Wireless Pers Commun (2011) 60:125–143
DOI 10.1007/s11277-011-0252-4

Rate-Control and Queueing of Backhaul Downstream Traffic for Mobile Wireless Systems

Young-June Choi · Kyungtae Kim · Kang G. Shin

Published online: 3 March 2011

Abstract A backhaul network in mobile wireless systems consists of lower-level base stations (BSs) and upper-level access routers (ARs). While the legacy model considers only a queue at a BS for downstream traffic, we focus on queues at both the BS and the AR, hence calling it the *split two-level queueing (S-2Q) model*. The transmission rate of the backhaul link between a BS and an AR can be adjusted to stablize the BS queue. We develop a queue-aware rate control algorithm for the backhaul link such that BS queues will suffer neither buffer overflow nor underflow due to drastic short-term variations in wireless channel condition. We then derive the stability conditions of BS queues and propose two strategies, each applicable to handoff and normal (non-handoff) users separately. For handoff users, it is desirable that the BS queue buffers as few packets as possible to improve handoff performance. For normal users, it is desirable that the BS queue buffers as many packets as possible to exploit multiuser diversity of opportunistic scheduling. Our in-depth simulation results have shown that the proposed algorithm stabilizes the BS queue for normal users and reduces handoff latency to 0.8 second from 3 seconds for handoff users.

Keywords Backhaul networks · Wireless mobile networks · Queueing · Rate control

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2010-(C1090-1021-0011)).

Y.-J. Choi (✉)
Ajou University, Woncheon-dong, Yeongtong-gu, Suwon 443-749, Korea
e-mail: choiyj@ajou.ac.kr

K. Kim
NEC Laboratories America, 4 Independence way, Princeton, NJ 08540, USA
e-mail: kyungtae@nec-labs.com

K. G. Shin
The University of Michigan, 2260 Hayward ave, Ann Arbor, MI 48109, USA
e-mail: kgshin@eecs.umich.edu

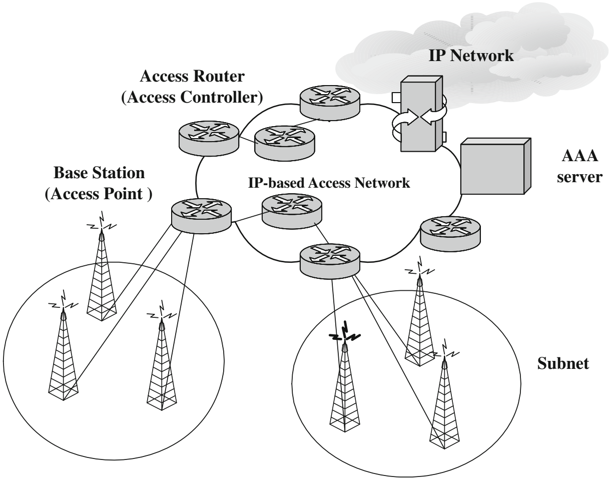


Fig. 1 A backhaul access network considered for mobile wireless systems

1 Introduction

For efficient radio resource and mobility management, mobile wireless networks have employed a hierarchical backhaul structure. In 2G or 3G cellular systems [1], a set of Base-Transceiver Stations (BTSs) are managed by a Base-Station Controller (BSC). In real applications, a BTS plays the limited role of physical transmitter and receiver with some simple functions such as fast power control, while a BSC manages most of the radio resources. In 3G-LTE (Long-Term Evolution) systems that consist of base stations (BSs) and Radio-Network Controllers (RNCs) where a BS manages more radio resources. Recently, the WiMAX forum has defined an Access Service Network (ASN) which is the backhaul network that connects multiple BSs to an ASN gateway [2]. Mobile terminals (MTs) are thus connected to a BS wirelessly, while the BS is connected to a wired backhaul network. Throughout this paper, we will use the term "access router" (AR) to represent the upper entity—such as the RNC in 3G systems or the ASN gateway in WiMAX systems—that controls multiple BSs underneath.

In mobile wireless networks as shown in Fig. 1, downstream packets will be buffered at an AR and then at a BS before they are delivered to MTs, which we call a *split two-level queueing* (S-2Q) model. When downstream packets are transmitted, wireless links will become the bottleneck, since backbone networks and wired links will have significantly more bandwidth than wireless links. Since the data rate on a wired backhaul-link between an AR and a BS is usually much higher than that on a wireless-link between the BS and an MT, the downstream traffic is likely to be buffered at BSs.

We therefore propose to control the data rate of backhaul-links so as to protect BS queues from overflow or underflow. In our proposed approach, ARs will primarily buffer the downstream traffic and deliver the buffered traffic to BS queues based on feedback from each BS. A BS periodically measures its queue length and sends feedback to the corresponding AR in order to maintain a moderate queue length. We design an easy-to-implement algorithm and analyze the conditions of queue length in order to achieve the stability of a BS queue.

Further, by adapting the S-2Q model, we propose two strategies for handoff users/MTs and normal users (i.e., who are not involved with handoffs). For handoff users, it is desirable that the BS queue buffers as few packets as possible to improve handoff performance. For normal users, it is desirable that the lower queue buffers as many packets as possible without incurring buffer overflow. This is because channel-aware scheduling, also called *opportunistic scheduling*, can exploit multiuser diversity when a BS queue holds many packets from different users [3,4]. The proposed mechanisms are applied and evaluated for IEEE 802.16 systems [5].

To best of our knowledge, this is the first attempt to address the backhaul link control for two-level queueing of downlink traffic in mobile wireless networks. In [6,7], radio resource allocation has been investigated by exploiting the hierarchical backhaul structure. In recent IP-based access networks, as the bandwidth of backhaul networks is related to deployment cost, sizing backhaul links has been addressed in [8,9]. There has also been extensive research on the topology design of backhaul access networks (see [10] and references therein), and mesh backhaul networks (e.g., [11,12]). In this paper, we do not deal with topological problems such as tree-based or mesh networks, but use a simple topology where each BS is logically connected to an AR. Our approach can be extended to such multihop-based backhaul networks, but such an extension is part of our future work.

The remainder of this paper is organized as follows. In Sect. 2, we discuss the advantages of the S-2Q model over the existing legacy model. In Sect. 3, we present a rate-control algorithm for backhaul links and derive the conditions of its stable operation. In Sect. 4, user-adaptive application of the algorithm is described for two user types, handoff and normal user types, followed by our proposed user-split network architecture. In Sect. 5, the proposed algorithm is simulated to demonstrate its ability of handling the two user types. Finally, the paper concludes with Sect. 6.

2 Split Two-Level Queueing

Although in a backhaul network of cellular systems, both BSs and an AR can buffer downstream packets, only queueing at each BS is considered while ignoring the queueing at the AR. That is, packets are assumed to be buffered only at BS queues; we call it the *legacy model*. In this legacy model, packet drops due to buffer overflow will occur only at BS queues, while in the S-2Q model, the packet drops can be controlled by the upper queue of an AR, because the AR can deal with the entire traffic within the subnet below itself. Since the AR is usually equipped with a large buffer that would otherwise be given to its subordinate BSs, the flexibility in managing a given buffer is enhanced, thus reducing the probability of buffer overflow even if bursty data traffic is destined for one particular BS. Usually, data traffic often suffers buffer overflow due to its burstiness before reaching the last-hop wireless links.

