to the sum of offloading SMD's sojourn time in the queueing system for uplink transmission T_u and downlink transmission T_d, where the sojourn time of uplink or downlink transmission includes both the waiting time Tuw for uplink or T_{dw} for downlink and service time T_{us} for uplink or T_{ds} for downlink. On the other hand, the power consumption of an offloading requirement on a SMD corresponds to the sum of power consumed during uplink and downlink transmission, and also during the waiting period for remote cloud computation. Specifically, the power consumption for uplink transmission includes both the power consumed for transmitting computation task with power level $P(t)$ during service time T_{us} and for receiving downlink control signaling during waiting time T_{uw}. Meanwhile, the power consumption for downlink transmission includes the power consumed for

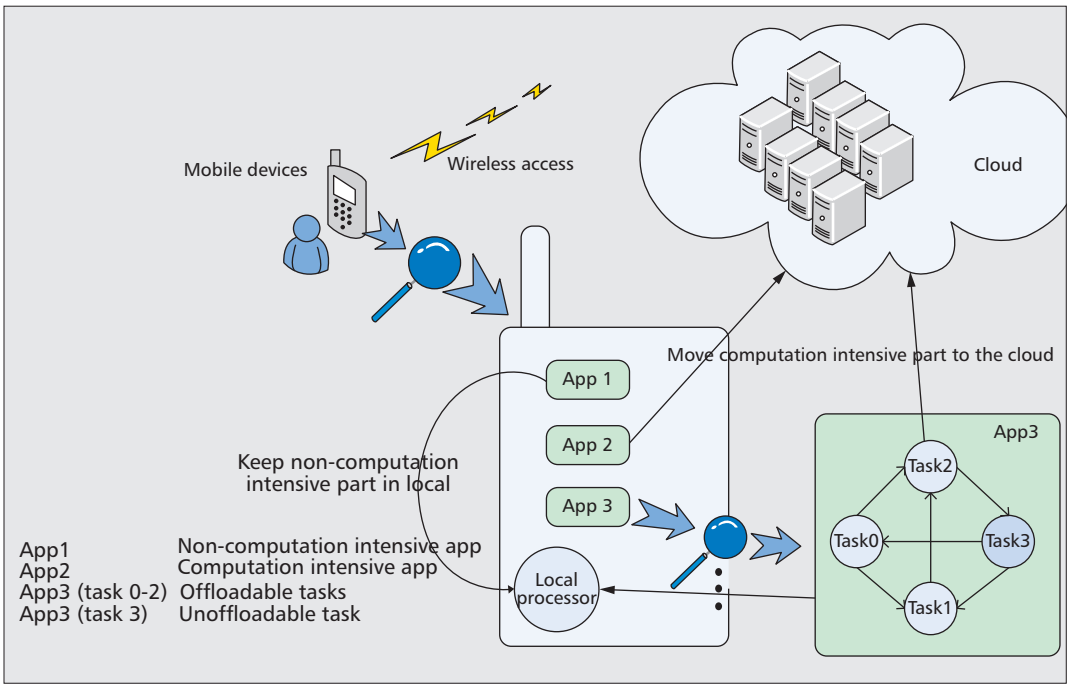

Figure 3. *An sample of partial offloading procedure.*

receiving the computation results and downlink control signaling during the downlink sojourn time T_d. Therefore, the power consumption and wireless transmission delay incurred by an offloading requirement are determined by the power control and resource allocation algorithms.

Challenges in HetNet for MCC

In our proposed framework of HetNet for MCC, every functional block needs to make a decision based on the current system state to obtain the optimized performance. However, the current state-of-art techniques for each functional block face several challenges in order to make the optimized decision, which are discussed in this section and summarized in Table 1.

Offloading Decision

As shown in Fig. 3, there are generally two kinds of offloading techniques, called as full offloading and partial offloading.

