An Integrated Bandwidth Adaptation Scheme for Multimedia Wireless Networks and its Connection-Level Performance Analysis

Ning Lu and John Bigham

Original scientific paper

Abstract— **This paper presents an integrated bandwidth adaptation scheme for multimedia wireless networks using application utility functions. With the proposed scheme, each call in the network is assigned a utility function according to its adaptive characteristics. Depending on the network load the allocated bandwidth of ongoing calls are upgraded or degraded dynamically so that the achieved utility of the network is maximized. Appropriate call admission control and bandwidth reservation policies are also incorporated into the scheme to provide QoS guarantees to the new and handoff calls, respectively. Extensive simulation experiments have been conducted to evaluate the connection-level performance of the proposed scheme. Results show that our bandwidth adaptation scheme is effective in both increasing the utility and bandwidth utilization of wireless networks while keeping the call blocking and handoff dropping probabilities substantially low.**

Index Terms: **bandwidth adaptation, connection-level performance, multimedia wireless networks, QoS, utility function.**

I. INTRODUCTION

In recent years, there has been an explosive demand for multi-class traffic (voice, video, and data), especially bandwidth-intensive multimedia traffic in wireless networks. Different classes of traffic over networks require various amounts of bandwidth and it is important to provide QoS guarantees to the multi-class traffic according to their bandwidth needs. QoS provisioning has been extensively studied for wireline networks; however, due to the diverse application QoS requirements and the scarcity of wireless link bandwidth, the QoS issues in wireless networks are much more challenging than their wired counterpart. Thus efficient resource management becomes a key factor in enhancing the system performance of wireless networks.

Bandwidth adaptation is one of the most promising methods to provide QoS guarantees in wireless networks. In the traditional non-adaptive network environment, once a call is admitted its allocated bandwidth is fixed throughout its lifetime; when a new or handoff call requests a certain amount of bandwidth, the network rejects the call if there is no sufficient bandwidth available. However with adaptive bandwidth allocation, when a new or handoff call arrives to a congested network the allocated bandwidth of ongoing calls can be degraded to smaller values to accept the new or handoff call; thereby reducing call blocking and handoff dropping probabilities. On the other hand, when an ongoing call is terminated due to its completion or outgoing handoff, the released bandwidth can be used to increase the network bandwidth utilization by upgrading other ongoing calls which have not received their maximum bandwidth requirements.

A. Related Work

The concept of bandwidth adaptation was originally introduced in wired networks to overcome the problem of network congestion. More recently, some adaptive bandwidth management and QoS provisioning schemes for wireless networks have been reported in the literature [1, 3 – 6, 8 – 10]. Oliveira et al. [9] propose an admission control scheme based on adaptive bandwidth reservation to provide QoS guarantees for multimedia traffic in high-speed wireless cellular networks. The proposed scheme allocates bandwidth to a call in the cell where there is a new or handoff request and reserves bandwidth dynamically in all neighbouring cells according to the network conditions. Bandwidth reservation in all neighbouring cells guarantees QoS; however, it often results in underutilization of resources as the mobile user hands off to only one of the cells. Moreover the allocated bandwidth of the call is fixed during the stay in a cell and it can be changed only at a handoff. In the work of El-Kadi et al. [5], an effective rate-based borrowing scheme (RBBS) is provided for multimedia wireless networks. In the case of insufficient bandwidth, in order not to deny service to requesting calls, bandwidth can be borrowed on a temporary basis from existing calls to accept the new or handoff call. Although the scheme is adaptive, it does not include a quantitative measure (e.g. QoS satisfaction or revenue curve) to reflect the importance of different calls. In [1], Bharghavan et al. present the TIMELY adaptive resource management architecture and algorithms for resource reservation and

Manuscript received February 02, 2006 and revised May 31, 2006. This research was supported in part by the Office of Naval Research, USA. Contract number: N00014-03-1-0323.

N. Lu is with the MPI-QMUL Information Systems Research Centre, Macao Polytechnic Institute, Macao, China and the Department of Electronic Engineering, Queen Mary, University of London, UK (e-mail: ning.lu@elec.qmul.ac.uk).

J. Bigham is with the Department of Electronic Engineering, Queen Mary, University of London, UK.

adaptation in mobile computing environments. The architecture has four layers – link, reservation, adaptation and transport – all of which perform resource adaptation in a coordinated manner to solve the problems introduced by scarce and dynamic network resources. A revenue model for resource usage is introduced and a weighted version of the max-min fair rate adaptation algorithm is proposed to distribute resources among the adaptable flows in order to maximize revenue increase. However, the multi-layer feature has made the bandwidth adaptation work at the expense of high message overhead. Curescu et al. [4] introduces a utility-maximization bandwidth allocation scheme. Interestingly, the scheme takes into account the age of the connections to reflect the sensitivity of the connections to the bandwidth re-allocation by assuming that the duration of every connection can be estimated. However, the feasibility of estimating connection duration is doubtful in real-time wireless network environments since the connection duration is updated dynamically with bandwidth re-allocation. Another bandwidth adaptation scheme using quantitative QoS measure can be found in [10]. A revenue-based model is used to describe the bandwidth degradation process for both real-time and non-real-time traffic and an effective algorithm has been proposed to maximize the total network revenue. The scheme gives handoff calls higher priorities over new calls to provide them QoS guarantees. However, it does not reserve bandwidth for real-time handoff calls thus it risks an inability of meeting their QoS requirements when the network is heavily overloaded since real-time connections are sensitive to delays and the lack of bandwidth during handoff can cause them to be dropped.