The burstiness of data traffic has long been observed, e.g., self-similarity of Internet traffic [13]. Figure 2 shows a typical traffic pattern of CBR voice traffic and bursty data traffic. While voice traffic consumes bandwidth with small variations, data traffic does with large variations. Sometimes, the rate of arriving data traffic exceeds the medium transmission rate, so a buffer overflow becomes inevitable in the legacy model.

Figure 3 shows the behavior of the legacy and proposed models. In the legacy model, most packets are buffered at BS queues, thus sometimes being dropped at BS queues, while in the proposed S-2Q model, most traffic is controlled by an AR queue, and hence, a BS may suffer few queue drops. In the latter case, a BS may encounter a queue overflow as a result of the AR's failure to control the backhaul-link data rate. On the other hand, a BS's queue may also become empty (i.e., a queue underflow) even when the AR buffers some traffic destined for this AP.

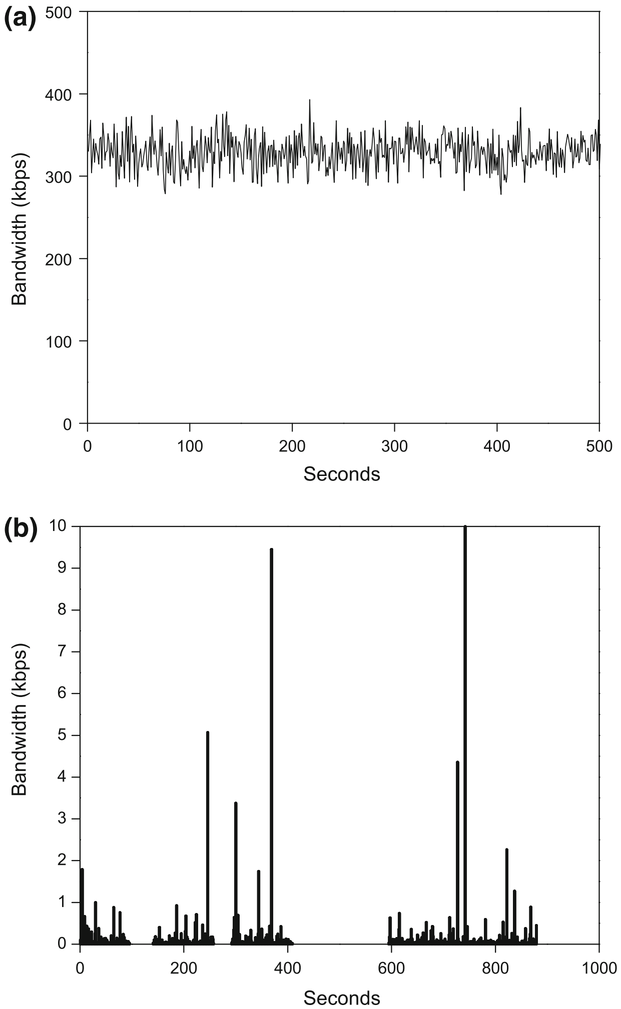


Fig. 2 Comparison of traffic burstiness: voice traffic (*upper*) and data traffic (*lower*)

Let us first consider a buffer overflow in the legacy model. When a burst of data packets arrive at an access network, a buffer overflow may occur at a BS. Suppose there are N BSs under an AR and let L^+ be the maximum queue size. Then, the probability of no buffer overflow is $Pr\{q_i^{BS} \leq L^+\}$, where q_i^{BS} is the queue size of BS$_i$. Equivalently, we consider the S-2Q model where an AR has a buffer of size $N \cdot L'$ and each BS has a buffer of size L, when $L + L' = L^+$. Then, the total buffer space of an AR and BSs, $N \cdot (L + L')$, is the same as $N \cdot L^+$, the sum of BSs' buffers in the legacy model. The probability of no buffer overflow is given as $Pr\{\sum_{i=1}^{N} q_i^{AR} \leq N \cdot L',\ q_i^{BS} \leq L\}$, where q_i^{AR} is the queue size for BS$_i$'s traffic in the AR.

Using these probabilities, one can derive the following condition:

$$Pr\left\{\sum_{i=1}^{N} q_i^{AR} \leq N \cdot L', q_i^{BS} \leq L\right\} \tag{1}$$

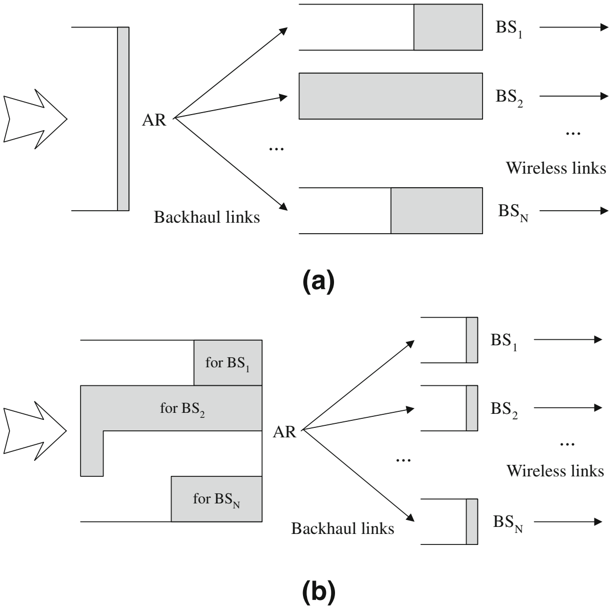


Fig. 3 Two simple models of backhaul queueing: **a** the legacy model, **b** the proposed S-2Q model

$$> Pr\left\{q_1^{AR} + q_1^{BS} \le L + L', \ldots, q_N^{AR} + q_N^{BS} \le L + L'\right\} \tag{2}$$

$$= Pr\left\{q_1^{BS} \le L^+\right\} \cdots Pr\left\{q_N^{BS} \le L^+\right\}. \tag{3}$$

In the legacy model, Eq. (2) is equal to Eq. (3) as $q_i^{AR} = 0$. Equation (3) represents the probability of no buffer overflow in the legacy model, which is clearly smaller than that in the S-2Q model represented by Eq. (1). One can also intuitively see this relationship because keeping a buffer in the AR increases the flexibility (i.e., degrees of freedom) for buffer management. In the legacy model, a buffer overflow occurs whenever excess traffic is generated toward at least one BS, while in the S-2Q model, it does not always occur.

This advantage in the S-2Q model is achieved with stability, when the backhaul-link rate is controlled properly. An algorithm for controlling the backhaul link and the condition of queue stability will be derived in the following sections.

