Location/Mobility-Dependent Bandwidth Adaptation in QoS-Sensitive Cellular Networks

Sunghyun Choi
Philips Research-USA, Briarcliff Manor, New York
E-mail: sunghyun.choi@philips.com

Kang G. Shin
The University of Michigan, Ann Arbor, Michigan
E-mail: kgshin@eecs.umich.edu

Abstract— **Wireless/mobile networks are highly dynamic due to (1) time-varying and location-dependent channel conditions and (2) user mobility. These characteristics cause wireless link bandwidth, one of the scarcest and most precious resources, to fluctuate severely. Presented is how to manage the wireless link bandwidth allocated to each connection *adaptively* in this highly dynamic environment so as to maximize the service provider's total reward/revenue. Bandwidth adaptation can be triggered by either arrival/departure of a connection (due to setup, hand-off, or termination) or change of the location-dependent channel condition. Using simulations, we demonstrate how the proposed scheme works, and its superiority to a non-adaptive scheme in terms of link utilization and aggregate reward/revenue.**

I. Introduction

Recently, there has been a rapid growth of research and development efforts to provide mobile users a means of seamless communication via wireless media. This has made it possible to implement and deploy current cellular systems, PCS, and wireless LANs. Compared to their wired counterpart, wireless/mobile networks must deal with several problems which make it very difficult to provide Quality-of-Service (QoS): (1) resources (e.g., bandwidth) in wireless networks are scarcer than in wired networks; (2) wireless channels are prone to location-dependent, bursty, and time-varying errors; (3) users tend to move around during a communication session causing hand-offs between cells. Due to these distinct characteristics, it is essential to develop mechanisms tailored to support QoS for mobile users. The key to support wireless/mobile QoS is how to manage wireless link bandwidth efficiently in a highly dynamic environment with time-varying channel conditions and user mobility.

In view of rapidly-fluctuating wireless link bandwidths, we use an *adaptive QoS* concept in which a connection's required bandwidth is not fixed at a single value, but is given as a *range* of bandwidth, $[b_{min}, b_{max}]$. Ideally, it is desirable to always provide the maximum required bandwidth b_{max} for each connection. However, as a user moves around, or due to dynamically-fluctuating channel and network conditions, this may not always be possible. The minimum required bandwidth b_{min} can be considered as the bandwidth required to support the lowest-level QoS the mobile user can "live with." Hence, we can use the minimum required bandwidth b_{min} for admitting/rejecting each new/handed-off connection.

Using this adaptive QoS concept, it is possible to utilize bandwidth more efficiently, and thus increase the service provider's reward/revenue, while reducing the number of hand-off drops and new connection blocks. In this paper, we present and evaluate an adaptive QoS framework and bandwidth-adaptation mechanisms which maximize the service provider's reward/revenue. While such adaptive QoS concepts/mechanisms were studied by others [3], [4], [7], distinct features of our scheme include: (1) considering bandwidth fluctuation and adaptation due to location-dependent channel condition variation; (2) penalizing connections which initiate adaptation actions too frequently; and (3) incorporating application-specific adaptation constraints, such as how often and how much of adaptation can be made.

This paper is organized as follows: Section II describes the wireless system model under consideration. Our adaptive bandwidth-allocation mechanism is presented in Section III. Using simulations, the performance of our scheme is evaluated, and compared with that of a non-adaptive scheme in Section IV. The paper concludes with Section V.

II. System Description

A. Cellular Networks

We will consider a cellular network, in which a mobile communicates with another party in the network, via a base station (BS) while staying in the cell of the BS. When a user moves to an adjacent cell in the middle of a communication session, a hand-off will enable the mobile to maintain connectivity to its communication partner. The cellular system uses a fixed channel allocation (FCA) scheme, and cell i has a wireless link capacity $C(i)$. The unit of link bandwidth is BU. Let $C_{i,j}$ be connection j in cell i, and $\mathbf{S}_i$ be the set of indices of connections in cell i.