B. Our Contributions

In this paper, we present an integrated bandwidth adaptation scheme for multimedia wireless networks using application utility functions. The objective of our scheme is to strike the balance among multiple connection-level QoS requirements of multimedia wireless networks under the constraints of limited and varying bandwidth resources. Within the proposed scheme, a call admission control policy is provided for the new calls, a bandwidth adaptation algorithm is proposed for the connected calls to maximize the network utility, and a bandwidth reservation mechanism is applied for real-time handoff calls to prevent them from being dropped.

Moreover, even though the revenue/utility function has been applied for the bandwidth allocation problems in [1, 3, 4, 6, 8, 10], none of these schemes provide explicit formulation of utility functions to capture the adaptive nature of the application. For example, reference [10] adopts the Sigmoid functions, reference [6] uses the linear and convex functions and reference [4] assigns utility functions using subjective values from the authors' experiments. This paper classifies the traffic into different classes and formulates the utility function for each traffic class according to their adaptive characteristics.

The rest of the paper is organized as follows. Section II describes the utility-based adaptive traffic model including the concept of utility function and the characteristics of the traffic classes used in our study. In Section III we give a detailed description and formulation of the bandwidth adaptation problem and propose a utility-maximization bandwidth adaptation algorithm. Section IV introduces the new call admission and handoff call management mechanisms. The simulation model and numerical results are presented in Section V. Finally Section VI gives the concluding remarks.

II. UTILITY-BASED ADAPTIVE TRAFFIC MODELING

A. The Concept of Utility Function

Utility was originally used in economics and has been brought into networking research in recent years [3, 4, 8, 11, 12]. It represents the "level of satisfaction" of a user or the performance of an application. A utility function here is a curve mapping bandwidth received by the application to its performance as perceived by the user. It is monotonically non-decreasing; in other words, more bandwidth allocation should not lead to degraded application performance. The key advantage of the utility function is that it can inherently reflect the user's QoS requirements and quantify the adaptability of the application. The shape of the utility function varies according to the adaptive characteristics of the application.

B. Traffic Classes and Their Utility Functions

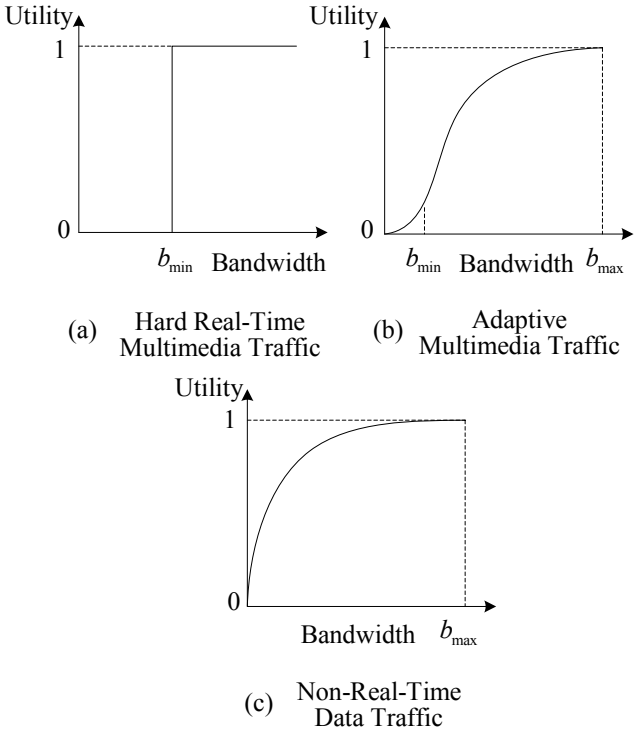

Fig. 1. Multi-class traffic utility functions

In this paper, the traffic offered to the network is assumed to belong to two classes:

- Class I – real-time multimedia traffic, and
- Class II – non-real-time data traffic.

Class I traffic can be further classified into two subclasses – hard real-time multimedia traffic and adaptive multimedia traffic.

Hard real-time multimedia traffic refers to the multimedia applications with stringent bandwidth requirements. A call belonging to this traffic class requires strict end-to-end performance guarantees and does not show any adaptive properties. It is not allowed to enter the network if the minimum bandwidth requirement cannot be met. Once accepted, the bandwidth is fixed during the call's lifetime; more bandwidth allocation will not lead to performance enhancement while any bandwidth degradation will cause the QoS (utility) to drop to zero. Examples include audio/video phone, video conference and telemedicine [9, 11]. We use the following utility function to model hard real-time multimedia traffic:

$$u(b) = \begin{cases} 1, \text{ when } b \geq b_{\min} \\ 0, \text{ when } b < b_{\min} \end{cases}$$

where $b_{\min}$ is the minimum bandwidth requirement. The general shape of the hard real-time multimedia traffic utility function is depicted in Fig. 1 (a).

Adaptive multimedia traffic refers to multimedia applications that can adapt to various network loads. In case of congestion, they can gracefully adjust their transmission rates such that the QoS received by end-users is still acceptable. However, this type of traffic requires the network to provide a minimum level of performance guarantees; if the bandwidth is reduced below some threshold the quality becomes unacceptable. Typical examples are interactive multimedia services and video on demand [9, 11]. We use the following utility function to model adaptive multimedia traffic:

$$u(b) = 1 - e^{-\frac{k_1 b^2}{k_2 + b}}$$

where k_1 and k_2 are tunable parameters which determine the shape of the utility function and ensure that when the maximum requested bandwidth is received, $u \approx 1$. The similar utility function has also been used to model adaptive multimedia traffic in [2, 11]. The utility function of adaptive multimedia traffic takes the shape like Fig. 1 (b).