3 Rate-Control Algorithm

3.1 Motivation

For cost-efficient implementation of the S-2Q model, packets after the BS's buffer is filled up are buffered at an AR queue. When bursty traffic is delivered to a BS, the BS may encounter a queue overflow as a result of the AR's failure to control the backhaul-link rate. On the other hand, a BS's queue may also become empty (i.e., queue underflow) even when the AR buffers too much of traffic destined for this BS.[1] For the proper operation of the S-2Q model, an AR

[1] This situation is equivalent to a non-work-conserving server, i.e., a server is idle even when packets are available for transmission. Note that a queue underflow can occur when there is no traffic to send at the AR

should control the backhaul-link rate such that its subordinate BSs will suffer neither queue underflow nor overflow. If a wireless-link rate remains constant and is also predictable, an AR can send traffic to BSs at a constant rate. But wireless links are usually unreliable and channel conditions are unpredictable, so an AR should adapt the backhaul-link rate to the instantaneous wireless-link rate.

When the overall channel condition becomes good, the wireless-link rate will increase, so the backhaul-link rate should be increased. In contrast, when the overall channel condition becomes poor and transmission errors occur, the wireless-link rate will decrease, so the backhaul-link rate should be reduced. To overcome the difficulty in predicting future wireless channel conditions, we control the backhaul-link rate based on the measurement of queue length.

Hence, we devise an adaptive queue-aware rate-control algorithm that adjusts the backhaul-link rate from an AR queue to BS queues, based on the measurement of queue length.[2] For simplicity, we assume that an AR has sufficient queue space and each of its subordinate BSs has a queue of maximum length L. Let the backhaul-link rate and wireless-link rate be λ_i and r_i, respectively, when the link is connected to BS$_i$.

3.2 Algorithm

If the queue size of a BS is too small, even non-work-conserving transmission may be possible, i.e., a BS queue might experience underflow despite the fact that its AR buffers traffic for the BS. To avoid this situation, we define a threshold $q_{\min}$. If the current queue length is $<q_{\min}$, the BS requests its AR to increase the transmission rate by α. On the other hand, to avoid buffer overflow, we define another threshold $q_{\max}$. When the BS's queue length is $>q_{\max}$, the BS requests its AR to decrease the transmission rate by $1/\beta$. We use $\alpha = \beta = 2$ which is a reasonable choice, because various data rates in a cell are often increased or decreased by a binary exponent, given the nature of modulation and coding.

Based on the measurement of queue length, each BS provides feedback to its AR. A BS need not report this information if $q_{\min} \leq q_i \leq q_{\max}$, but must report it if $q_i < q_{\min}$ or $q_i > q_{\max}$. Hence, we can design two types of signaling messages from a BS to its AR: *req_underflow_preventing* and *req_overflow_preventing*. This way, a minimal size of a feedback message is delivered, even without containing such information as queue length or average wireless link rate, so the overhead of implementation is kept low. Upon receiving data from the AR, the BS inspects its current queue length and sends the feedback again if $q_i < q_{\min}$ or $q_i > q_{\max}$. Otherwise (i.e., if there is no feedback), the AR increases or decreases λ_i by multiplying it by θ. Two strategies of setting θ will be addressed in the next section. The AR will thus adjust the transmission rate according to our rate-control algorithm as follows.

```
BS operation: after receiving data
if $q_i < q_{\min}$
    send a req_underflow_preventing
else if $q_i > q_{\max}$
```

Foonote 1 continued
queue as well as the BS queue. Throughout this paper, a buffer underflow means the non-work-conserving underflow.

[2] We consider the aggregated queue length of all flows, not per-flow queue length, not only because the per-flow queue length incurs much more overhead for exchanging messages in designing our rate-control algorithm, but also because fair queueing can also be achieved by scheduling or shaping traffic.

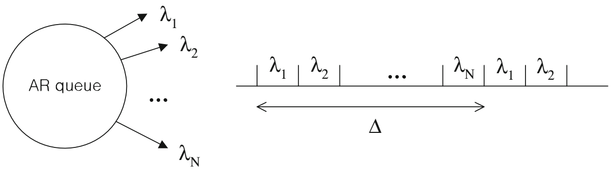

AR queue
λ_1
λ_2
...
λ_N
λ_1 | λ_2 | ... | λ_N | λ_1 | λ_2
Δ

Fig. 4 Scheduling over backhaul links at an AR

```
    send a *req_overflow_preventing*
end if
```

<u>AR operation: before sending data</u>

```
if receive a *req_underflow_preventing*
    if $\lambda_i = 0$
        $\lambda_i \leftarrow r_{\min}$
    else
        $\lambda_i \leftarrow \min(2\lambda_i, r_{\max})$
    end if
else if receive a *req_overflow_preventing*
    if $\lambda_i = r_{\min}$
        $\lambda_i \leftarrow 0$
    else
        $\lambda_i \leftarrow \max(\lambda_i/2, r_{\min})$
    end if
else
    $\lambda_i \leftarrow \min(\lambda_i * \theta, r_{\max})$
end if
```

In the normal operation of adaptive modulation and coding, there are a minimum data rate $r_{\min}$ and a maximum data rate $r_{\max}$, supported in a system, e.g., $r_{\min} = 38.4$ Kbps and $r_{\max} = 2457.8$ Kbps in cdma2000 1x EV-DO [14]. Therefore, we limit λ_i by $r_{\min} \leq \lambda_i \leq r_{\max}$. We also set $\lambda_i = 0$ for the worst case when all the MTs in a cell cannot receive data for a while because of their bad channel conditions. Thus, λ_i becomes 0 upon receiving a *req_overflow_preventing* message when λ_i was $r_{\min}$, and λ_i becomes $r_{\min}$ upon receiving a *req_underflow_preventing* message when λ_i was 0.

3.3 Stability Conditions

In practice, packets are periodically transferred from an AR queue to its subordinate BS queues, as an AR runs a scheduling algorithm for many BSs as shown in Fig. 4. The packet inter-arrival time can vary according to the AR's queueing and scheduling policies. Suppose that there is a maximum interval Δ when a specific scheduling policy is used at an AR. The reason for considering the maximum interval is to obtain the condition of $q_{\min}$ and $q_{\max}$ in the worst case when a BS's queue overflow or underflow can occur. For analytical simplicity, we only use a fixed Δ.

At an arbitrary time t_o, let $\lambda_i(t_o)$ be the average rate of downstream traffic arrival at a BS. Then, a certain amount $\lambda_i(t_o) \cdot \Delta$ of traffic is newly buffered at the queue. Let $r_i(t_o)$ be the average wireless-link rate during $[t_o, t_o + \Delta)$, then an amount $r_i(t_o) \cdot \Delta$ of traffic is delivered over the wireless downlink. After this time interval, the queue length will decrease if $\lambda_i \le r_i$. Otherwise, it will increase. We will henceforth omit the BS index i for notational simplicity.

A problem in designing our algorithm is how to set $q_{\min}$ and $q_{\max}$ such that the BS queue will suffer neither underflow nor overflow. Our solution to this problem is given in the next propositions, proved in Appendices 1 and 2.