Full Offloading — When full offloading, all computation tasks of mobile applications are moved from the local MD to the remote cloud. This may significantly reduce the implementation complexity of MDs, which makes the MDs lighter and smaller. However, different computation tasks of a single application may have different characteristics that make them more or less suitable for offloading. Therefore, full offloading is not always the optimal choice for MDs.

Partial Offloading —Due to its flexibility, partial offloading has gained more attention in the field of offloading research, where an application is partitioned into several computation tasks including un-offloadble ones (e.g., those handle user interaction) and offloadable ones. Before a task is executed, it may require certain amount of data from other tasks. Therefore, data migration via wireless networks is needed if this task and the other tasks are executed at different locations (i.e., MD or cloud server node).

For both offloading techniques, a controller in the SMD needs to determine whether to offload the application or some of its computation tasks to the cloud when there is a request for application execution. The offloading decision can be either made statically before execution, or dynamically at runtime to cope with the dynamic processing loads on SMD and cloud server nodes and the changing wireless channel status and wireless network traffic load.

In [10], a simple model is provided to analyze the energy saving achieved by computation offloading. Suppose a computation task requires C instructions, define

M as the speeds in instructions per second of the MD,

S as the speeds in instructions per second of the cloud server,

D as the bytes which the cloud server and MD exchanged,

B as transmission rate of the MD via wireless network,

P_c as the energy consumed by MD for computing,

P_i as the energy consumed by MD while being idle,

P_{tr} as the energy consumed by MD for sending and receiving data.

Then, the amount of energy saved is

$$P_c \times \frac{C}{M} - P_t \times \frac{C}{S} - P_{tr} \times \frac{D}{B}.$$

A basic principle for offload decision is derived, i.e., offloading is beneficial when large amounts

In our proposed framework of HetNet for MCC, every functional block needs to make a decision based on the current system state to obtain the optimized performance. However, the current state-of-art techniques for each functional block face several challenges in order to make the optimized decision.

The most important and widely studied QoS metric is the response delay or execution time of the application, where the remote execution time includes transmission delay of both the wireless network and the Internet, and task execution delay of the remote cloud.

of computation C are needed with relatively small amounts of communication D. However, this simple model is insufficient in providing an optimized solution for offloading. Other existing work can be seen as improved versions of this simple model from various aspects, which are discussed as follows.

Optimization Objectives — The simple model tries to get the maximum benefit by balancing the computation and communication power. However, good QoS performance of the services has to be guaranteed from the view of user experience. In other words, the computation-intensive tasks can be offloaded to the cloud only when the QoS is ensured. The most important and widely studied QoS metric is the response delay or execution time of the application, where the remote execution time includes transmission delay of both the wireless network and the Internet, and task execution delay of the remote cloud. Generally speaking, there are five classes of optimization objectives considering all possible combinations of these two metrics:

1 Optimizing energy consumption,
2 Optimizing execution time,
3 Optimizing energy consumption under execution time constraint,
4 Optimizing execution time under energy consumption constraint,
5 Optimizing both energy consumption and execution time.

Objective 3) is addressed in [6, 7, 9, 11]. Objective 5) is a nontrivial multi-objective optimization problem. Reference [8] addressed objective 5) by setting the optimization objective as the total energy consumption plus the product of total remote execution time and a predefined delay/wait cost. By varying the delay cost, different power/energy versus delay trade-off balances can be obtained.

Energy and Speed Tradeoffs — In practice, the value of M is dependent on the value of P_c since low speed of the MD's processor can provide significant energy savings, while the value of B is dependent on the value of P_{tr} since the transmission rate of a wireless link is a function of its signal-to-interference-noise-ratio (SINR) and transmission bandwidth, where the SINR value is determined by the transmit power, channel gain, and noise plus interference power. The correlation between every pair of variables is not explored in the simple model. In [8], the optimal values of P_c and P_{tr} are determined where it is assumed that $B = P_{tr}/(P_{tr} + i)$ with i representing the uplink channel stress. The closed-form solutions of two constrained dynamic optimization problems are derived in [9] to set the optimal clock frequency of local processor and data transmission rate of each time slot within the application execution interval, so that the power consumption is minimized within a delay constraint. The offload decision is made by choosing the execution location with smaller energy consumption.