Class II traffic refers to the data applications which are rather tolerant of delays. In case of congestion, it is acceptable to buffer non-real-time data at a network node (such as a base station) and transmit them at a slower rate. For Class II traffic, it is assumed that there is no minimum required bandwidth since it can tolerate relatively large delays. Examples are email, paging/fax, file transfer/retrieval, and remote login [9, 11]. We use the following utility function to model Class II traffic [2, 11]:

$$u(b) = 1 - e^{-\frac{kb}{b_{max}}}$$

where k is a tunable parameter which determines the shape of the utility function and ensures that when the maximum requested bandwidth is received, $u \approx 1$. The utility function of non-real-time data traffic is shown in Fig. 1 (c).

C. Adaptive Traffic Model

Adaptive multimedia can be classified into discrete adaptive multimedia and continuous adaptive multimedia. In discrete adaptive multimedia, the bandwidth takes on a set of discrete values in approaches such as layered coding (or hierarchical coding) [16, 17], whereas in continuous adaptive multimedia, the bandwidth takes on continuous values [18, 19]. Discrete adaptive multimedia is a subset of continuous adaptive multimedia; therefore in this paper we only consider continuous adaptive multimedia. Our multimedia wireless network model consists of real-time multimedia traffic and non-real-time data traffic. Each call in the network is assigned a utility function according to the adaptive characteristics of its traffic class. To keep the traffic model simple, we quantize each utility function into linear piece-wise segments by dividing its utility range into a fixed number of equal intervals. Assume that the utility function of the i-th ongoing call is $u_i(b_i)$, after quantization it becomes a linear piece-wise function represented by a set of $<$bandwidth, utility$>$ points:

$$u_i(b_i) = (< b_{i,1}, u_{i,1} >, < b_{i,2}, u_{i,2} >, \ldots, < b_{i,K_i}, u_{i,K_i} >)$$

where K_i is the maximum $<$bandwidth, utility$>$ level and $u_i(b_{i,j}) = u_{i,j}$ is the achieved utility for the allocated bandwidth $b_{i,j} \in (b_{i,1}, b_{i,2}, \ldots, b_{i,K_i})$ $(1 \leq j \leq K_i)$.

With adaptive bandwidth allocation paradigm, the allocated bandwidth of an ongoing call can be dynamically changed in the set of $(b_{i,1}, b_{i,2}, \ldots, b_{i,K_i})$ during its lifetime depending on the network load. If the network is under-loaded, every call will be allocated its maximum bandwidth b_{i,K_i}; otherwise, depending on how much the network is overloaded, one or more calls will be allocated a bandwidth less than b_{i,K_i}. Note that although for hard real-time multimedia traffic their allocated bandwidth cannot be changed we can regard them as a special case of adaptive traffic.

III. BANDWIDTH ADAPTATION

A. Problem Formulation

In our wireless network model, bandwidth adaptation is performed based on each individual cell with fixed bandwidth capacity B. It can be decomposed into two processes – bandwidth upgrades and bandwidth degrades. Bandwidth upgrades are triggered by call departure events and bandwidth degrades are triggered by call arrival events. Call departure events include call completion events (a call within the cell terminates) and outgoing handoff events (a call leaves its current cell). Call arrival events include new call arrival events (a new call is generated within the cell) and incoming handoff events (a handoff call arrives to the cell).

A.1 Bandwidth Upgrades

Assume that in an overloaded cell, when a call is terminated or handed off to another cell, there are n ongoing calls that have not received their maximum bandwidth. The released bandwidth (denoted by β) can be utilized to upgrade these ongoing calls. Denote the utility function of the i-th ongoing call as $u_i(b_i)$ $(1 \le i \le n)$ and its current allocated bandwidth as β_i, thus the upgradable utility function of the i-th ongoing call can be written as $u_i^{\uparrow}(b_i^{\uparrow}) = u_i(\beta_i + b_i^{\uparrow})$ $(0 \le b_i^{\uparrow} \le b_{i,K_i} - \beta_i)$ where b_{i,K_i} is the maximum bandwidth requirement. The objective of bandwidth upgrades is to find the bandwidth upgrades profile $\{b_i^{\uparrow}\}$ for all ongoing calls to maximize the sum of their utilities subject to bandwidth constraints, i.e.

$$\text{maximize: } \sum_{i=1}^{n} u_i^{\uparrow}(b_i^{\uparrow})$$

$$\text{subject to: } \sum_{i=1}^{n} b_i^{\uparrow} \le \beta$$

$$\text{and } 0 \le b_i^{\uparrow} \le b_{i,K_i} - \beta_i$$

A.2 Bandwidth Degrades

Consider a saturated cell with n ongoing calls, when a new or handoff call arrives the bandwidth of ongoing calls can be degraded to smaller values to accommodate the new or handoff call. Denote the utility function of the i-th ongoing call as $u_i(b_i)$ $(1 \le i \le n)$ and its current allocated bandwidth as β_i, thus the degradable utility function of the i-th ongoing call can be written as $u_i^{\downarrow}(b_i^{\downarrow}) = u_i(\beta_i - b_i^{\downarrow})$ $(0 \le b_i^{\downarrow} \le \beta_i)$; also denote the utility function of the new or handoff call as $u_{n+1}(b_{n+1})$. The objective of bandwidth degrades is to find the bandwidth degrades profile $\{b_i^{\downarrow}\}$ for ongoing calls and the allocated bandwidth b_{n+1} for the new or handoff call to maximize the utility sum of all calls subject to bandwidth constraints, i.e.