B. Adaptive Error Handling

Our scheme uses adaptive error control to handle time-varying channel conditions. This scheme may be based on adaptive modulation, e.g., the physical layer of IEEE 802.11 [2], or adaptive usage of error-control codes [6], or both. A common characteristic of these techniques is that the better the channel condition, the more efficiently the channel resources can be used, i.e., a higher transmission rate can be achieved with the same bandwidth. We represent the *bandwidth usage efficiency* of a connection with r_c (≤ 1), i.e., $b_g = r_c w_a$, where w_a is the bandwidth allocated to this connection and b_g is the actual throughput (or rate) perceived by (or granted to) this connection. Note that r_c is a time-varying function of the location of the connection and its environment, and will be determined by the underlying error-control scheme.[1]

For the given link capacity $C(i)$ of cell i, the sum $W_{a,i}$ of bandwidths allocated to all the connections in the cell should be bounded by $C(i)$:

$$W_{a,i} = \sum_{j \in \mathbf{S}_i} w_a(C_{i,j}) \leq C(i), \tag{1}$$

The work reported in this paper was supported in part by AFOSR under Grant No. F49620-00-1-0327.

[1] Conceptually, r_c can be interpreted as the channel code rate of the code being used if an adaptive error-control strategy is used.

0-7803-7005-8/01/$10.00 © 2001 IEEE

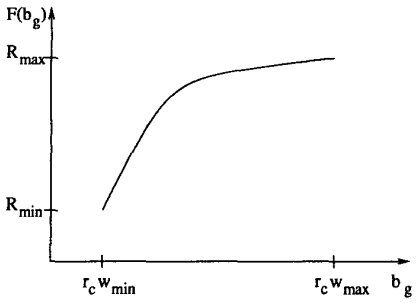

Fig. 1. The reward function $F(b_g)$, where $b_g = r_c w_a$, and $w_{max\ (min)} = b_{max\ (min)}/r_c$.

where $w_a(C_{i,j})$ is the bandwidth allocated to connection $C_{i,j}$. Note that the sum, $B_{g,i} = \sum_{j \in \mathbf{S}_i} b_g(C_{i,j})$, of the user-perceived actual throughput, where $b_g(C_{i,j})$ is the actual throughput perceived by connection $C_{i,j}$, and so its upper-bound, which is actually the user-perceived link capacity, is time-varying and dependent on the bandwidth allocated to each connection, i.e., $w_a(C_{i,j})$. We assume that r_c can have a number of different discrete values, i.e., $r_c \in \{r_1\ (= r_{min}), r_2, \cdots, r_n\ (= r_{max})\}$ since only limited sets of modulations and error-control codes are used due mainly to the resultant computational complexity proportional to the set size. The underlying adaptive error-control scheme should be good enough to handle long-term variations of the channel, e.g., resulting from shadowing and signal attenuation, while it may not handle short-term variations, e.g., resulting from short-term fading. Short-term variations can be handled by proper MAC and packet scheduling/retransmission. These combined strategies render a specific (average) amount b_g of actual throughput perceived by each connection at the cost of bandwidth $(1 - r_c)w_a$. Note that the bandwidth usage efficiency r_c should reflect the time-averaged — not instantaneous — behavior of the channel a connection is experiencing.

C. Adaptive Application and Reward/Revenue Models

We assume that each connection is specified by a *range*, $[b_{min}, b_{max}]$, of the required actual throughput. For example, real-time multimedia traffic can be represented and transported as multi-layer scalable flows [5]. For coded video and audio, for instance, it is possible to change the output rate by adjusting some encoding parameters. A non-real-time connection might require the minimum throughput it can "live with." On the other hand, some non-real-time connections may not require any guaranteed throughput at all. All the traffic of this type can be handled in an aggregate manner without per-connection bandwidth allocation. More specifically, the required actual throughput of a connection is given by a set of discrete throughput values, $\mathbf{B} = \{b_1(= b_{min}), b_2, \cdots, b_m(= b_{max})\}$, depending on the underlying application. Note that the required bandwidth range, i.e., $[w_{min}, w_{max}]$, where $w_{max\ (min)} = b_{max\ (min)}/r_c$, of a connection with the required actual throughput range $[b_{min}, b_{max}]$ is time-varying, depending on the value of r_c.