Application Partition — In the simple model, the amount of data exchange between the server cloud and MD D is assumed to be a fixed value. However, this is not true in partial offloading. As shown in Fig. 3, since the execution of computation task 2 needs data input from computation task 0 and 1, the amount of data exchange D for the remote execution of task 2 depends on the execution locations of task 0 and task 1. In [6] and [7], a call graph and a profile tree are constructed, respectively, to provide a global view of the application behavior and specify the amount of data migration needed between different tasks. A dynamic partition algorithm based on Lyapunov optimization is proposed in [11] to make the offloading decision for all the computation tasks.

Dynamic Variation of Environment — The simple model targets to find the optimal strategy by using the transmission rate and execution time of computation statically. However, as the execution time changes at different execution instances due to varying workload of the MD and cloud server nodes, it is hard to find an accurate static information. Moreover, the dynamic variation of the wireless channel condition along with the changing traffic load of the wireless network make the transmission rate highly dynamic. In [11], dynamic offloading is studied considering the dynamic arrival of application execution requirement and the variation of data transmission rate at different execution period. However, the data rate is assumed to remain constant within a single application execution period. In [8] and [9], the wireless channel variation within the execution period of an application is modeled by two-state Markov chains.

Although existing work has made the above improvements to the simple model from various aspects, there are still some open problems, especially when the offloading data is transmitted via HetNet. First, the offloading decision is made based on an estimation of the SMD's power consumption and execution time when the application is executed locally or remotely. Existing work seldom considers the impact of wireless network traffic load on the above two parameters for remote execution. However in practice, the traffic load has an important influence over the remote execution time through its impact on the waiting time of the queuing system corresponding to the wireless transmission. On the other hand, the power consumption for remote execution may also be impacted by the traffic load of the SMD's serving cell and neighboring cells. This is because the power consumption is dependent on the channel gains of the uplink wireless channel at the scheduling instants of the SMD, which are likely to be affected by the traffic load. For example, the probability of the SMD being scheduled in a time slot with more favorable channel conditions becomes larger with the increasing traffic load due to the multiuser diversity effect if the channel-aware scheduling algorithms are applied. Moreover, increased traffic load at the neighboring cells may result in the larger interference opportunities for the SMD's transmission, which in turn leads to larger power consumption. We perform simulation on a HetNet consisting of one macrocell and three pico cells and also on a macro-only network, where users with computation

offloading requirement of 10Kb dynamically arrive to both networks. The users leave the system when they finished transmission. The system bandwidth of both HetNet and macro-only network are 10MHz. We adopt the round robin scheduling algorithm for both the HetNet and macro-only network. Moreover, cell range expansion is used for cell association in HetNet. Figure 4 shows the average delay and power consumption per user for transmission of the offloading data with varying arrival rates. It reveals that the delay and power consumption increase with the arrival rate in HetNet, and the transmission delay in the HetNet is smaller than that of the macro-only network at the cost of little increase in the power consumption. Apart from the above problem, existing work in literature on offloading decision seldom considers the cost charged and the penalty rates refunded by the wireless network for providing and violating the QoS guarantee.