$$\text{maximize: } \left(\sum_{i=1}^{n} u_i^{\downarrow}(b_i^{\downarrow}) \right) + u_{n+1}(b_{n+1})$$

$$\text{subject to: } \left(\sum_{i=1}^{n} (\beta_i - b_i^{\downarrow}) \right) + b_{n+1} \le B$$

$$\text{and } 0 \le b_i^{\downarrow} \le \beta_i$$

B. Our Proposed Utility-Maximization Algorithm

The objective of bandwidth upgrades and bandwidth degrades is to maximize the utility sum of n and $n+1$ utility functions respectively subject to bandwidth constraints. Without loss of generality, we propose an algorithm to maximize the total utility of n utility functions. Finding optimal solutions for such a problem is NP-hard and has exponential time complexity [7]. Since bandwidth adaptation should be performed in real time to support the frequent resource fluctuations in wireless networks, we design an efficient search-tree based algorithm to obtain near-optimal solutions.

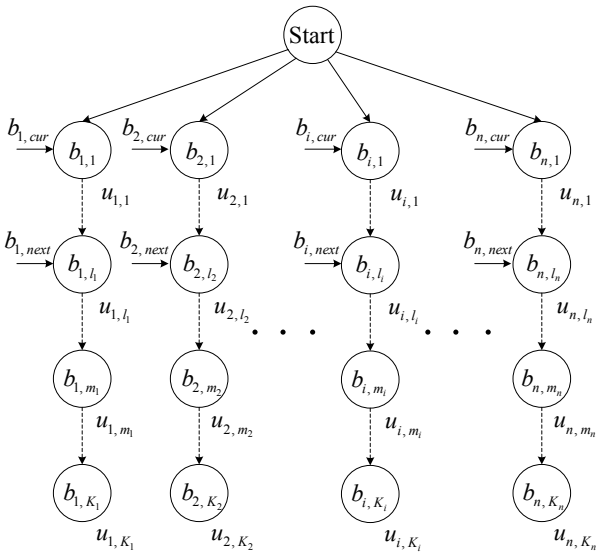

Fig. 2. Bandwidth allocation tree

TABLE I
NOTATIONS FOR THE UTILITY-MAXIMIZATION ALGORITHM

$b_{i,cur}$	the current allocated bandwidth of the i-th call
$u_{i,cur}$	the current achieved utility of the i-th call
$b_{i,next}$	the next possible allocated bandwidth of the i-th call
$u_{i,next}$	the next possible achieved utility of the i-th call
$b_{i,temp}$	the temporary allocated bandwidth of the i-th call
$u_{i,temp}$	the temporary achieved utility of the i-th call
$b_{i,req}$	the required bandwidth for upgrading the current allocated bandwidth of the i-th call to its next bandwidth level
$b_{req,\max}$	the maximum required bandwidth for upgrading the current allocated bandwidth of the call to its next bandwidth level among all n ongoing calls, i.e. $b_{req,\max} = \max\{b_{i,req}\}$ $(1 \le i \le n)$
$r_{i,cur}$	the current utility generation ratio of the i-th call
$r_{i,temp}$	the temporary utility generation ratio of the i-th call
$r_{i,\max}$	the maximum utility generation ratio of the i-th call
B	the total available bandwidth to be allocated
b_{avail}	the current available bandwidth to be allocated

The algorithm can be formulated by a tree as illustrated in Fig. 2. The linear piece-wise utility function of each call is represented by a branch in the tree and the <bandwidth, utility> points of each utility function are represented by the nodes in the branch. The nodes of each branch are laid out downwards to reflect the bandwidth allocation order i.e. the second <bandwidth, utility> point of a utility function is connected to the first <bandwidth, utility> point and so on. The tree contains n branches and the number of branches is the same as the number of utility functions. Each branch (utility function) is associated with ten variables $b_{i,cur}$, $u_{i,cur}$, $b_{i,next}$, $u_{i,next}$, $b_{i,temp}$, $u_{i,temp}$, $b_{i,req}$, $r_{i,cur}$, $r_{i,temp}$ and $r_{i,\max}$ (all notations used can be found in Table I). The algorithm allocates the bandwidth in a greedy fashion based on the concept of utility generation ratio, i.e. giving priority to the <bandwidth, utility> points with higher utility generation ratio. The utility generation ratio r is calculated by dividing the utility increase by bandwidth increase between two <bandwidth, utility> points. The pseudo-code of the algorithm is as follows:

(1) $b_{avail} = B$
for each call-i
initialize $b_{i,cur}$, $u_{i,cur}$, $b_{i,next}$, $u_{i,next}$, $b_{i,temp}$ and $u_{i,temp}$ to be at the first level
$r_{i,temp} = 0$
$r_{i,\max} = 0$
$b_{i,req} = b_{i,cur+1} - b_{i,cur}$
$r_{i,cur} = (u_{i,cur+1} - u_{i,cur}) / b_{i,req}$

(2) for each call-i
while $(b_{i,temp} < b_{i,K_i})$
increase $b_{i,temp}$ by one level
$r_{i,temp} = (u_{i,temp} - u_{i,cur}) / (b_{i,temp} - b_{i,cur})$
if $(r_{i,temp} > r_{i,\max})$
$b_{i,next} = b_{i,temp}$
$r_{i,\max} = r_{i,temp}$

