Downstream Backhaul Packet Control for Mobile Wireless Networks

Young-June Choi
Ajou University
Suwon 443-749, South Korea
choiyj@ajou.ac.kr

Kyungtae Kim
NEC Laboratories America
Princeton, NJ 08540, U.S.A.
kyungtae@nec-labs.com

Kang G. Shin
The University of Michigan
Ann Arbor, MI 48109, U.S.A.
kgshin@eecs.umich.edu

***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 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.**

I. Introduction

For efficient radio resource and mobility management, mobile wireless networks have employed a hierarchical backhaul structure. In 2G or 3G cellular systems, 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), a BS—not the upper entity, RNC—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 [1]. 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, 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 [2], [3]. The proposed mechanisms are applied and evaluated for IEEE 802.16 systems [4].

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 [5], [6], 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 [7], [8]. There has also been extensive research on the topology design of backhaul access networks (see [9] and references therein), and mesh backhaul networks (e.g., [10], [11]). 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 Section II, we discuss the advantages of the S-2Q model over the existing legacy model. In Section III, we present a rate-control algorithm for backhaul links and derive the conditions of its stable operation. In Section IV, 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 Section V, the proposed algorithm is simulated to demonstrate its ability of handling the two user

types. Finally, the paper concludes with Section VI.

II. 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 [12]. 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.

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 $Pr\{q_i^{BS} \le L^+\}$, 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 $Pr\{\sum_{i=1}^{N} q_i^{AR} \le N \cdot L',\ q_i^{BS} \le L\}$, 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:

$$Pr\{\sum_{i=1}^{N} q_i^{AR} \le N \cdot L',\ q_i^{BS} \le L\} \quad (1)$$

$$> Pr\{q_1^{AR}+q_1^{BS} \le L+L', \cdots, q_N^{AR}+q_N^{BS} \le L+L'\} \quad (2)$$

$$= Pr\{q_1^{BS} \le L^+\} \cdots Pr\{q_N^{BS} \le L^+\}. \quad (3)$$

In the legacy model, Eq. (2) is equal to Eq. (3) as $q_i^{AR} = 0$. Eq. (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.

III. Rate-Control Algorithm

A. 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 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$_i$.

B. 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.

[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 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.

BS operation: after receiving data

1: if $q_i < q_{\min}$
2: send a *req_underflow_preventing*
3: else if $q_i > q_{\max}$
4: send a *req_overflow_preventing*
5: end if

AR operation: before sending data

1: if receive a *req_underflow_preventing*
2: if $\lambda_i = 0$
3: $\lambda_i \leftarrow r_{\min}$
4: else
5: $\lambda_i \leftarrow \min(2\lambda_i, r_{\max})$
6: end if
7: else if receive a *req_overflow_preventing*
8: if $\lambda_i = r_{\min}$
9: $\lambda_i \leftarrow 0$
10: else
11: $\lambda_i \leftarrow \max(\lambda_i/2, r_{\min})$
12: end if
13: else
14: $\lambda_i \leftarrow \min(\lambda_i * \theta, r_{\max})$
15: end if

Fig. 1. The proposed rate-control algorithm for backhaul links.

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, given in Fig. 1.

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.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, as shown in Fig. 1.

C. 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. 2. 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.

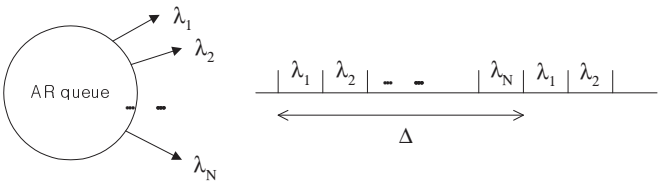

Fig. 2. Scheduling over backhaul links at an AR.

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 $\lambda_i \leq r_i$. 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.

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

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

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

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

Proposition 1 is proved in Appendix A. Proposition 2 is similarly proved and Proposition 3 is straightforward from both propositions, so we omit the proofs. The above propositions ensure our rate-control algorithm to achieve stability in queue length.

IV. 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.

A. 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 [13].

To prevent packet loss during a handoff, "packet buffering-and-forwarding" method has been considered [14]. 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 [4]. 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.