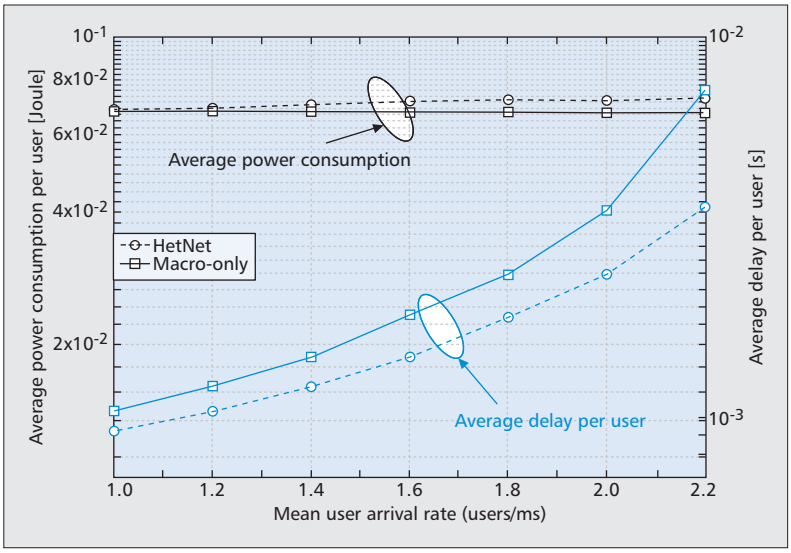

Figure 4. *Average delay and power consumption of offloading data transmission in HetNet and Macro-only network.*

Admission Control and Cell Association

Admission Control — Admission control algorithms are utilized to ensure that admittance of a new connection into a resource constrained network does not violate the SLAs guaranteed by the network to both the new connection and the already admitted connections. There are several criteria that AC algorithms use for accepting or rejecting a connection, e.g., transmit power, network load, achievable throughput, QoS, and so on [12]. The design objectives of the AC algorithms are the minimization of both false rejections and false admissions. False rejections lead to the unnecessary blocking of a connection whose requirements could be met by the network if admitted. On the other hand, false admission of a new connection which should not have been admitted into the network leads to the dropping or QoS degradation of the new and/or existing connections due to the limited network capacity.

In HetNet, AC algorithms not only need to check the status of its serving cell, but also the status of the adjacent cells due to two reasons. First, AC algorithms need to ensure that sufficient resources are available for handovers. It is well known that handover originated connections are more sensitive and should have higher priority than new connections, because dropping an existing connection is undesirable from the user's point of view. Second, the amount of interference introduced by the new connection to the adjacent cells should also be taken into account when making an accept/reject decision. Although there have been some research on the AC algorithms for heterogenous networks composed of different radio access networks, e.g, cellular, WiFi, etc., there are few studies on the AC of HetNet with small cells. In [13], the admission control problem for hybrid access in OFDMA-based femtocell networks is studied, where macrocell MDs can establish connections with femtocell eNodeBs to improve their QoSs.

Cell Association — In HetNet, a fundamental problem is the one of associating MDs, either with the macro eNodeB or with a low power eNodeB (pico, femto, etc.). The dynamic cell association takes its decision by considering the long-term signal measurements and the total number of MDs within the network. Therefore, it is not expected to be performed very often, but only when the conditions have significantly changed. Compared with macrocell-only networks, HetNets are much more sensitive to the cell association policy because of the massive disparities in cell sizes. There are a few recent investigations on the cell association problem with the following objectives:

Best SNR: In LTE networks, a MD is associated with the eNodeB whose signal is received with the largest average strength. However, it leads to a load imbalance between the macro and low power eNodeBs, which limits the MD throughput.

Load Balance: A new method referred to as Range Expansion has been proposed within 3GPP by adding a fixed bias to the measured received (Rx) power of pico eNodeBs. Thus, a larger fraction of MDs is served by low power eNodeBs, and a better load balancing is achieved between the macro and low power eNodeBs. However, with this static modification to the cell association, certain MDs that are connected with the low power eNodeBs receive stronger signals from the macro eNodeB and the achieved better load balancing may not compensate a lower SINR resulting from the increased interference level.

Throughput Maximization under Static Interference: The goal is to associate MDs to one of the macrocell or small cell eNodeBs in order to maximize the sum rate or minimum rate of all MDs [14]. These methods improve on the Range Expansion method by exploiting the tradeoff between load balancing and interference. However, the interference model in these studies are assumed to be static, i.e., neighboring eNodeBs are always active and causing interference.