Proposition 1 *The worst-case condition of $q_{\min}$ to avoid queue underflow (i.e., non-work-conserving) is $q_{\min} \ge \left(\lfloor \log_2 \frac{r_{\max}}{r_{\min}} \rfloor \cdot r_{\max} + r_{\min}\right) \cdot \Delta$.*

Proposition 2 *The worst-case condition of $q_{\max}$ to avoid queue overflow is $q_{\max} \le L - (r_{\max} - r_{\min}) \cdot \Delta$.*

Since $q_{\max}$ should be greater than $q_{\min}$, we derive a condition on L as:

Proposition 3 $L > \left(\lfloor \log_2 \frac{r_{\max}}{r_{\min}} \rfloor + 1\right) \cdot r_{\max} \cdot \Delta$.

The above propositions ensure our rate-control algorithm to achieve stability in queue length.

4 User-Adaptive Application of S-2Q Model

The S-2Q model can be applied to handoff users and normal users in different ways. We address the issues of both user types and propose a user-adaptive S-2Q model based on user classification.

4.1 Handoff Users

When multiple BSs are subordinate to an AR, establishing a subnet, there are two types of handoff operations: (i) handoffs within the same subnet (intra-subnet handoffs) and (ii) handoffs between different subnets (inter-subnet handoffs). MTs need not change a layer-3 connection as long as they move around within one subnet, but need to change a layer-2 connection. Thus, one solution to mobility management is performing only layer-2 handoffs within a subnet for intra-subnet handoffs and layer-3 operations for inter-subnet handoffs. Here we focus only on intra-subnet handoffs, because inter-subnet handoffs requires solutions to various implementation issues in layer 3 including mobile IP. Details on IP mobility can be found from [16].

To prevent packet loss during a handoff, "packet buffering-and-forwarding" method has been considered [17–19]. In the legacy model, some packets can remain after the receiver MT has already moved to another cell. The packets that arrived at the old BS will be forwarded to the new BS to avoid packet loss; otherwise, the packets will be dropped. Although some packets destined for a specific MT were delivered to the correct BS when they passed through the AR, the MT may have already moved to another cell while the packets are being buffered at the old BS. Even if the packets are forwarded to the new cell the MT has already moved to, the queueing delay lengthens the handoff latency. Moreover, these packets will be delivered out of order, and hence, MTs suffer performance degradation.

This problem can be alleviated in the S-2Q model, when an AR buffers most of packets and BSs buffer small number of packets. The queueing delay occurs at an AR instead of a BS, but the AR can update the MT's location information before delivering most downstream traffic to BSs. Whenever an MT moves within a subnet, most traffic destined for the MT need not be forwarded to a new BS for a handoff, because the AR can transmit the traffic to the correct BS directly.

The S-2Q model also facilitates other solutions for smooth handoffs. For example, if packets are multicast to both the old and new BSs, an MT can transmit or receive to/from both BSs at the same time. This mechanism is applicable to the S-2Q model, rather than the legacy model, because an AR can transmit packets to multiple BSs. This mechanism has been introduced as a *macro-diversity handoff procedure* in IEEE 802.16e systems [15]. Although it consumes twice more bandwidth at both wireless and wired links, handoff users will experience a smooth handoff.

In our evaluation, we consider the simplest solution that the packets buffered at the old BS are dropped during the handoff, which is the default operation in IEEE 802.16 systems. As in the case of packet buffering-and-forwarding, the old BS will not observe significant packet drops if most packets are buffered at an AR queue in the S-2Q model.

These advantages are also gained in inter-subnet handoffs, if the packet are buffered at ARs. Then, the packets buffered at an old AR can be forwarded to a new AR after making the layer-3 change, and there are no dropped or forwarded packets at the old BS. In the case of legacy model, however, the packets were probably sent to the old BS before making a layer-3 handoff. The packets will be dropped or re-routed by the AR to a new subnet, which lengthens the handoff latency.

4.2 Normal Users

User mobility does not only cause handoffs at the network level, but also causes wireless channels to vary rapidly at the system level. Basically, wireless channels are attenuated monotonically with the distance between a transmitter and a receiver (i.e., path loss). When MTs move around, multipath fading, also known as *fast fading*, makes the wireless channel fluctuate in addition to the dynamically-changing channel condition due to path loss and shadowing.

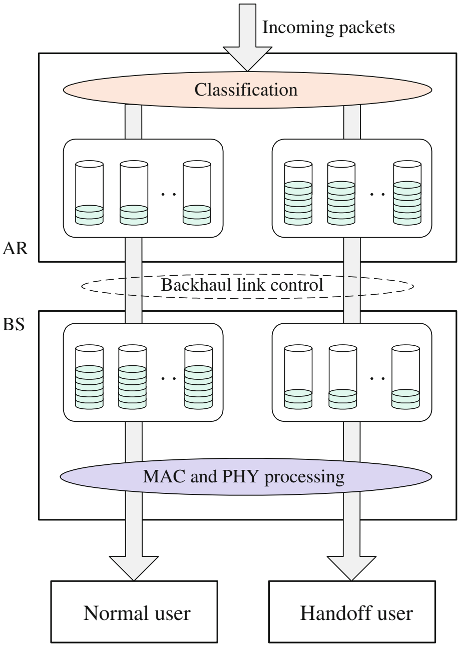
To exploit the varying channel conditions, a BS can allocate its wireless channel to the MT that has the best channel quality in a cell. Owing to adaptive modulation and coding, wireless networks can support various data rates according to the reported channel quality. Hence, this channel-aware scheduling, also called *opportunistic scheduling*, has been widely used (e.g., [3,4]). Exploitation of this nature of wireless channels increases cell throughput, resulting in *multiuser diversity*.

However, multiuser diversity over wireless links may not be fully utilized, if most packets are buffered in ARs and there is not too much of traffic from different users in a BS queue. This will hinder the wireless link scheduler at a BS from taking advantage of multiuser diversity. It is well-known that the average data rate over a wireless link increases with the number of users whose downlink traffic is buffered at a BS [4].

4.3 Framework of User-Adaptive S-2Q

The backhaul-link rate control algorithm can be applied to handoff users and normal users in different ways. To achieve a small queue length at BSs especially for handoff users, we must decrease the BS's queue length as long as the queue doesn't experience underflow. So, in our algorithm, before sending data to BS$_i$, an AR decreases λ_i by a constant ratio θ, i.e., $0 < \theta < 1$ unless it received other signaling messages from BS$_i$.

Fig. 5 The user-adaptive S-2Q model



On the other hand, for normal users, a BS queue should keep a sufficiently large queue for multiuser diversity without suffering any overflow. To achieve multiuser diversity, it will be beneficial to increase the BS's queue length before the queue experiences overflow. Thus, the condition of θ should be $\theta > 1$.

To manage handoff users and normal users separately, we also design the framework of user-adaptive S-2Q as shown in Fig. 5. The handoff user type is identified via such information as signal strength or signal-to-interference-and-noise ratio (SINR) from neighboring BSs, when each MT is capable of measuring SINR from its own BS as well as a neighboring BS.