A reward function is chosen such that the service provider earns $F(b_g)$ units of reward/profit by providing actual throughput $b_g \in \mathbf{B}$ to a connection or an application. The reward function F is connection-specific, and can be an arbitrary non-decreasing function of b_g, depending on the adopted pricing scheme. Figure 1 shows an example reward function curve. Note that the reward function seen by the service provider (i.e., given as a function of the allocated bandwidth $w_a\ (= b_g/r_c)$) is a time-varying function (since it depends on the time-varying parameter r_c).

Depending on its application, a connection specifies (1) its desired minimum duration Δt between two consecutive adaptations, called the *desired minimum inter-adaptation interval*, and (2) the desired maximum change Δb on its granted throughput in one adaptation process, called the *desired maximum throughput adaptation*. These will henceforth be called the connection-specific *adaptation constraints*. For example, a real-time video connection might require Δt to be several seconds and Δb to be a moderate positive number since too frequent and drastic fluctuation of the perceived video quality is not desirable, while a non-real-time connection probably specifies them to be zero since the more bandwidth is assigned, the better in this case. Note that these two values specify how often and how much a connection's perceived actual throughput can be adapted. In summary, connection $C_{i,j}$ is specified by a set: $\{\mathbf{B}_{i,j}, F_{i,j}(\cdot), \Delta t_{i,j}, \Delta b_{i,j}\}$.

III. Adaptive QoS via Bandwidth Management

This section describes our framework for bandwidth management to support adaptive QoS in cellular networks so as to maximize the service provider's total reward subject to connection-specific adaptation constraints.

A. Bandwidth-Adaptation Mechanism

In a cellular network, the BS of a cell has complete authority to determine its link condition and control the link bandwidth allocated to each connection (and the bandwidth reserved for hand-offs [1]). In this paper, we concentrate on the adaptation actions initiated or ordered by BSs due to the wireless link load fluctuations.

There are two types of adaptation actions: bandwidth *upgrade* and *downgrade*. Bandwidth downgrade actions are crucial since they directly affect the rejection of new connection requests and the dropping of handed-off/on-going connections. There are two cases that invoke a downgrade action: (1) the arrival of a new or handed-off connection at a cell, thus causing the sum of allocated bandwidth to exceed the link capacity; and (2) the channel condition degradation of a connection, thus requiring more bandwidth for error-control redundancy. An upgrade action, on the other hand, can be triggered when additional bandwidth becomes available as a result of (1) a connection termination, (2) outgoing hand-offs, or (3) an improved channel condition that decreases the need for a connection's error-control redundancy. However, unlike event-driven downgrade actions, upgrade actions should be performed more judiciously utilizing each connection's desired minimum inter-adaptation interval. This way we can reduce a wireless link bandwidth oscillation with frequent upgrade and downgrade actions, thus causing significant overhead and user-perceived quality to fluctuate severely, both of which are undesirable. Note that adaptation actions are not free; they too will consume resources.

B. Connection-Admission Control

As mentioned above, a downgrade action can be triggered by the admission of a new or handed-off connection. Described below is an admission-control process for both new and handed-off connections. Let $b_{min}(C_{i,j})$ and $b_{max}(C_{i,j})$ be $C_{i,j}$'s minimum and maximum required actual throughputs, respectively.