- Performance Optimization under Dynamic Interference: In a small cell network scenario, the number of MDs associated with a low

Through cooperation between different kinds of radio access nodes, HetNet can perform power control and resource allocation of multiple cells simultaneously to achieve maximum throughput and meet the power and QoS requirements for MCC applications.

power eNodeB is small and thus the interference situations may dynamically change, i.e. some low power eNodeBs may not be active in a certain period of time since there are no data packets to transmit for the MDs associated with them. A cell association policy for HetNet is proposed in [15] under the dynamic interference assumption which aims to optimize the average throughput or delay for the MDs.

When considering the support of MCC applications, several challenges exist for the AC and CA functions. First, in both the AC and CA functions, power consumption and transmission delay should be considered as the criteria for acceptance/rejection/association decision. Moreover, since the transmission delay involves both the computation task upload delay and computation results download delay, both the uplink and downlink traffic and wireless channel conditions have to be taken into account. Second, as a MD may roam to a neighboring cell during the offloading data transmission period, the AC function needs to ensure that the required QoS be met under this situation. In particular, as it is more intolerable for an offloading task to be interrupted at the result download phase than at the task upload phase due to the amount of time and resources already involved in the offloading data transmission, the avoidance of dropping a download connection for offloading should be considered with a higher priority. Finally, the interference introduced to the neighboring cells should be considered in the AC and CA functions.

Power Control and Resource Allocation

Through cooperation between different kinds of radio access nodes, HetNet can perform power control and resource allocation of multiple cells simultaneously to achieve maximum throughput and meet the power and QoS requirements for MCC applications. In HetNet, many wireless links potentially interfere with each other because their allocated resources are not fully orthogonal in order to better reuse the limited radio resources. For example, interference scenarios of femtocell can be considered as a special case of that of HetNet. The interference between macrocell and femtocell, i.e., inter-tier interference, arises from the fact that femtocells may utilize the spectrum already allocated to the macrocell. Meanwhile, all the femtocells can share the same radio resources for improving the resource usage, which may cause the interference between femtocells themselves, i.e., intra-tier interference. Without proper interference management methods, significant power is likely to be wasted in order to maintain an acceptable user performance. For example, high transmit power is usually radiated from a cell edge outdoor MD associated with a macrocell eNodeB to provide reliable communications. If no proper interference management method is applied, interference is possibly generated to nearby indoor MDs connected to a femtocell eNodeB in case that the whole or a part of frequency band is shared between the femtocell and macrocell. Thereupon, the femtocell MDs have to increase their transmission power to maintain the communication with their indoor femto eNodeB. In this situation, the overall energy efficiency of the network becomes even worse after deploying the femtocells. Interference management is therefore an important method to capitalize on the potential energy efficiency in HetNet. Various interference management techniques, such as power control and interference coordination, can be used to provide the promising energy-efficiency performances.

Power Control — Since the computation offloading decision is mainly concerned with the power consumption on the SMD for transmitting/receiving the offloading data to/from the cloud, we will focus on the uplink power control mechanisms as the power consumption for data reception is far less than that for data transmission.

In 3GPP LTE, Fractional Path Loss Compensation Power Control (FPC) mechanism is used for uplink power control, which is open-loop and based on the path loss measurement done by the MD but controlled with a factor α by the network. Meanwhile, a closed-loop power control mechanism can also be applied, where measurements by the eNodeBs are used to generate transmit power control commands that are send to the MD as part of the downlink control signalling [16]. However, the above power control mechanisms are based on an important assumption that there is a correlation between being close to the serving cell and being far away from neighbor cells, which may not be true in HetNet. Therefore, if the macrocells and small cells are co-channel deployed, the varying uplink interference across cells may impact the effectiveness of the FPC scheme. As a result, uplink power control in HetNet has been studied in recent years. For example, the concept of outgoing interference based power control has been proposed, where a MD generating high interference to other cells should transmit at low power, and vice versa. Also, a MD can also adjust its transmit power based on the incoming interference, and transmit at high power level if it is subject to serious interference [17].

Resource Allocation — Wireless network scheduling is an important resource allocation function. Since the MCC applications are usually delay and power sensitive, the scheduling function should be designed with the corresponding objectives. Most research on scheduling has been treating the downlink scenario with the following optimization objectives. Some comprehensive surveys are provided in [18].