(3) among all calls find the largest $b_{i,req}$ denoted by $b_{req,\max}$
if $(b_{req,\max} >= b_{avail})$
among all calls with $b_{i,req} >= b_{avail}$ find the call with the highest $r_{i,cur}$ denoted by call-k
$b_{k,cur} = b_{k,cur} + b_{avail}$
return $b_{i,cur}$ for each call

(4) among all calls find the call with the largest $r_{i,\max}$ denoted by call-j
if ($b_{avail} \geq (b_{j,next} - b_{j,cur})$)
$b_{avail} = b_{avail} - (b_{j,next} - b_{j,cur})$
$b_{j,cur} = b_{j,next}$
$b_{j,temp} = b_{j,next}$
$r_{j,temp} = 0$
$r_{j,\max} = 0$
$b_{j,req} = b_{j,cur+1} - b_{j,cur}$
$r_{j,cur} = (u_{j,cur+1} - u_{j,cur}) / b_{j,req}$
else
$b_{j,cur} = b_{j,cur} + b_{avail}$
return $b_{i,cur}$ for each call

(5) for call-j found in Step (4)
while $(b_{j,temp} < b_{j,K_j})$
increase $b_{j,temp}$ by one level
$r_{j,temp} = (u_{j,temp} - u_{j,cur}) / (b_{j,temp} - b_{j,cur})$
if $(r_{j,temp} > r_{j,\max})$
$b_{j,next} = b_{j,temp}$
$r_{j,\max} = r_{j,temp}$

(6) Go to Step (3)

In Step (1), the algorithm initializes the associated variables and calculates the current utility generation ratio $r_{i,cur}$ for each branch.

In Step (2), for each branch the algorithm increases $b_{i,temp}$ by one level and calculates the utility generation ratio $r_{i,temp}$. If $r_{i,temp}$ is greater than $r_{i,\max}$ the algorithm replaces $r_{i,\max}$ with $r_{i,temp}$ and upgrades $b_{i,next}$ to $b_{i,temp}$. The above process is repeated until every node of the branch has been investigated and the node with the maximum utility generation ratio is the next possible bandwidth allocation node in its corresponding branch.

Step (3) checks if the available bandwidth b_{avail} is not more than $b_{req,\max}$. If this is the case, b_{avail} is allocated to the call with the highest current utility generation ratio among all calls with $b_{i,req} \geq b_{avail}$.

The algorithm keeps track of $r_{i,\max}$ for each branch. In Step (4) it chooses the branch with the highest $r_{i,\max}$ and upgrades its current and temporary allocated bandwidth to its next possible bandwidth allocation if there is enough bandwidth available.

Subsequently, for the selected branch, it has the same current and next possible bandwidth allocation. Therefore, in Step (5) the algorithm updates the next possible bandwidth allocation for this branch using the approach stated in Step (2). After finding the next possible bandwidth allocation for this branch, the algorithm goes back to Step (3) to execute the above procedure repeatedly until there is no sufficient bandwidth available (we assume that the total available bandwidth B cannot satisfy the maximum bandwidth requirements of all calls).

IV. New Call Admission And Handoff Call Management

Having introduced the utility-maximization bandwidth adaptation algorithm, this section describes the new call admission and handoff call management mechanisms.

A. New Call Admission

The objective of new call admission is to provide QoS guarantees for the new calls while efficiently utilizing network resources. Specifically, a call admission algorithm decides whether a new call can be admitted into the network with its satisfied QoS while not violating the QoS constraints of the existing calls. The decision is made based on the requested bandwidth of the new call, the availability of network resources as well as the current bandwidth allocation of the existing calls.

With our new call admission policy when a new call requests admission into the network, the cell first attempts to allocate the desired amount of bandwidth to the new call. If the desired amount of bandwidth is not available, the bandwidth adaptation algorithm is invoked to free some bandwidth from the existing ongoing calls. After bandwidth adaptation if the sum of the available bandwidth in the cell plus the freed bandwidth according to the bandwidth adaptation algorithm can satisfy the desired bandwidth of the new call, the call is admitted; otherwise, the call is blocked.

B. Handoff Call Management

To protect real-time handoff calls from being dropped we have applied bandwidth reservation in our scheme. However, bandwidth reservation should be carefully used due to the fact that it can decrease the bandwidth utilization of the networks. Our handoff management policies differentiate between Class I and Class II traffic. A certain amount of bandwidth is reserved exclusively for Class I traffic handoff calls; the motivation is that real-time traffic would suffer an actual loss by being dropped. For a Class I handoff call, the proposed scheme works as follows. If the sum of the available free and reserved bandwidth in the cell that the handoff call is moving into is greater than or equal to the minimum required bandwidth of the call, the handoff call is accepted. If not, the bandwidth adaptation algorithm is performed to free more bandwidth; after bandwidth adaptation if the total available bandwidth can satisfy the minimum bandwidth requirement the handoff call is accepted; otherwise, it is rejected.

The reserved bandwidth is not available to Class II traffic because it is assumed that a Class II call, although inconvenienced by being dropped, would be able to resume its transmission at a later time without any significant loss due to its elastic characteristics. For a Class II handoff call, it is accepted as long as there is some bandwidth available in the cell that the call is moving into. It will only be dropped when there is no bandwidth available in the cell at all. Bandwidth degradation of a Class II handoff call results in a slower transmission rate and, thus a longer transmission delay. However, Class II is for non-real-time data calls, and the impact of increased delay should not be significant [9].