The admission-control process checks if it is always possible to provide each connection its minimum required actual throughput. The admission test for a handed-off connection C_{ho} in cell i can be represented by

$$\sum_i b_{min}(C_{i,j})/\hat{r}_c(C_{i,j}) + b_{min}(C_{ho})/r_{min} \leq C(i), \quad (2)$$

where the *worst-case* bandwidth usage efficiency $\hat{r}_c$ is given by

$$\hat{r}_c(C_{i,j}) = \begin{cases} r_c(C_{i,j}), & \text{if } C_{i,j} \text{ is stationary,} \\ r_{min}, & \text{if } C_{i,j} \text{ is moving.} \end{cases} \quad (3)$$

That is, the $\hat{r}_c$ value of a stationary connection is determined at its current bandwidth usage efficiency while that of a moving connection is determined at the minimum bandwidth usage efficiency r_{min}. The rationale behind using $\hat{r}_c(C_{i,j})$ instead of $r_c(C_{i,j})$ in Eq. (2) is that $r_c(C_{i,j})$ of a moving connection is supposed to vary over time, and the connection might be dropped in a cell when its channel condition gets worse. For the same reason, r_{min} is used for a handed-off connection. When a newly-admitted connection is moving, we assume the connection to be stationary if its bandwidth usage efficiency does not vary and it does not hand off during a certain time period. The channel a stationary connection experiences could vary over time due to its time-varying environment. However, the "average" channel behavior will not change much, so r_c — which is determined based on the time-averaged behavior of a channel — will stay virtually constant.

Admission control of new connections is based on the scheme proposed in [1], in which the adaptive QoS concept was not used. A portion of the link capacity is set aside in each cell for possible hand-offs from its adjacent cells to keep the hand-off dropping probability below a target value. This reserved bandwidth can be used only for hand-offs from adjacent cells, but *not* for admitting new connections in the cell. A newly-requested connection C_{new} in cell i requires an admission test:

$$\sum_i b_{min}(C_{i,j})/\hat{r}_c(C_{i,j}) + b_{min}(C_{new})/r_{min} \leq C(i) - W_{r,i}, \quad (4)$$

where $W_{r,i}$ is the *target reservation bandwidth* — the required bandwidth to be reserved for hand-offs — in cell i. Upon arrival of a new connection request, $W_{r,i}$ is updated predictively and adaptively — before performing the admission test (4) on the request — depending on the neighbors' loads. For the admission decision, in addition to checking bandwidth availability in cell i as in Eq. (4), some neighbors of cell i also check their bandwidth availability. See [1] for more details on this.

The maximum actual throughput $b_{max}(C_{new})$ is assigned initially to the newly-admitted connection. Then, connection-specific parameters are defined as:

$$\Delta t^+_{new} := \Delta t^-_{new} := \Delta t_{new}, \quad (5)$$

$$\Delta b^+_{new} := \Delta b^-_{new} := \Delta b_{new}, \quad (6)$$

where Δt^+ (Δt^-) is the *target* minimum time for an upgrade (downgrade) action measured since the last adaptation, and Δb^+ (Δb^-) is the *target* upper-bound on the actual throughput increase (decrease) in each upgrade (downgrade) action. If there happens to be a bandwidth demand conflict after admitting a new or handed-off connection, a downgrade action is invoked to resolve it.

C. Adaptation Rule

Basic questions regarding bandwidth adaptation are (1) when to adapt, (2) which connections to be adapted, and (3) how much to adapt. Our bandwidth adaptation will attempt to maximize the aggregate reward from already-admitted connections while satisfying connection-specific adaptation constraints. To better explain this adaptation rule, we define two parameters for each connection, i.e., the upgrade slope $\Delta R^+_{i,j}$ and the downgrade slope $\Delta R^-_{i,j}$ of connection $C_{i,j}$'s reward function $F_{i,j}$ are, respectively, defined by

$$\Delta R^+_{i,j} = \begin{cases} 0, & \text{if } l \text{ is the maximum,} \\ \frac{F_{i,j}(r_c(C_{i,j})w_{l+1}) - F_{i,j}(r_c(C_{i,j})w_l)}{w_{l+1} - w_l}, & \text{otherwise,} \end{cases} \quad (7)$$

$$\Delta R^-_{i,j} = \begin{cases} \infty, & \text{if } l = 1, \\ \frac{F_{i,j}(r_c(C_{i,j})w_l) - F_{i,j}(r_c(C_{i,j})w_{l-1})}{w_l - w_{l-1}}, & \text{otherwise,} \end{cases} \quad (8)$$

where l is the level of the actual bandwidth received by connection $C_{i,j}$, i.e., $b_g(C_{i,j}) = b_l$ or $l = level(C_{i,j})$.