Maximum throughput and fairness tradeoff. The channel variation due to the fast fading effects of wireless channel is exploited as an opportunity by the scheduler, which tends to choose the MD with the best channel condition for transmission at every time slot. However, it should also provide a fair share of resources to all the MDs in the long term.

Queue stability. The aim of the scheduler is to ensure the queues' stability without any knowledge of arrival and channel statistics if indeed stability can be achieved under any policy. Therefore, the scheduler takes into not only the channel state but also the queue state, and

the examples are MaxWeight and Exponential rule, etc.

Minimum delay or queue overflow probability. Since stability is a weak form of performance optimality, some research work focus on scheduling polices that minimize the overall average delay (per data unit) seen by the MDs; or scheduling policies which minimize the probability that either the sumqueue or the largest queue overflows a large buffer, such as the Log rule.

Energy efficiency and delay tradeoff. Energy efficiency is an important concern in the design of modern wireless systems. However, it is often achieved at the cost of increasing delay due to two reasons. First, since the required transmission power is a convex function of the communication rate, this implies that transmitting data at low rates over a longer duration is more energy efficient as compared to high rate transmissions. Second, since the wireless channel is time-varying and as good channel conditions require less transmission power, scheduling MDs only at good channel conditions leads to reduced energy cost. Therefore, the problem of minimizing energy subject to delay constraint or minimizing delay subject to energy constraints are treated for single user [19] and multiuser wireless systems [20].

A mixture of optimization objectives for different traffic. Since there is usually a mixture of elastic and real-time traffic in realistic wireless networks, several studies that addresses the problem of simultaneously supporting these traffic have been studied, ensuring that the real-time traffic receiving their desired QoS while the elastic traffic achieves the maximum possible throughput.

Compared with the large amount of research on downlink scheduling, considerably less work has been dedicated to the uplink. LTE uses the SC-FDMA radio access technology for its uplink transmission. As a result, resources assigned to the same MD must be contiguous in the frequency domain. This contiguous allocation constraint limits the scheduling flexibility and makes the above optimization problems more complex for the uplink scenario. As the optimal solutions would mostly be NP-hard, the proposed scheduling algorithms are often based on heuristics yielding reasonable system performance under practical circumstances [21].

The resource allocation function should also consider interference coordination by allocating different resources between neighboring eNodeBs in the time or frequency domains in order to mitigate co-channel interference. The main challenge lies in the fact that the location and coverage areas of small cells such as femtocells are uncertain, and the traffic load in small cells are less aggregated due to the much fewer served MDs per cell compared to the macro cells. The coordination requires communication between different network nodes in order to (re)configure radio resources. Based on the needs of the inter-sites communication interval, most interference coordination schemes may be categorized into two classes, i.e., semi-static and dynamic schemes.

When considering the support of MCC applications with power and QoS constraints, the above power control, scheduling and interference coordination methods in (heterogeneous) wireless networks are inadequate. The scheduling algorithms should consider the delay and energy efficiency tradeoff for the support of MCC applications. However, the existing studies mostly focus on the downlink scenario while the MCC application cares about the power consumed by the MD instead of the eNodeB. Therefore, an uplink scheduling algorithm with delay and power guarantee has to be designed, where the contiguous allocation constraint has to be considered. Furthermore, considering those MDs in the wireless networks without MCC applications, the scheduling algorithm should consider the mix traffic scenario. Finally, the scheduling algorithm should be optimized jointly with the power control and interference coordination functions in HetNet to achieve the maximum spectrum efficiency while guaranteeing the power and QoS requirements for the offloading connections.