V. Simulation Experiments

A. Simulation Model

A 6×6 wrap-around cellular network model consisting of 36 cells has been built to evaluate the performance of our proposed bandwidth adaptation scheme. Each cell (base station) has a bandwidth capacity of 30 Mbps with 5% bandwidth reserved for Class I handoff calls; the diameter of the cell is 1 km. New call arrivals are assumed to follow a Poisson distribution with mean rate λ and the call holding time is assumed to follow an exponential distribution with mean $1/\mu$. Mobile terminals can travel in one of six directions with equal probability and their speed is uniformly distributed between 10 and 60 miles per hour.

TABLE II
Traffic Characteristics for Our Simulation

Applic Group	Traffic Class	Bandwidth Requirement (Desired Bandwidth)	Average Connection Duration	Example	Utility Function (b is Mbps)
1	I	30 Kbps (30 Kbps)	3 minutes	Voice Service & Audio-phone	$\begin{cases}1, b \ge 0.03\\ 0, b < 0.03\end{cases}$
2	I	256 Kbps (256 Kbps)	5 minutes	Video-phone & Video-conference	$\begin{cases}1, b \ge 0.25\\ 0, b < 0.25\end{cases}$
3	I	1 - 6 Mbps (3 Mbps)	10 minutes	Interact. Multimedia & Video on Demand	$1-e^{-\frac{1.045b^2}{2.166+b}}$
4	II	0 - 20 Kbps (10 kbps)	30 seconds	E-mail, Paging & Fax	$1-e^{-\frac{4.6b}{0.02}}$
5	II	0- 512 Kbps (256 kbps)	3 minutes	Remote Login & Data on Demand	$1-e^{-\frac{4.6b}{0.5}}$
6	II	0 - 10 Mbps (5 Mbps)	2 minutes	File Transfer & Retrieval Service	$1-e^{-\frac{4.6b}{10}}$

Six representative groups of traffic belonging to the two traffic classes introduced in Section II are carefully chosen for the simulation. They are typical traffic seen in wireless networks and have been widely used by others [4, 5, 9]. Each traffic group is associated with an appropriate utility function and calls belonging to the same group are assumed to have the same utility function. The details about how to choose utility functions and calculate their tunable parameters for different groups of calls can be found in [11]. Table II shows the exact characteristics of the traffic and all six groups of traffic are generated with equal probability.

To highlight the performance of our proposed scheme, we compare it with a non-adaptive scheme and a well-known adaptive scheme RBBS [5]. The non-adaptive scheme is simulated assuming that a call must be allocated its maximum bandwidth to be admitted and once accepted its bandwidth cannot be changed throughout the lifetime. With RBBS when there is a new or handoff call request under network congestion, the bandwidth can be borrowed from ongoing calls on a temporary basis to accept the new or handoff call.

B. Numerical Results

In the experiments, apart from the traditional connection-level performance metrics i.e. call blocking and handoff dropping probabilities, we introduce a new performance metric called average cell utility (cell utility means the utility sum of all ongoing calls in a given cell). Average cell utility is calculated as follows: every time the bandwidth adaptation occurs, the achieved cell utility is re-calculated and added to the total accumulated cell utility. At the end of the simulation the average cell utility is calculated by dividing the total accumulated cell utility by the bandwidth adaptation frequency.

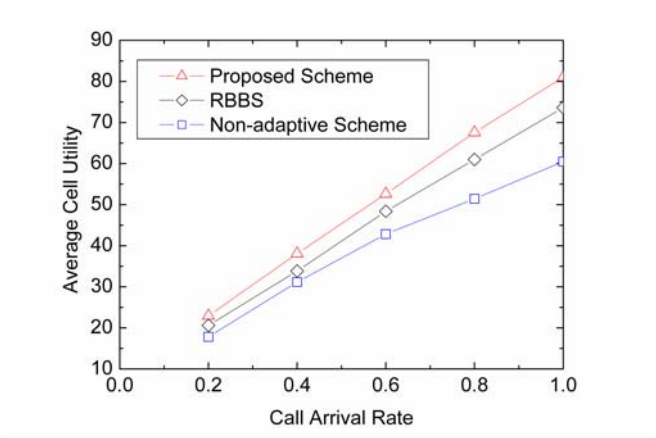

Fig. 3. Average cell utility

Fig. 3 compares the average cell utility of the three schemes as a function of the call arrival rate. It can be seen that our scheme achieves higher utility than the other two schemes since it works in the fashion to maximize the utility sum of all ongoing calls in each individual cell. The trend becomes more evident when the call arrival rate increases. For example, at the call arrival rate of 0.2 (calls/sec/cell), the average cell utility of our scheme is slightly higher than that of RBBS and about 28% higher than that of the non-adaptive scheme; when the call arrival rate increases to 1.0, the average cell utility of our scheme has become about 10% higher than that of RBBS and 34% higher than that of the non-adaptive scheme.