B. 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.

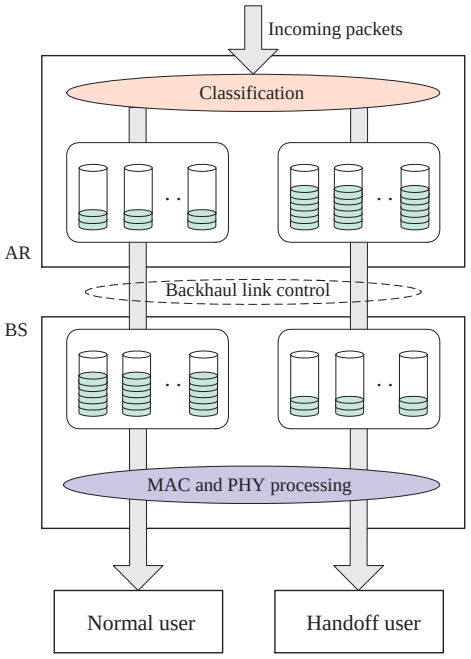

Fig. 3. The user-adaptive S-2Q model.

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., [2], [3]). 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 [3].

C. 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.

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

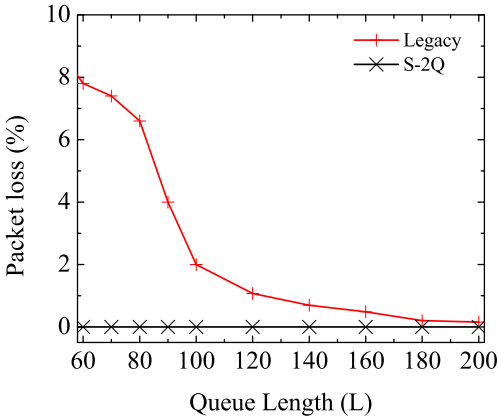

Fig. 4. Comparison of packet drops in the legacy and S-2Q strategies.

Fig. 3. 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.

When MTs are located in cell-edge areas, they can also be managed separately in order to solve the well-known inter-cell interference problem and enhance cell-edge performance. Those users are likely to be supported by dynamic frequency reuse [6] or macro-diversity [4], [15] that does not exploit opportunistic scheduling, so the handoff user type can subsume such cell-edge users as well as handoff users.

V. 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 [16]. MTs download FTP traffic from a server, and the propagation delay between the FTP server and the AR is set to 30 msec. As the backhaul link has sufficient bandwidth, its propagation delay is assumed to be 1 msec. The simulation time is 300 seconds. 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 msec. 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.

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. Fig. 4 plots the packet-drop rate in the entire subnet as a function of L. As argued in Section II, 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.

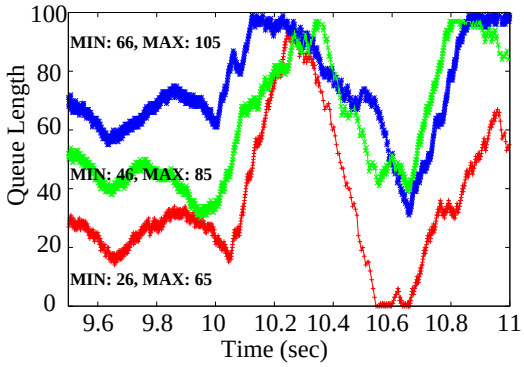

Fig. 5. Effects of different combinations of $q_{\min}$ and $q_{\max}$ on queue length in the S-2Q model.

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*. Fig. 5 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 second starting from the 10 second point, so the queue length begins to increase at the 10 second 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 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.

Fig. 6 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.

VI. 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

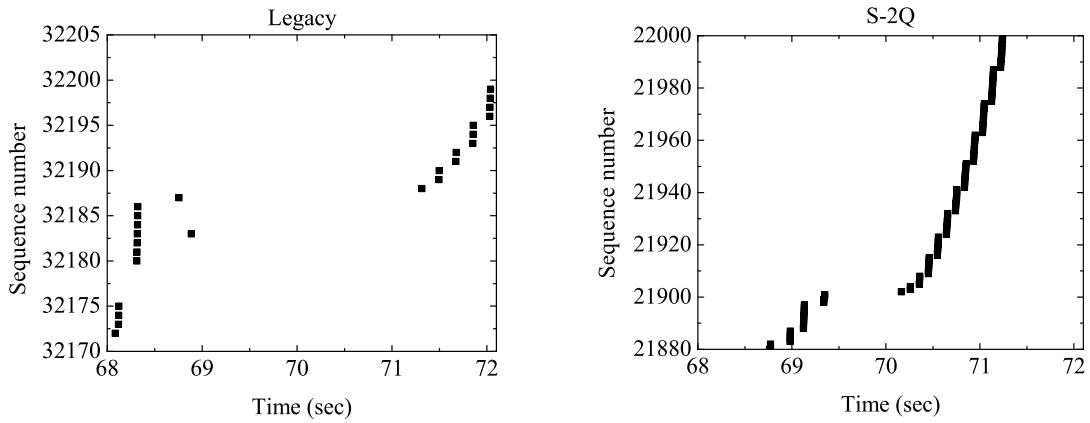

Fig. 6. TCP sequence number during a handoff.