Conclusion

HetNet consisting of macrocells and small cells is expected to provide the wireless connection anywhere and anytime. With the rapid development of HetNet, the mobile users may enjoy the cloud services with good user experience regardless of spectrum scarcity. For this purpose, we propose a framework of HetNet for MCC, which is based on the power and QoS negotiation between the offloading decision module and the wireless HetNet. However, several challenges have to be addressed before the users can truly enjoy the life from MCC applications, which are discussed in the article. Future research on HetNet for MCC will be conducted based on a tight coupling of the unique characteristics of the MCC applications and the wireless heterogeneous networks. Specifically, the computation offloading decision should be made considering the power and QoS guarantee from the HetNet along with the cost and refunded penalty rates. The wireless HetNet, on the other hand, should apply the related radio resource management functions, e.g., admission control and resource allocation, to provide the power and QoS guarantee of the MCC applications whenever feasible.

With the rapid development of HetNet, the mobile users may enjoy the cloud services with good user experience regardless of spectrum scarcity. For this purpose, we propose a framework of HetNet for MCC, which is based on the power and QoS negotiation between the offloading decision module and the wireless HetNet.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 61272168), Program for New Century Excellent Talents in University (NCET-11-0600), the Fundamental Research Funds for the Central Universities (No. 2012JBM003), the State Key Laboratory of Rail Traffic Control and Safety (Contract No. RCS2011ZT005), Beijing Jiaotong University, and the Key Grant Project of Chinese Ministry of Education (No. 313006).

References

[1] H. T. Dinh *et al.*, "A Survey of Mobile Cloud Computing: Architecture, Applications, and Approaches," *Wireless Commun. and Mobile Computing*, Oct 2011.

[2] K. Zheng *et al.*, "Interference Coordination for OFDM-based Multihop Cellular Networks Towards LTE-Advanced," *IEEE Wireless Commun.*, vol. 18, no. 1, Feb. 2011, pp. 54–63.

[3] K. Zheng *et al.*, "Energy-Efficient Wireless In-Home: the Need for Interference-Controlled Femtocells," *IEEE Wireless Commun.*, vol. 18, no. 6, Dec. 2011, pp. 36–44.
[4] Y. Ye *et al.*, "A Framework for QoS and Power Management in A Service Cloud Environment with Mobile Devices," *IEEE Int'l. Symp. Service Oriented System Engineering (SOSE)*, 2010, June 2010, pp. 236–43.
[5] M. Abundo, V. Cardellini, and F. Lo Presti, "An MDP-based Admission Control for a QoS-Aware Service-Oriented System," *2011 IEEE 19th Int'l. Wksp. Quality of Service (IWQoS)*, June 2011, pp. 1–3.
[6] E. Cuervo *et al.*, "MAUI: Making Smartphones Last Longer with Code Offload," *Proc. 2010 Int'l. Conf. Mobile Syst.*, App., Services, 2010, pp. 49–62.
[7] B. G. Chun *et al.*, "CloneCloud: Elastic Execution Between Mobile Device and Cloud," *Proc. 6th Conf. Computer Systems (EuroSys)*, Apr. 2011, pp. 301–14.
[8] S. Gitzenis and N. Barnbos, "Joint Task Migration and Power Management in Wireless Computing," *IEEE Trans. Mobile Computing*, vol. 8, no. 9, Sept. 2009, pp. 1189–204.
[9] Y. Wen, W. Zhang, and H. Luo, "Energy-Optimal Mobile Application Execution: Taming Resource-Poor Mobile Devices with Cloud Clones," *Proc. IEEE Infocom*, 2012, pp. 2716–20.
[10] K. Kumar and Y.-H. Lu, "Cloud Computing for Mobile Users: Can Offloading Computation Save Energy?," *Computer*, vol. 43, no. 4, Apr. 2010, pp. 51–56.
[11] D. Huang, P. Wang, and D. Niyato, "A Dynamic Offloading Algorithm for Mobile Computing," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, June 2012, pp. 1991–95.
[12] E. Z. Tragos, G. Tsiropoulos, G. T. Karetsos, and S. A. Kyriazakos, "Admission Control for QoS Support in Heterogeneous 4G Wireless Networks," *IEEE Network*, May/June 2008, pp. 30–37.
[13] L. B. Le *et al.*, "Joint Load Balancing and Admission Control in OFDMA-Based Femtocell Networks," *IEEE ICC 2012*, 2012, pp. 5135–39.
[14] S. Corroy, L. Falconetti, and R. Mathar, "Cell Association in Small Heterogeneous Networks: Downlink Sum Rate and Min Rate Maximization," *IEEE WCNC 2012*, 2012, pp. 898–902.
[15] Y. Zhang *et al.*, "Performance Analysis of user Association Policies in Small Cell Networks using Stochastic Petri Nets," *IEEE ICC Wksp. 2013*.
[16] A. Simonsson, and A. Furuskar, "Uplink Power Control in LTE C Overview and Performance," *IEEE 68th Vehic. Tech. Conf., VTC 2008-Fall*, Sept. 2008, pp. 1–5.
[17] Y. Wang, and S. Venkatraman, "Uplink Power Control for Heterogeneous Networks in LTE," *2012 IEEE Globecom Workshops (GC Wksps.)*, Dec. 2012, pp. 592–97.
[18] M. Andrews, "A Survey of Scheduling Theory in Wireless Data Networks," *Wireless Communications (The IMA Volumes in Mathematics and Its Applications)*, vol. 143, 2007, pp. 1–17.
[19] J. Lee and N. Jindal, "Energy-Efficient Scheduling of Delay Constrained Traffic over Fading Channels," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1866–75, Apr. 2009.
[20] M. J. Neely, "Optimal Energy and Delay Tradeoffs for Multi-User Wireless Downlinks," *IEEE Trans. Info. Theory*, vol. 53, no. 9, Sept. 2007, pp. 1–17.
[21] S.-B. Lee *et al.*, "Proportional Fair Frequency-Domain Packet Scheduling for 3GPP LTE Uplink," *IEEE Infocom 2009*, pp. 2611–15.