4.4 Implementation Issues

When MTs are located in cell-edge areas, they can also be managed separately in order to solve the well-known intercell interference problem and enhance cell-edge performance. Those users are likely to be supported by dynamic frequency reuse [7] or macro-diversity [15,20] that does not exploit opportunistic scheduling, so the handoff user type can subsume such cell-edge users as well as handoff users.

A queueing strategy for the upper queue at an AR may affect the overall performance significantly, especially for handoff users whose packets are mostly buffered at an AR. The upper queue may employ a rate shaping technique, thus reducing the burstiness of TCP traffic. The shaper at the upper queue can transmit the same amount of traffic for every flow that is destined for a BS, so the fairness of TCP flows can be improved. This traffic shaping reduces the complexity in designing a scheduler for BSs, because the BS scheduler may not consider the queue length together with fairness which is already considered by the upper queue. The improved performance of TCP fairness is observed by handoff users in our evaluation results.

We evaluate our mechanisms in IEEE 802.16 systems [5] that are regarded as a starting point of 4G networks in the WiMAX group. The WiMAX standard has defined three different profiles for an Access Service Network (ASN) which is the backhaul network that connects multiple BSs to an ASN gateway [2]. In Profile A, which is now defunct, most radio resources are managed by the ASN gateway as in traditional cellular networks and thus has a hierarchical structure. In Profile B, the functionalities of a BS and an ASN gateway are co-located on the same platform/solution, which makes the architecture flat. Profile C also defines a hierarchical architecture where all the radio resource management functions are performed at the BS but BSs are controlled by an ASN gateway. As Profile C is popular, our mechanism between an AR and its subordinate BSs can be applicable to an ASN gateway and its subordinate BSs. Similarly, it is applicable to a set of a RNC and subordinate e-Node Bs in LTE and LTE-Advanced systems. Especially, femto BSs are equipped with limited capability, so our approach will be useful for the hierarchical management between a macrocell and several femtocells. Our mechanism can also be applied to existing mobile wireless networks such as cdma2000 EV-DO and WCDMA-HSDPA in evolution towards all-IP B3G (beyond 3G) systems.

5 Evaluation Result

Using *ns*-2-based simulation, we evaluated the performance of the proposed S-2Q model in an IEEE 802.16 OFDM/TDD (time division duplexing) environment with a 5 Mhz channel [21]. MTs download FTP traffic from a server, and the propagation delay between the FTP server and the AR is set to 30 m s. As the backhaul link has sufficient bandwidth, its propagation delay is assumed to be 1 m s. The simulation time is 300 s. From the measured wireless-link rate, we set $r_{\min}$ and $r_{\max}$ to 1 Mbps and 15 Mbps, respectively. Δ is set to 12 m s. The queue length represents the number of packets in a queue, where packets are queued each in a unit of 1500 bytes. In both models, we set the maximum queue lengths of a BS and an AR at L and $10L$, respectively, but the backhaul-link rate control algorithm was not applied to the legacy model. L is set to 100 by default, since L should be greater than 60 in our setting according to Proposition 3. From Propositions 1 and 2, we use $q_{\min} = 46$ and $q_{\max} = 85$. From our extensive simulation, the desirable operating region of θ is found to lie between 0.7 and 0.9 for the handoff user type and between 1.1 and 1.3 for the normal user type. Since its effect on performance is insignificant, we do not show it here, and we only set $\theta = 0.8$ or 1.2.

Figure 6 shows a sample of the number of users whose downlink traffic is buffered at a BS, when 8 MTs in a cell are FTPing files and the backhaul-link rate control algorithm is used with $\theta = 0.8$ or $\theta = 1.2$. As expected, when θ is 0.8, the number of such users is kept very small. In contrast, the number is increased sufficiently when $\theta = 1.2$ for multiuser diversity.

Now, we show the performance of normal users in the S-2Q model in comparison with the legacy case where there is no buffering at the AR. To consider a situation of the overloaded cell, four of ten BSs have 15 FTP connections while the others have 5 FTP connections, when the AR has 10 subordinate BSs. Figure 7 plots the packet-drop rate in the entire subnet as a function of L. As argued in Sect. 2, the BS queue in the legacy case suffers from significant packet drops (mostly at the overloaded BSs that support 15 FTP connections), especially when the BS's queue is small, but this problem does not appear in the S-2Q model. This result confirms that the S-2Q model is better in managing the total buffer space in a subnet, and the legacy model, on the other hand, suffers packet losses when a BS is overloaded. When $L = 100$, we compare the change of queue length at a BS that is supporting 15 FTP connections in Fig. 8. The queue length often hits the ceiling owing to a queue overflow in the legacy model, but the queue length stays moderate and stable in the S-2Q model.

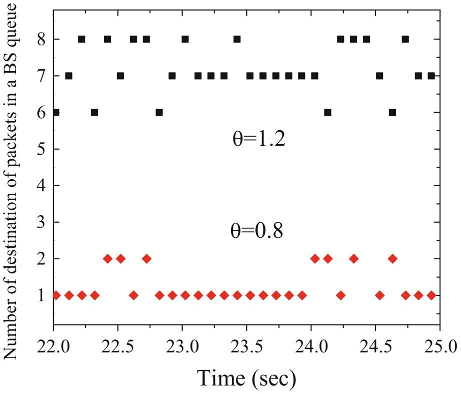


Fig. 6 A sample of the number of users whose downlink traffic is buffered at a BS

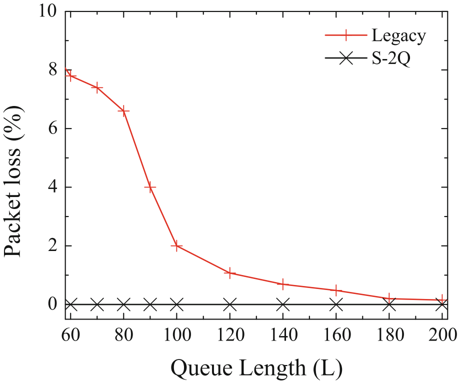


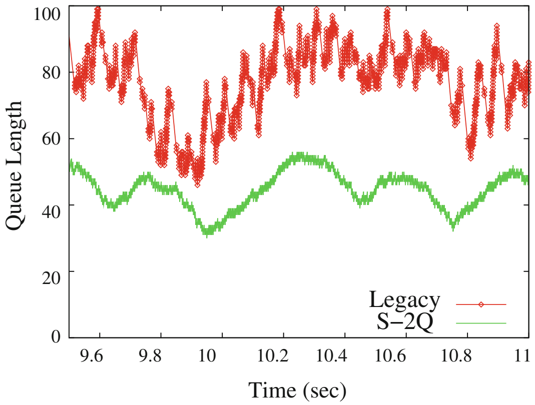
Fig. 7 Comparison of packet drops in the legacy and S-2Q strategies

We evaluated the operation of the S-2Q model with $q_{\min}$ and $q_{\max}$ that do not satisfy the conditions in Propositions 1 and 2. Figure 9 shows the change of queue length at a BS (supporting 15 FTP connections) for three combinations of $q_{\min}$ and $q_{\max}$. The condition of the wireless channel has been poor for 0.3 s starting from the 10 s point, so the queue length begins to increase at the 10 s point and then decreases. When $q_{\min} = 26$, far less than the given condition of $q_{\min} \geq 46$, a queue underflow occurs. On the other hand, when $q_{\max} = 105$ (it actually exceeds the queue limit of 100) that is far greater than the given condition of $q_{\max} \leq 86$, a queue overflow occurs. In summary, the S-2Q model works well with our rate-control algorithm under the given conditions.