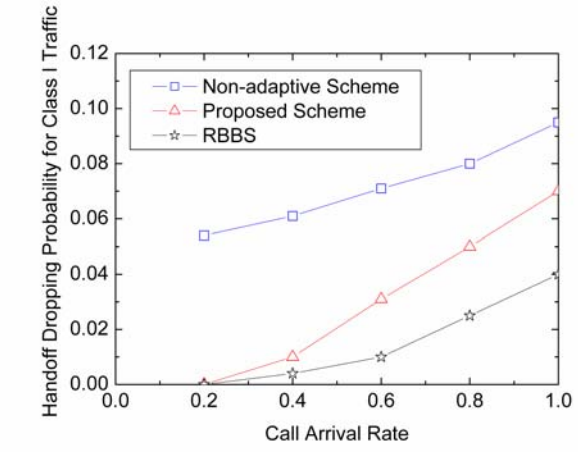

Fig. 4. Handoff dropping probability for Class I traffic

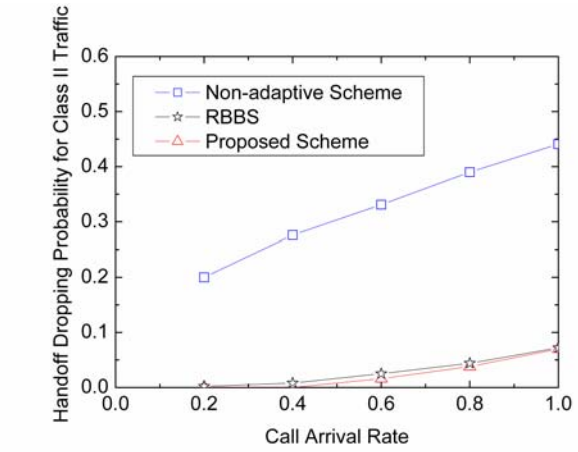

Fig. 5. Handoff dropping probability for Class II traffic

Figs. 4 and 5 illustrate the handoff dropping probabilities for Class I and Class II traffic, respectively. For Class I handoff calls the non-adaptive scheme has the poorest performance due to the fact that it does not utilize the bandwidth flexibility of ongoing calls to free bandwidth to accept the handoff calls under network congestion. Compared with the non-adaptive scheme, the Class I traffic dropping probabilities of the proposed scheme and RBBS are both low (RBBS slightly outperforms our scheme) because both schemes not only degrade ongoing calls to free bandwidth when the network is overloaded but also give Class I traffic handoff calls exclusive use of the reserved bandwidth to protect them from the actual loss by being dropped. In terms of Class II traffic dropping probabilities, our scheme has the best performance; even though the reserved bandwidth is not available to the handoff calls of Class II traffic, the dropping ratio of our proposed scheme is still substantially low because their handoff calls do not have minimum bandwidth requirements and can be accepted as long as there is some free

bandwidth available in the cell. The non-adaptive scheme features an extremely high dropping probability because with non-adaptive scheme Class II calls are assumed to be non-adaptive and their minimum bandwidth requirements are equal to their maximum bandwidth requirements. Thus Class II handoff calls are more likely to be dropped without accessing the reserved bandwidth pool.

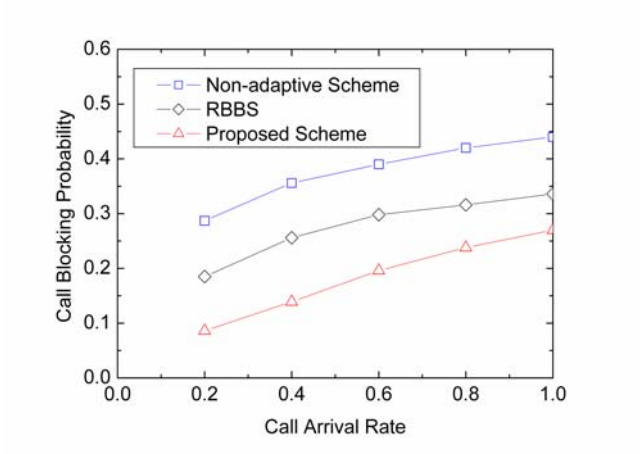

Fig. 6. Call blocking probability

Fig. 6 compares the call blocking probabilities of the three schemes. The results show how our scheme allows a significant improvement in the blocking ratio while keeping the handoff dropping probability low. The proposed scheme outperforms RBBS partially because with RBBS every time the bandwidth adaptation happens only one share (a share is a pre-defined parameter) of the adaptive bandwidth can be borrowed from each ongoing call while with the proposed scheme the allocated bandwidth of ongoing calls can be degraded down to the minimum level in one step to accommodate the new calls.

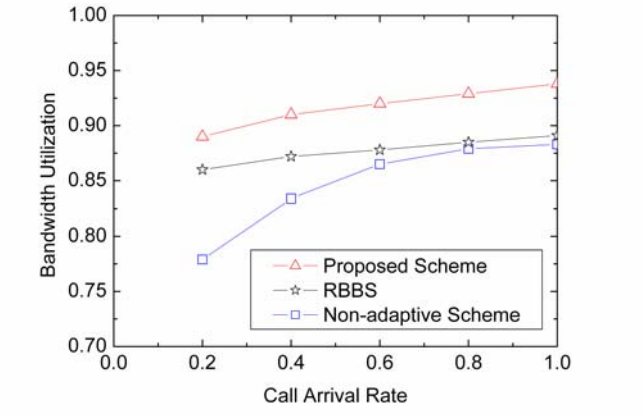

Fig. 7. Bandwidth utilization

Fig. 7 demonstrates the values of bandwidth utilization for the three schemes. It can be observed that when the call arrival rate increases, the bandwidth utilization value becomes higher for all schemes. At the maximum call arrival rate 1.0, the bandwidth utilization of the proposed scheme comes close to equalling the bandwidth outside of the reserved pool. The results for the non-adaptive scheme are worse than for the other two because the non-adaptive scheme cannot take advantage of the adaptive feature of the calls to utilize more network bandwidth.