BIOGRAPHIES

LEI LEI received a B.S. degree in 2001 and a Ph.D. degree in 2006, respectively, from Beijing University of Posts & Telecommunications, China, both in telecommunications engineering. From July 2006 to March 2008, she was a postdoctoral fellow at Computer Science Department, Tsinghua University, Beijing, China. She worked for the Wireless Communications Department, China Mobile Research Institute from April 2008 to August 2011. She has been an Associate Professor with the School of Electronic and Information Engineering, Beijing Jiaotong University since September, 2011. Her current research interests include performance evaluation, quality-of-service and radio resource management in wireless communication networks.

KAN ZHENG [SM'09] (kzheng@ieee.org) received the B.S., M.S. and Ph.D. degree from Beijing University of Posts & Telecommunications (BUPT), China, in 1996, 2000 and 2005, respectively, where he is currently associate professor. He worked as a researcher in the companies including Siemens, Orange Labs R & D (Beijing), China. His current research interests lie in the field of wireless communications, with emphasis on heterogeneous networks and M2M networks.

JIADI CHEN received her B.S. degree from Chongqing University & Posts and Telecommunications (CUPT), China, in 2010. She is currently a candidate for Ph.D. in the Key Lab of Universal Wireless Communications, Ministry of Education, BUPT. Her research interests include performance analysis of wireless networks, resource allocation and scheduling algorithm.

HANLI MENG received her B.S. degree from Beijing University of Posts & Telecommunications (BUPT), China, in 2013. She is currently a candidate for M.S in the Key Lab of Universal Wireless Communications, Ministry of Education, BUPT. Her research interests include resource allocation and scheduling algorithm in wireless networks.

ZHANGDUI ZHONG received the B.Eng. and M.Sc. degrees from Northern Jiaotong University (currently Beijing Jiaotong University), Beijing, China, in 1983 and 1988, respectively. He has been a Professor with the School of Electronic and Information Engineering, Beijing Jiaotong University, since 2000. He has authored seven books and over 150 technical papers in the field of wireless communication. His current research interests include wireless communication theory for railway systems, wireless ad hoc networks, channel modeling, radio resource management, intelligent transportation systems, and GSM-R systems.