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 A
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 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,

$$
\begin{aligned}
q_{\min} &\ge \sum_{i=0}^{m} (\lambda(t_o + i \cdot \Delta) - r(t_o + i \cdot \Delta)) \cdot \Delta && (4)\\
&= r_{\max} \cdot \Delta + \sum_{i=1}^{m} \left(r_{\max} - 2^{i-1} r_{\min}\right) \cdot \Delta && (5)\\
&= ((m+1) \cdot r_{\max} - (2^m - 1) \cdot r_{\min}) \cdot \Delta && (6)\\
&\ge \left(\lfloor \log_2 \frac{r_{\max}}{r_{\min}} \rfloor \cdot r_{\max} + r_{\min} \right) \cdot \Delta. && (7)
\end{aligned}
$$

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

REFERENCES

[1] WiMAX Forum, "*WiMAX Forum Network Architecture*," Rel. 1, ver. 1.2, Jan. 2008.

[2] A. Jalali, R. Padovani, and R. Pankaj, "Data Throughput of CDMA-HDR a High Efficiency-High Data Personal Communication Wireless System," in Proc. *IEEE VTC-Spring*, Tokyo, Japan, May 2000.

[3] Xin Liu, Edwin K. P. Chong, and Ness B. Shroff, "Opportunistic Transmission Scheduling with Resource-Sharing Constraints in Wireless Networks," *IEEE JSAC*, vol. 19, no. 10, pp. 2053-2064, Oct. 2001.

[4] IEEE 802.16e-2005, "*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"*, Feb. 2006.

[5] Suman Das, Harish Viswanathan, Gee Rittenhouse, "Dynamic Load Balancing Through Coordinated Scheduling in Packet Data Systems," in Proc. *IEEE INFOCOM 2003*, San Francisco, CA, USA, Mar. 2003.

[6] Guoqing Li and Hui Lu, "Downlink Radio Resource Allocation for Multi-cell OFDMA System," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3451-3459, Dec. 2006.

[7] David T. Chen and Ivan N. Vukovic, "CDMA 1X Radio Access Network IP Backhaul Sizing Analysis," in Proc. *IEEE Globecom 2002*, Taipei, Taiwan, Nov. 2002.

[8] Rangsan Leelahakriengkrai, *et al*., "Performance Analysis of 1X EV-DO Systems Under Realistic Traffic Models and Limited-Size IP Backhaul," in Proc. *IEEE Asia Pacific Conf. Commun. 2004*, Beijing, China, Aug. 2004.

[9] David Amzallag, Joseph (Seffi) Naor, and Danny Raz, "Algorithmic Aspects of Access Networks Design in B3G/4G Cellular Networks," in Proc. *IEEE INFOCOM 2007*, Anchorage, AK, USA, May 2007.

[10] Harish Viswanathan and Sayandev Mukherjee, "Throughput-Range Tradeoff of Wireless Mesh Backhaul Networks," *IEEE JSAC*, vol. 24, no. 3, pp. 593-602, March 2006.

[11] Girija Narlikar, Gordon Wilfong, and Lisa Zhang, "Designing Multihop Wireless Backhaul Networks with Delay Guarantees," in Proc. *IEEE INFOCOM 2006*, Barcelona, Spain, April 2006.

[12] Mark E. Crovella and Azer Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes," *IEEE Trans. Networking*, vol. 5, no. 6, pp. 835-846, Dec. 1997.

[13] Debashis Saha, Amitava Mukherjee, Iti Saha Misra, and Mohuya Chakraborty, "Mobility Support in IP: A Survey of Related Protocols," *IEEE Network*, vol. 18, no. 6, Nov. 2004, pp. 34-40.

[14] Ramon Caceres and Venkata N. Padmanabhan, "Fast and Scalable Wireless Handoffs in Support of Mobile Internet Audio," *Mobile Networks and Applications*, vol. 3, no. 4, pp. 351-363, 1998.

[15] R. Bernhardt, "Macroscopic Diversity in Frequency Reuse Radio Systems," *IEEE J. Select. Areas Commun.*, vol. 5, no. 5, pp. 862-870, Jun. 1987.

[16] WiMAX forum, "*Mobile WiMAX - Part I: A Technical Overview and Performance Evaluation*," Aug. 2006.