When a new or handed-off connection is admitted, or a connection's channel condition gets deteriorated, the aggregate bandwidth allocation in the cell might become larger than the link capacity. Then, the allocated bandwidths of some connections (possibly including the newly-admitted or handed-off connection) should be reduced. The downgrade process in cell i works as shown in Fig. 2, where $\tau_{i,j}$ is the time elapsed since $C_{i,j}$'s last bandwidth adaptation, and $\beta_{i,j}$ is its actual throughput decrease during the current downgrade. If $C_{i,j}$ never experienced an adaptation before, $\tau_{i,j}$ is set to the time elapsed since its setup.

As clear from the pseudo-code, while meeting its adaptation constraints, a connection with the minimum downgrade slope is chosen and downgraded until the aggregate bandwidth allocation becomes less than, or equal to, the link capacity. If it is not possible to find a connection satisfying the adaptation constraints, the constraints are "loosened" until such a connection is found, and then this connection's constraints are tightened, i.e., Δt^- (Δb^-) is increased (decreased), to compensate for the violation of the constraints later. We believe that keeping all the on-going connections (or accommodating a new connection) is more important than not violating the connections' adaptation constraints. This is why Δt and Δb is said to be *desired* (as opposed to "required") values in Section II. If a newly-admitted connection initiates a downgrade action, then its Δb^- (Δt^-) is set to ∞ ($-\infty$) to render it unlimited adaptation from its maximum actual throughput provided initially.

$\Delta t_{adj} := 0$; $\Delta b_{adj} := 0$;
while ($W_{a,i} > C(i)$) {
  if ($\Delta t_{adj} = \max_j \Delta t^-_{i,j}$) $\Delta b_{adj} := \Delta b_{adj} + 1$ (BUs);
  while ($\nexists$ k s.t. $\Delta R^-_{i,k} \neq \infty$,
    $\tau_{i,k} > \Delta t^-_{i,k} - \Delta t_{adj}$, and $\beta_{i,k} < \Delta b^-_{i,k} + \Delta b_{adj}$)
    $\Delta t_{adj} := \Delta t_{adj} + 1$ (sec);
  choose j with min $\Delta R^-_{i,j}$,
    $\tau_{i,j} > \Delta t^-_{i,j} - \Delta t_{adj}$, and $\beta_{i,j} < \Delta b^-_{i,j} + \Delta b_{adj}$;
  $W_{a,i} := W_{a,i} - (w_l - w_{l-1})$, where $l = level(C_{i,j})$;
  $lev(C_{i,j}) := lev(C_{i,j}) - 1$;
  if ($\Delta t_{adj} \neq 0$) $\Delta t^-_{i,j} := \Delta t^-_{i,j} + \Delta t_{adj}$;
  else $\Delta t^-_{i,j} := \Delta t_{i,j}$;
  if ($\Delta b_{adj} \neq 0$)
    if ($\Delta b^-_{i,j} > 1$ (BU)) $\Delta b^-_{i,j} := \Delta b^-_{i,j} - 1$ (BU);
  else $\Delta b^-_{i,j} := \Delta b_{i,j}$;
}

Fig. 2. Resolving a bandwidth demand conflict via downgrade.

if ($C_{i,j}$ initiated a downgrade) {
  $\Delta t^+_{i,j} := \Delta t^+_{i,j} + 1$ (sec);
  if ($\Delta b^+_{i,j} > 1$) $\Delta b^+_{i,j} := \Delta b^+_{i,j} - 1$ (BU);
}

Fig. 3. Penalizing an on-going connection initiating a downgrade action.