We further investigate the performance of handoff users, together with the impact on TCP performance in the proposed S-2Q model. Eight MTs FTP files from a BS, and one of them represented by "TCP flow 8" is moving to a neighboring BS. To investigate the effect of TCP, we make all flows follow the same handoff user type with $\theta = 0.8$. In the S-2Q model, the AR queue employs an additional traffic shaper on top of simple FIFO scheduling.

Figure 10 compares TCP throughput in the legacy model and the S-2Q model. In the legacy model case, TCP flow 8 experiences throughput degradation for more than 3 seconds during a handoff, while in the S-2Q model case, the degradation is much less. Here, flow 8 achieves higher throughput after moving to the new cell because the cell is lightly-loaded. The result also reveals throughput unfairness among TCP flows in the legacy model. In the S-2Q model, however, TCP flows share bandwidth resource fairly, because the upper (AR)

Fig. 8 Comparison of queue length in the legacy and S-2Q models



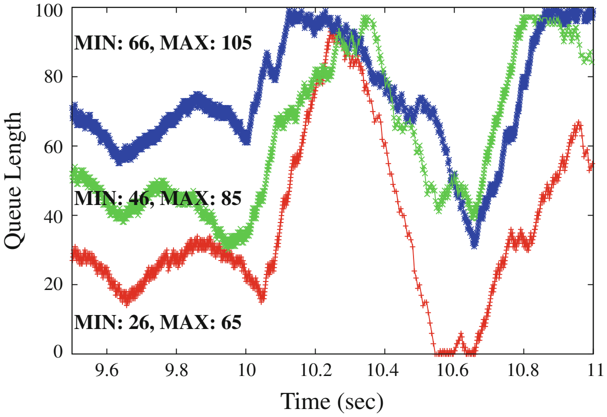


Fig. 9 Effects of different combinations of $q_{\min}$ and $q_{\max}$ on queue length in the S-2Q model

queue plays a role of TCP traffic shaper such that the amount of delivered traffic will be the same for every flow.

Figure 11 further compares the TCP sequence numbers of flow 8 during the handoff. In both models, handoff latencies are 3.0 and 0.8 seconds, respectively. As stated earlier, IEEE 802.16 systems do not support packet buffering-and-forwarding at BSs, thereby dropping the packets buffered at the old BS during the handoff. Thus, the MT in the legacy model experiences a longer handoff latency and more packet losses. In contrast, the packets in the S-2Q model are mostly buffered at the AR, so the MT experiences a shorter handoff latency with few packet losses.

In the above experiment, the TCP congestion window is also plotted in Fig. 12. During the handoff, flow 8 experiences a burst of packet drops that cause its TCP congestion window to drop to 1. In the legacy model, more than 10 packets were dropped, while in the S-2Q model, only one packet was usually dropped, because one TCP ACK packet is lost in the uplink. The loss of one ACK packet occurs as the corresponding BS hands off the connection to a new BS before receiving the final TCP ACK.

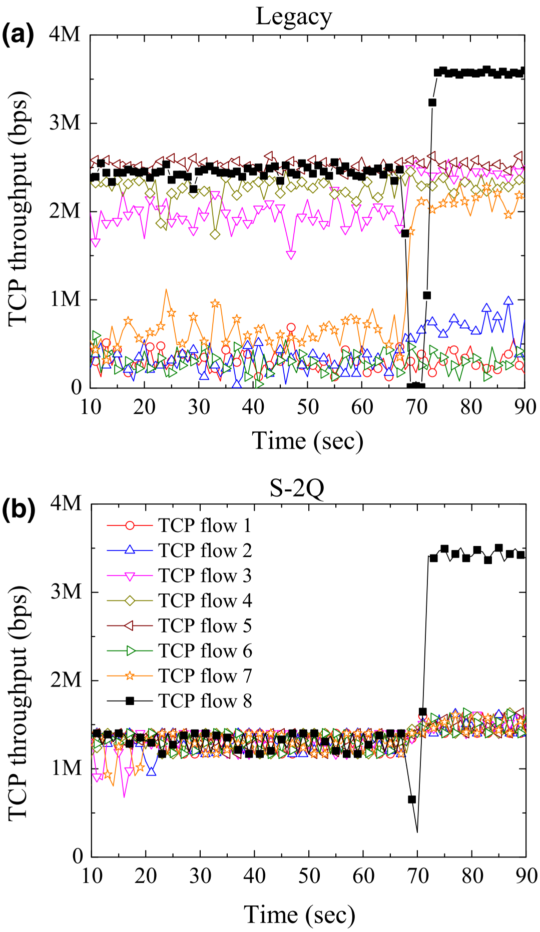


Fig. 10 TCP throughput during a handoff

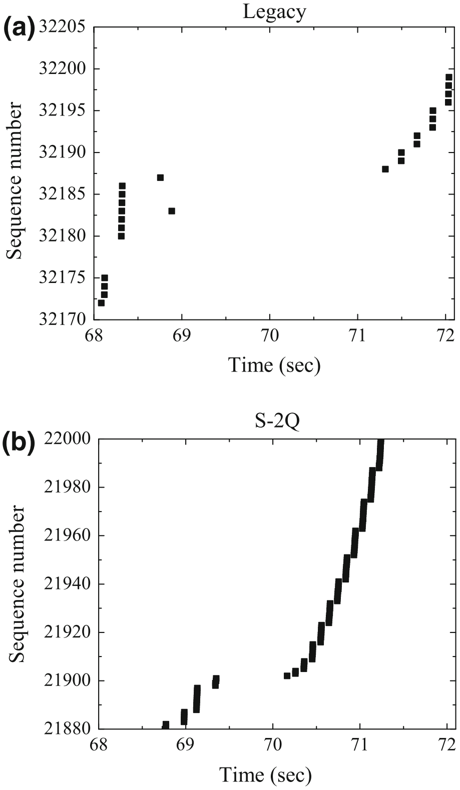
6 Conclusion

In this paper, we addressed the problem of buffering downstream traffic at ARs and BSs in the backhaul access networks for mobile wireless communications. ARs buffer packets and forward them to BSs with the proposed rate-control algorithm in the S-2Q model such that the BS queues will achieve stability even in the presence of wireless channels with fluctuating conditions. We also devised a framework for managing handoff users separately from normal users in order to apply different objectives of the BS queues. The separate management for handoff users will be a promising solution in next-generation wireless systems to handle the potential problems of cell-edge users who are exposed to inter-cell interference.