VI. CONCLUDING REMARKS

In this paper we have proposed an integrated bandwidth adaptation scheme for multimedia wireless networks. When the network is overload, the bandwidth adaptation can be performed to free bandwidth from existing ongoing calls to accept the new and handoff calls. We classify the traffic into different classes and assign a utility function to each call depending on the adaptive characteristics of its traffic class. Our main contribution is a novel, search-tree based bandwidth adaptation algorithm which aims to maximize the total utility of the network. To achieve better overall connection-level performance, a call admission control policy is also deployed to provide QoS guarantees to the new calls. Moreover, a fixed percentage of bandwidth in each cell is reserved exclusively for real-time handoff calls to prevent them from being dropped since they are rather sensitive to delays. Simulation experiments have been conducted to evaluate the performance of the proposed scheme by comparing it with a non-adaptive scheme and RBBS. The results have demonstrated the excellent overall connection-level performance of our proposed scheme.

REFERENCES

[1] V. Bharghavan, K. W. Lee, S. W. Lu, S. W. Ha, J. R. Li, and D. Dwyer, "The Timely Adaptive Resource Management Architecture", IEEE Personal Communications Magazine, 5(4): 20-31, Aug. 1998.

[2] L. Breslau and S. Shenker, "Best-effort Versus Reservations: A Simple Comparative Analysis", ACM Computer Communications Review, 28: 131-143, Sept. 1998.

[3] Y. Cao and V. O. K. Li, "Utility-oriented Adaptive QoS and Bandwidth Allocation in Wireless Network", IEEE International Conference on Communications (ICC), 2002, vol. 5, pp. 3071-3075.

[4] C. Curescu and S. Nadjm-Tehrani, "Time-Aware Utility-based Resource Allocation in Wireless Networks", IEEE Transactions on Parallel and Distributed Systems, 16 (7): 624 - 636, July 2005.

[5] M. El-Kadi, S. Olariu, and H. Abdel-Wahab, "A Rate-based Borrowing Scheme for QoS Provisioning in Multimedia Wireless Networks", IEEE Transactions on Parallel and Distributed Systems, 13(2):156–167, Feb. 2002.

[6] T. Kwon, Y. Choi, and S. K. Das, "Bandwidth Adaptation Algorithms for Adaptive Multimedia Services in Mobile Cellular Networks", Wireless Personal Communications, vol. 22, no. 3, pp.337-357, Sept. 2002.

[7] C. Lee, J. P. Lehoczky, R. Rajkumar and D. P. Siewiorek, "On Quality of Service Optimization with Discrete QoS Options", Proceedings of the IEEE Real-Time Technology and Applications Symposium, June 1999

[8] R. Liao and A. T. Campbell, "A Utility-based Approach for Quantitative Adaptation in Wireless Packet Networks", Wireless Networks, 7(5): 541- 557, 2001.

[9] C. Oliviera, J. Kim, and T. Suda, "An Adaptive Bandwidth Reservation Scheme for High Speed Multimedia Wireless

Networks," IEEE Journal on Selected Areas in Commmunications, vol. 16, pp. 858-874, Aug. 1998.
[10] S. K. Das, S. K. Sen, K. Basu, and H. Lin, "A Framework for Bandwidth Degradation and Call Admission Control Schemes for Multiclass Traffic in Next-Generation Wireless Networks", IEEE Journal on Selected Areas in Commmunications, 21: (10), pp. 1790 - 1802, Dec. 2003.
[11] V. Rakocevic, Dynamic Bandwidth Allocation in Multi-Class IP Networks using Utility Functions, pp. 90 – 100, PhD thesis, Queen Mary, University of London, 2002.
[12] D. Reininger, R. Izmailov, B. Rajagopalan, M. Ott and D. Raychaudhuri, "Soft QoS Control in the WATMnet Broadband Wireless System," IEEE Personal Communications, pp. 34-43, Feb. 1999.
[13] S. Shenker, "Fundamental Design Issues for the Future Internet", IEEE Journal of Selected Areas in Communications, 13(7): 1176-1188, Sept. 1995.
[14] A. Sinha and A. A. Zoltners, "The Multiple-Choice Knapsack Problem", Operation Research, 27: 503–515, 1979.
[15] H. Kellerer, U. Pferschy and D. Pisinger, "Knapsack problems", Springer-Verlag, 317-322, 2004.
[16] S. McCanne, M. Vetterli, V. Jacobson, "Low complexity video coding for receiver-driven layered multicast", IEEE Journal on Selected Areas in Communications, 15(6): 983 – 1001, 1997.
[17] J. Hartung, A. Jacquin, J. Pawlyk, K. Shipley, "A Real-time Scalable Video Codec for Collaborative Applications over Packet Networks", Proceedings of ACM Multimedia, pp. 419 – 426, 1998.
[18] S. Chakrabarti , R. Wang, "Adaptive Control for Packet Video", Proceedings of IEEE International Conference on Multimedia Computing and Systems, pp. 56 – 62, 1994.
[19] N. Duffield, K. Ramakrishnan, A. Reibman, "SAVE: an algorithm for smoothed adaptive video over explicit rate networks", IEEE/ACM Transaction on Networking, 6(6): 717 – 728, 1998.

Ning Lu received his B.S. degree from Beijing University of Posts and Telecommunications, China in 2001 and M.S. degree from Queen Mary, University of London, UK in 2002, respectively. He is currently a research assistant in the MPI-QMUL Information Systems Research Centre, Macao Polytechnic Institute, Macao, China and a Ph.D student in the Department of Electronic Engineering, Queen Mary, University of London, UK. His research interests include call admission control and adaptive resource management for QoS support in multimedia wireless networks.

John Bigham is a Reader in the Department of Electronic Engineering, Queen Mary, University of London, UK. His research interests are centred round the development of intelligent support for communications and network resource management using agent technology and learning techniques. He has developed several agent systems for the control of fixed and wireless networks. Recent work has concentrated on techniques for improving QoS and capacity in wireless networks in the commercial and military domains.