As shown in Fig. 3, when on-going connection $C_{i,j}$ initiated a downgrade action due to its hand-off or perceived channel deterioration, it is "penalized" to have less chance to upgrade/upgrade in future. This will reduce the fluctuation of wireless link bandwidth. An upgrade action is triggered when the total allocated bandwidth is less than the unreserved bandwidth, i.e., $W_{a,i} < C(i) - W_{r,i}$.

The upgrade process in cell i works as shown in Fig. 4. Basically, while satisfying its adaptation constraints, a connection with the maximum upgrade slope is chosen, and upgraded as long as the aggregate allocated bandwidth is less than the unreserved bandwidth. When no adaptation action was invoked for a time period T_{adapt}, which is the maximum cell inter-adaptation interval, the process in Fig. 5 is invoked to reallocate bandwidths among the existing connections. This selects those connections satisfying adaptation constraints with the maximum upgrade slope and the minimum downgrade slope, and upgrades and downgrades their allocated bandwidths, respectively, if it results in a positive gain of reward. Even without any link-bandwidth fluctuation, each connection's status in terms of adaptation constraints varies with time, e.g., a connection becomes available for adaptation by passing Δt^+ or Δt^- since the last adaptation, so this adaptation increases the aggregate reward.

IV. Comparative Performance Evaluation

A. Simulation Assumptions and Specifications

In our simulation environment, mobiles are traveling along a straight road (e.g., cars on a highway). This environment is the simplest in the real world, representing a one-dimensional cellular system. We make the following assumptions:

A1. The cellular system is composed of 10 linearly-arranged cells with 1 km-diameter, where two end-most cells are connected together to form a ring structure.

A2. Connection requests are generated according to a Poisson process with rate λ (connections/second) in each cell. A newly-generated connection can appear anywhere in the cell with an equal probability.

index = 1;
while (*index* = 1 and
  $\exists$ k s.t. $\Delta R^+_{i,k} \neq 0$, $\tau_{i,k} > \Delta t^+_{i,k}$, and $\beta_{i,k} < \Delta b^+_{i,k}$) {
  choose j with max $\Delta R^+_{i,j}$, $\tau_{i,j} > \Delta t^+_{i,j}$, and $\beta_{i,j} < \Delta b^+_{i,j}$;
  $l = lev(C_{i,j})$;
  if ($W_{a,i} + (w_{l+1} - w_l) < C(i) - W_{r,i}$) {
    $W_{a,i} := W_{a,i} + (w_{l+1} - w_l)$;
    $lev(C_{i,j}) := lev(C_{i,j}) + 1$;
  }
  else *index* = 0;
}

Fig. 4. Allocating the residual bandwidth to connections.

index = 1;
while (*index* = 1) {
  choose j with max $\Delta R^+_{i,j}$, $\tau_{i,j} > \Delta t^+_{i,j}$, and $\beta_{i,j} < \Delta b^+_{i,j}$;
  choose k with min $\Delta R^-_{i,k}$, $\tau_{i,k} > \Delta t^-_{i,k}$, and $\beta_{i,k} < \Delta b^-_{i,k}$;
  $W'_{a,i} := W_{a,i} + (w_{l_1+1} - w_{l_1}) - (w_{l_2} - w_{l_2-1})$,
    where $l_1 = lev(C_{i,j})$ and $l_2 = lev(C_{i,k})$;
  if ($\Delta R^+_{i,j} > \Delta R^-_{i,k}$ and $W'_{a,i} \leq C(i) - W_{r,i}$) {
    $W_{a,i} := W'_{a,i}$;
    $lev(C_{i,j}) := lev(C_{i,j}) + 1$;
    $lev(C_{i,k}) := lev(C_{i,k}) - 1$;
  }
  else *index* = 0;
}

Fig. 5. Re-allocating bandwidths to connections.