Appendix 1: Proof of Proposition 1

Assume that a *req_underflow_preventing* message was generated in $[t_o, t_o + \Delta)$. To avoid buffer underflow, $q_{\min}$ must be greater than $(r(t_o) - \lambda(t_o)) \cdot \Delta$ in case of $r(t_o) > \lambda(t_o)$ when the queue length keeps decreasing. Although λ increases exponentially below the queue length

Fig. 11 TCP sequence number during a handoff



of $q_{\min}$, it is possible that $r(t_o + m \cdot \Delta)$ is still greater than $\lambda(t_o + m \cdot \Delta)$ for an integer m. Thus, we consider the worst case of $r(t_o) = r(t_o + \Delta) = \cdots = r(t_o + m \cdot \Delta) = r_{\max}$ and $\lambda(t_o) = 0$, when $q(t) < q_{\min}$ for $t_o \le t < t_o + (m+1) \cdot \Delta$ and the queue length keeps decreasing.

The range of m is given when λ is less than r. According to our proposed algorithm, $\lambda(t_o + \Delta) = r_{\min}$, $\lambda(t_o + 2\Delta) = 2r_{\min}$, and thus, $\lambda(t_o + m \cdot \Delta) = 2^{m-1} r_{\min}$. We find the greatest m that satisfies $\lambda(t_o + m \cdot \Delta) < r_{\max}$, and yields $2^{m-1} r_{\min} < r_{\max} \le 2^m r_{\min}$. Thus,

$$q_{\min} \ge \sum_{i=0}^{m} (\lambda(t_o + i \cdot \Delta) - r(t_o + i \cdot \Delta)) \cdot \Delta \tag{4}$$

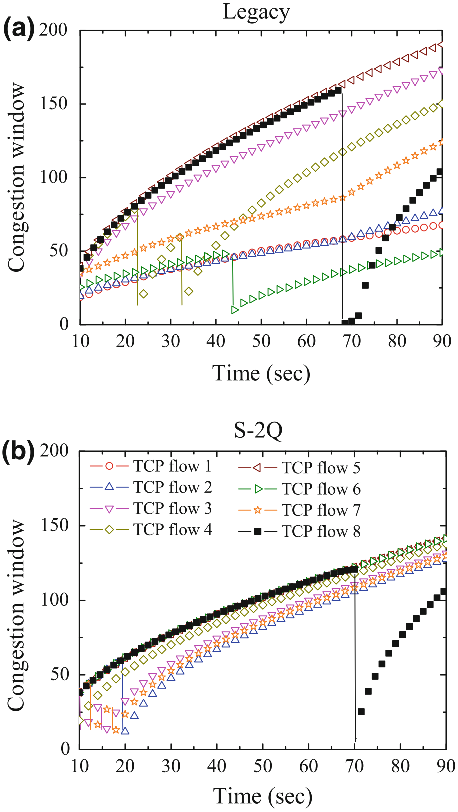
$$= r_{\max} \cdot \Delta + \sum_{i=1}^{m} \left(r_{\max} - 2^{i-1} r_{\min} \right) \cdot \Delta \tag{5}$$

$$= \left((m+1) \cdot r_{\max} - (2^m - 1) \cdot r_{\min} \right) \cdot \Delta \tag{6}$$

$$\ge \left(\lfloor \log_2 \frac{r_{\max}}{r_{\min}} \rfloor \cdot r_{\max} + r_{\min} \right) \cdot \Delta. \tag{7}$$

In Eq. (7), the equality holds when $\lfloor \log_2 \frac{r_{\max}}{r_{\min}} \rfloor = \log_2 \frac{r_{\max}}{r_{\min}}$.

Fig. 12 TCP congestion window size variation during a handoff: **a** legacy model, **b** S-2Q model



Appendix 2: Proof of Proposition 2

Assume that a *req_overflow_protecting* message was generated in $[t_o, t_o+\Delta)$. To avoid queue overflow immediately after a new arrival at $t_o + \Delta$, the residual queue length $L_{AP} - q_{\max}$ must be greater than $(\lambda(t_o+\Delta) - r(t_o)) \cdot \Delta$. As in the proof of Proposition 1, we consider the worst case in a similar way $r(t_o) = r(t_o + \Delta) = \cdots = r(t_o + m \cdot \Delta) = 0$ and $\lambda(t_o) = r_{\max}$, when $q(t) \geq q_{\max}$ for $t_o \leq t \leq t_o + (m+1) \cdot \Delta$ and the queue length keeps increasing.

According to our algorithm, $\lambda(t_o+\Delta) = \lambda(t_o)/2$, and thus, $\lambda(t_o+m\cdot\Delta) = \lambda(t_o)/2^m$. We find the greatest m that satisfies $\lambda(t_o + (m-1)\cdot\Delta) > r_{\min}$. That is, $\lambda(t_o+m\cdot\Delta) = r_{\min}$ and $\lambda(t_o + (m+1)\cdot\Delta) = 0$ according to our algorithm. Thus, $r_{\max}/2^{m-1} > r_{\min} \geq r_{\max}/2^m$. Similarly to the proof of Proposition 1, *we have*

$$L_{AP} - q_{\max} \geq \sum_{i=1}^{m+1} (\lambda(t_o+i\cdot\Delta) – r(t_o+(i-1)\cdot\Delta))\cdot\Delta \tag{8}$$

$$= \left(\sum_{i=1}^{m-1} r_{\max}/2^i + r_{\min}\right)\cdot\Delta \tag{9}$$

$$= \left(\left(1 - \frac{1}{2^{m-1}}\right) r_{\max} + r_{\min}\right)\cdot\Delta \tag{10}$$