A3. Mobiles can travel in either of two directions with an equal probability with a speed chosen randomly between 40 and 60 (km/hr). Each mobile will run straight through the road with the chosen speed, i.e., mobiles will never turn around.

A4. Each connection's lifetime is exponentially-distributed with mean 120 (seconds).

A5. A connection's required actual throughput set is given by $\mathbf{B} = \{1\ (= b_{min}), 2, 3, 4\ (= b_{max})\}$ (BUs). The reward function $F(\cdot)$ is given by $F(1) = 1$, $F(2) = 1.5$, $F(3) = 1.8$, and $F(4) = 2$. The desired minimum inter-adaptation interval $\Delta t = 5$ (sec). The desired maximum throughput adaptation $\Delta b = 2$ (BUs).

A6. A connection's bandwidth usage efficiency is given by $r_c = 0.95$ when its mobile is within 0.25 km range from a BS, and $r_c = 0.8$ otherwise.

A7. Each cell has a fixed link capacity 100 BUs, i.e., $C(i) = C = 100$ for all i.

Assumption **A6** represents the condition that (1) the farther from the BS, the worse the channel condition, which typically happens due to limited transmission power, and (2) two error-control codes are adaptively used to keep the user-perceived error probability below a given threshold.

The parameters used include: the target hand-off dropping probability $P_{HD,target} = 0.01$, and the maximum cell inter-adaptation interval $T_{adapt} = 10$ (sec). Other parameters relevant to the bandwidth reservation for hand-offs are assigned the same values used in [1] except for $N_{quad} = 10$. As a reference for comparison, we consider the performance of a non-adaptive scheme as well, in which each connection's required actual throughput is 1 (BU), and a fixed channel usage efficiency $r_c = 0.8$ is used irrespective of the connection's location, and the rest is the same as the scheme **AC3** in [1].

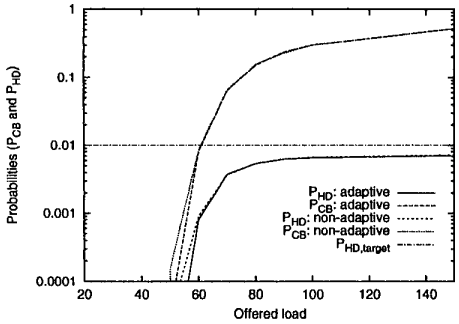

Fig. 6. P_{CB} and P_{HD} vs. offered load.

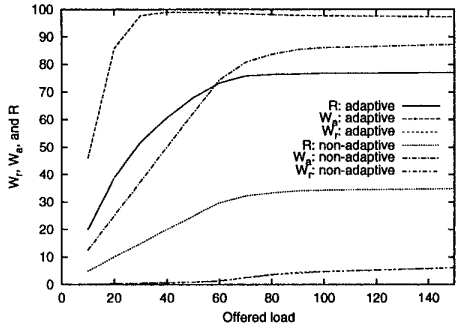

Fig. 7. W_r, W_a, and R vs. offered load.

B. Simulation Results and Discussion

Fig. 6 shows the performance of both the proposed and the non-adaptive schemes in terms of the new connection blocking probability P_{CB} and hand-off dropping probability P_{HD}. The *offered load* per cell, L, represents the aggregate throughput required on average to support the minimum required actual throughput to all existing connections in a cell, i.e., $L = \lambda \cdot 120$. First, we observe that P_{HD}'s for both schemes are bounded by $P_{HD,target}$ $(= 0.01)$ over the entire range of the offered load thanks to the bandwidth reservation for hand-offs. Moreover, both P_{CB} and P_{HD} of both schemes are almost the same, respectively. Obviously, this is because we use the minimum required actual throughput over the minimum bandwidth usage efficiency, i.e., $b_{min}(\cdot)/r_{min}$, for the admission control purpose with our adaptive scheme.

The utilization of the scarce wireless-link bandwidth and the aggregate revenue are also as important as connection admissibility. Fig. 7 shows the average revenue R, the average aggregate allocated bandwidth W_a, and the average bandwidth W_r reserved for hand-offs in each cell. First, we observe that W_a of our scheme is saturated at almost 100 BUs, which is the link capacity, i.e., our scheme can fully-utilize the link capacity. On the other hand, W_a of the non-adaptive scheme is saturated at around 88 BUs. The reason for this is because up to 8-BU bandwidth is reserved for hand-offs (i.e., $W_r \approx 8$ in a heavily-loaded region) on average with this scheme, and only the unreserved bandwidth can be utilized. We observe that W_r's of both schemes are about the same since our adaptive scheme again uses $b_{min}(\cdot)/r_{min}$ for the bandwidth reservation purpose. Note that W_r is adaptively determined based on the estimated user mobility and observed hand-off dropping events in each cell so as to keep P_{HD} below the target value, i.e., 0.01 in our study, according to [1].

Another interesting observation is that W_a of our scheme decreases slowly after peaking at around the offered load of 40. This is because W_r increases after passing the offered load of 40, and the reserved bandwidth cannot be utilized. However, we also observe that the sum of W_a and W_r is larger than the link capacity of 100 BUs with our scheme, e.g., at the offered load of 150, $W_a + W_r \approx 104$. The reason for this is that the reserved bandwidth $W_{r,i}$ in cell i affects only new connection-admission decisions and upgrade processes while the whole link capacity can be utilized after a downgrade process. By comparing the aggregate revenue R of the two schemes, we found that R of our scheme is limited by a larger value, and also increases monotonically till a larger offered load is reached. Since our revenue function has decreasing marginal revenues for an extra unit of actual throughput, the monotonic increase of R as the offered load increases implies that more connections are admitted. One can expect that for the offered load larger than 75, most connections are provided 1 BU which corresponds to one unit of revenue, then the difference of R and W_a comes from the minimum channel usage efficiency r_{min}, which is 0.8 for our simulation study, i.e., $W_a \cdot 0.8 \approx R$.

V. Concluding Remarks

We have proposed an architecture for wireless bandwidth allocation and management in a highly dynamic environment with user mobility and time-varying channel conditions. Each connection is specified by a set of an acceptable actual throughput range, a revenue function, and adaptation constraints defining how often and how much bandwidth adaptation can be made. The BS allocates bandwidth to each connection so as to maximize the aggregate revenue while attempting to meet the adaptation constraints. An adaptation action can be triggered due to a connection arrival/departure or a channel condition change. Our scheme also penalizes connections which initiate frequent downgrade/upgrade adaptations in order to reduce link bandwidth fluctuations. Using simulations, we demonstrated how the proposed scheme works, and showed its advantages over a non-adaptive scheme.

References

[1] S. Choi and K. G. Shin, "Predictive and adaptive bandwidth reservation for hand-offs in QoS-sensitive cellular networks," in *Proc. ACM SIGCOMM'98*, pp. 155–166, Sep. 1998.

[2] IEEE 802.11 WG, *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications*, 1999.

[3] R. R.-F. Liao and A. Campbell, "On programmable universal mobile channels in a cellular internet," in *Proc. ACM/IEEE MobiCom'98*, pp. 191–202, Oct. 1998.

[4] S. Lu, K.-W. Lee, and V. Bharghavan, "Adaptive service in mobile computing environments," in *Proc. IWQoS'97*, May 1997.

[5] M. Naghshineh and M. Willebeek-LeMair, "End-to-end QoS provisioning in multimedia wireless/mobile networks using an adaptive framework," *IEEE Communications Magazine*, pp. 72–81, Nov. 1997.

[6] D. Qiao and K. G. Shin, "A two-step adaptive error recovery scheme for video transmission over wireless networks," in *Proc. IEEE INFOCOM'00*, March 2000.

[7] A. K. Talukdar, B. R. Badrinath, and A. Acharya, "Rate adaptation schemes in networks with mobile hosts," in *Proc. ACM/IEEE MobiCom'98*, pp. 169–180, Oct. 1998.