$$\geq (r_{\max} - r_{\min})\cdot\Delta \tag{11}$$

References

1. Harri, H., & Toskala, A. (2004). *WCDMA for UMTS: Radio access for third generation mobile communications*. London: Wiley.
2. WiMAX Forum. (Jan, 2008). *WiMAX Forum Network Architecture* Rel. 1, ver. 1.2.
3. Jalali, A., Padovani, R., & Pankaj, R. (May, 2000). Data throughput of CDMA-HDR a high efficiency-high data personal communication wireless system. In *Proceedings of IEEE VTC-Spring*, Tokyo, Japan.
4. Xin, L., Chong, E. K. P., & Shroff, N. B. (2001). Opportunistic transmission scheduling with resource-sharing constraints in wireless networks. *IEEE JSAC, 19*(10), 2053–2064.
5. IEEE 802.16-REVd/D4-2004. (March, 2004). *Part16: Air Interface for Fixed Broadband Wireless Acces Systems*.
6. Suman, D., Viswanathan, H. & Rittenhouse, G. (Mar. 2003). Dynamic Load Balancing Through Coordinated Scheduling in Packet Data Systems. In *Proceedings of IEEE INFOCOM 2003*, San Francisco, CA, USA.
7. Guoqing, L., & Lu, H. (2006). Downlink radio resource allocation for multi-cell OFDMA system. *IEEE Transactions on Wireless Communications, 5*(12), 3451–3459.
8. David, T. C., & Vukovic, I. N. (Nov. 2002). CDMA 1X radio access network IP backhaul sizing analysis. In *Proceedings of IEEE Globecom 2002*. Taipei, Taiwan.
9. Rangsan, L., et al. (Aug. 2004). Performance Analysis of 1X EV-DO Systems Under Realistic Traffic Models and Limited-Size IP Backhaul. In *Proceedings of IEEE Asia Pacific Conference Communications 2004*. Beijing, China.
10. David, A., (Seffi) Naor, J. & Raz, D. (May 2007). Algorithmic aspects of access networks design in B3G/4G cellular networks. In *Proceedings of IEEE INFOCOM 2007*. Anchorage, AK, USA.
11. Harish, V., & Mukherjee, S. (2006). Throughput-range tradeoff of wireless mesh backhaul networks. *IEEE JSAC, 24*(3), 593–602.
12. Girija, N., Wilfong, G. & Zhang, L. (April 2006). Designing multihop wireless backhaul networks with delay guarantees. In *Proceedings of IEEE INFOCOM 2006*. Barcelona, Spain.
13. Mark, E. C., & Bestavros, A. (1997). Self-similarity in world wide web traffic: Evidence and possible causes. *IEEE Transactions on Networking, 5*(6), 835–846.
14. 3GPP2 C.S0024 Ver3.0. (Dec. 5, 2001). *cdma2000 High Rate Packet Data Air Interface Specification*.
15. IEEE 802.16e-2005. (Feb. 2006). *Part 16: Air interface for fixed and mobile broadband wireless access systems amendment for physical and medium access control layers for combined fixed and mobile operation in licensed bands*.
16. Debashis, S., Mukherjee, A., Saha Misra, I., & Chakraborty, M. (2004). Mobility support in IP: A survey of related protocols. *IEEE Network, 18*(6), 34–40.
17. Chun-Ting, C., & Shin, K. G. (Aug. 2005). Smooth handoff with enhanced packet buffering-and-forwarding in wireless/mobile networks. In *Proceedings of international conference on quality of service in heterogeneous wired/wireless networks (QShin)*. Orlando, FL, USA.
18. Ramon, C., & Padmanabhan, V. N. (1998). Fast and scalable wireless handoffs in support of mobile internet audio. *Mobile Networks and Applications, 3*(4), 351–363.
19. Chun-Ting, C., & Shin, K. G. (2005). An enhanced inter-access point protocol for uniform intra and intersubnet handoffs. *IEEE Transactions on Mobile Computers, 4*(4), 321–334.
20. Bernhardt, R. (1987). Macroscopic diversity in frequency reuse radio systems. *IEEE Journal of Selection in Areas Communications, 5*(5), 862–870.
21. WiMAX forum. (Aug. 2006). *Mobile WiMAX—Part I: A Technical Overview and Performance Evaluation*.

Author Biographies

Young-June Choi received B.S., M.S., and Ph.D. degrees from the Department of Electrical Engineering & Computer Science, Seoul National University, in 2000, 2002, and 2006, respectively. From Sept. 2006 through July 2007, he was a postdoctoral researcher at the University of Michigan, Ann Arbor, MI, USA. From 2007 through 2009, he was with NEC Laboratories America, Princeton, NJ, USA, as research staff member. He is currently an assistant professor at Ajou University, Suwon, Korea. His research interests include fourth-generation wireless networks, radio resource management, and cognitive radio networks. He was the recipient of the Gold Prize at the Samsung Humantech Thesis Contest in 2006.

Kyungtae Kim received the M.S. degree in computer science from Columbia University, New York and the Ph.D. degree at the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY. He is currently working with NEC Laboratories America, Inc., Princeton, NJ, in the areas of multimedia communication over the wireless network, mobility management, fixed mobile convergence, and mobile unified communication.

Kang G. Shin is the Kevin and Nancy O'Connor Professor of Computer Science and Founding Director of the Real-Time Computing Laboratory in the Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, Michigan. His current research focuses on computing systems and networks as well as on embedded real-time and cyber-physical systems, all with emphasis on timeliness, security, and dependability. He has supervised the completion of 67 PhDs, authored/coauthored more than 740 technical articles (263 of which are published in archival journals) and more than 20 patents or invention disclosures. He has co-authored (with C. M. Krishna) a textbook "Real-Time Systems," McGraw Hill, 1997. He has received numerous best paper awards, including the Best Paper from the 2010 USENIX Annual Technical Conference, the IEEE Communications Society William R. Bennett Prize Paper Award in 2003, the Best Paper Award from the IWQoS'03 in 2003, and an Outstanding IEEE Transactions of Automatic Control Paper Award in 1987. He has also coauthored papers with his students which received the Best Student Paper Awards from the 1996 IEEE Real-Time Technology and Application Symposium, and the 2000 UNSENIX Technical Conference. He has also received several institutional awards, including the Research Excellence Award in 1989, Outstanding Achievement Award in 1999, Service Excellence Award in 2000, Distinguished Faculty Achievement Award in 2001, and Stephen Attwood Award in 2004 from The University of Michigan (the highest honor bestowed to Michigan Engineering faculty); a Distinguished Alumni Award of the College of Engineering, Seoul National University in 2002; 2003 IEEE RTC Technical Achievement Award; and 2006 Ho-Am Prize in Engineering (the highest honor bestowed to Korean-origin engineers). He received the B.S. degree in Electronics Engineering from Seoul National University, Seoul, Korea in 1970, and both the M.S. and Ph.D. degrees in Electrical Engineering from Cornell University, Ithaca, New York in 1976 and 1978, respectively. From 1978 to 1982 he was on the faculty of Rensselaer Polytechnic Institute, Troy, New York. He also chaired the Computer Science and Engineering Division, EECS Department, The University of Michigan for three years beginning January 1991. He has held visiting positions at the U.S. Airforce Flight Dynamics Laboratory, AT&T Bell Laboratories, Computer Science Division within the Department of Electrical Engineering and Computer Science at UC Berkeley, and International Computer Science Institute, Berkeley, CA, IBM T. J. Watson Research Center, Software Engineering Institute at Carnegie Mellon University, HP Research Laboratories, Hong Kong University of Science and Technology, Ewha Womans University in Korea, and Ecole Polytechnique Federale de Lausanne (EPFL) in Switzerland. He is Fellow of IEEE and ACM, and overseas member of the Korean Academy of Engineering, has chaired numerous major conferences, including 2009 ACM Annual International Conference on Mobile Computing and Networking (MobiCom'09), 2008 IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON'08), the 3rd ACM/USENIX International Conference on Mobile Systems, Applications, and Services (MobiSys'05), 2000 IEEE Real-Time Technology and Applications Symposium (RTAS'00), and 1987 IEEE Real-Time Systems Symposium (RTSS). He also chaired the IEEE Technical Committee on Real-Time Systems during 1991-93, was a Distinguished Visitor of the Computer Society of the IEEE, an Editor of IEEE Trans. on Parallel and Distributed Computing, and an Area Editor of International Journal of Time-Critical Computing Systems, Computer Networks, and ACM Transactions on Embedded Systems. He has also served or is serving on numerous government committees, such as the US NSF Cyber-Physical Systems Executive Committee and the Korean Government R&D Strategy Advisory Committee.