DiffServ Resource Allocation for Fast Handoff in Wireless Mobile Internet

Yu Cheng and Weihua Zhuang, University of Waterloo

ABSTRACT

The next-generation wireless networks are evolving toward a versatile IP-based network that can provide various real-time multimedia services to mobile users. Two major challenges in establishing such a wireless mobile Internet are support of fast handoff and provision of quality of service (QoS) over IP-based wireless access networks. In this article, a DiffServ resource allocation architecture is proposed for the evolving wireless mobile Internet. The registration-domain-based scheme supports fast handoff by significantly reducing mobility management signaling. The registration domain is integrated with the DiffServ mechanism and provisions QoS guarantee for each service class by domain-based admission control. Furthermore, an adaptive assured service is presented for the stream class of traffic, where resource allocation is adjusted according to the network condition in order to minimize handoff call dropping and new call blocking probabilities.

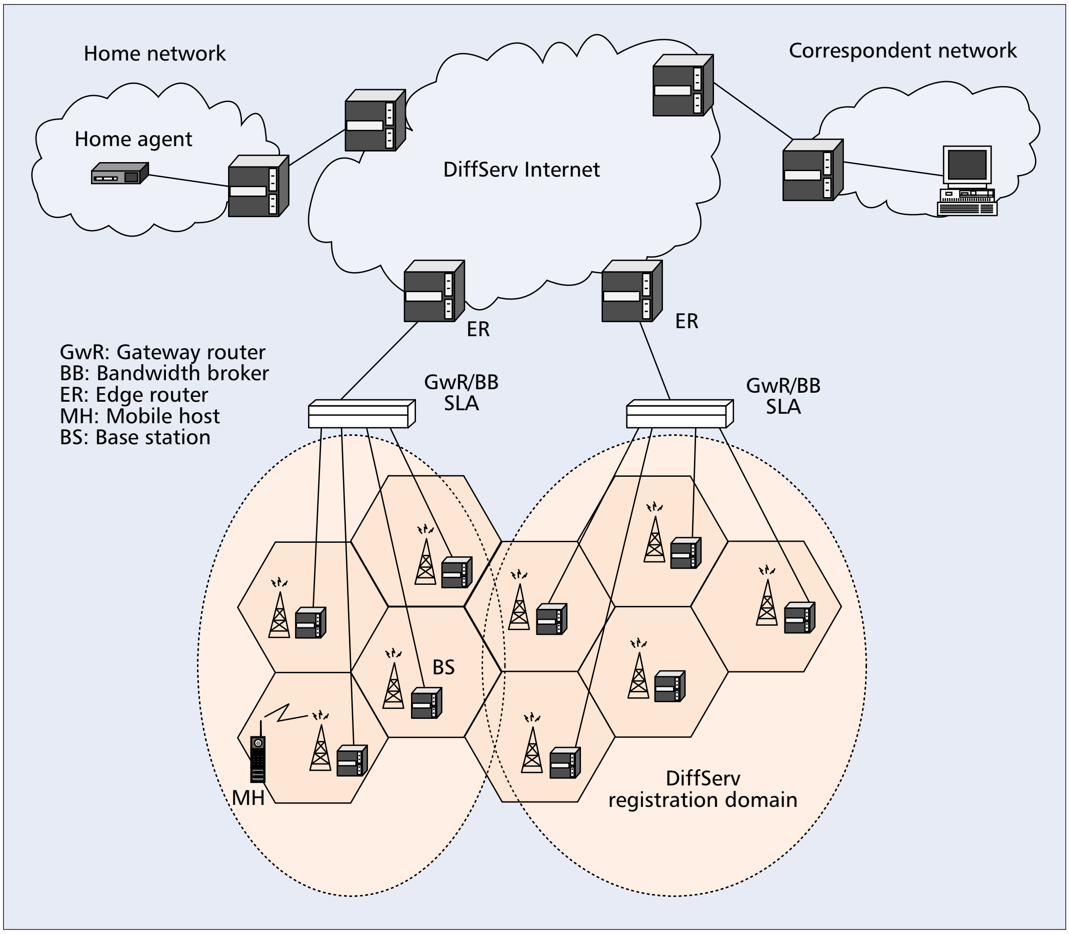
INTRODUCTION

Provision of various real-time multimedia services to mobile users is the main objective of the next-generation wireless networks, which will be IP-based and are expected to interwork with the Internet backbone seamlessly [1]. The establishment of such wireless mobile Internet is technically very challenging. Two major tasks are the support of fast handoff and the provision of quality of service (QoS) guarantee over IP-based wireless access networks.

The next-generation wireless networks will adopt micro/picocellular architectures for various advantages including higher data throughput, greater frequency reuse, and location information with finer granularity. In this environment, the handoff rate grows rapidly and fast handoff support is essential. Especially for real-time traffic, the handoff call processing should be fast enough to avoid high loss of delay-sensitive packets. To achieve fast handoff requires both a fast location/mobility update scheme and a fast resource allocation scheme. The popular scheme for fast location update is a *registration-domain-based* architecture. The radio cells (or the related base stations) in a geographic area are organized into a registration domain (e.g., a cellular IP network in the Cellular IP scheme [2], a foreign domain in the HAWAII approach [3], and a foreign network in the TeleMIP architecture [4]), and the domain connects to the Internet through a gateway [2] (a foreign root router [3] or a mobility agent [4]). When a mobile host (MH) moves into a registration domain for the first time, it will register the new care-of address (the address of the gateway) to its home agent. While it migrates within the domain, the mobility update messages will only be sent to the gateway, without registration with the home agent, often located far away. In the following, we present a resource allocation scheme over such a registration domain to achieve fast admission control, QoS guarantee, and high utilization of the scarce wireless frequency spectrum.

The integrated services (IntServ) and differentiated services (DiffServ) approaches [5] are the two main architectures for QoS provisioning in IP networks. The IntServ approach uses Resource Reservation Protocol (RSVP) to explicitly signal and dynamically allocate resources at each intermediate node along the path for each traffic flow. In this model, every change in an MH attachment point requires new RSVP signaling to reserve resources along the new path, which incurs latency in call admission control and is not suitable for fast handoff. Also, the heavy signaling overhead reduces the utilization efficiency of the wireless bandwidth. On the other hand, the DiffServ approach uses a much coarser differentiation model to obviate the above disadvantages, where packets are classified into a small number of service classes at the network edge. The packets of each class are marked and traffic conditioned by the edge router, according

0163-6804/02/$17.00 © 2002 IEEE



■ **Figure 1.** *A conceptual model of DiffServ registration-domain-based wireless network architecture.*

We consider wireless links as bottleneck links in the domain and the SLA is negotiated mainly based on the wireless resource availability. The gateway conditions the aggregate traffic for each service class according to the SLA resource commitments.

to the resource commitment negotiated in the service level agreement (SLA). In each core router, QoS for different classes is differentiated by different *per-hop behaviors* [5]. Resource allocation is performed by the *bandwidth broker* in a centralized manner, without dynamic resource reservation signaling and reservation status maintenance in the core routers. In this article the registration domain will be modeled as a DiffServ administrative domain, with the gateway router as the edge router connecting to the Internet backbone and the base stations as the edge routers providing Internet access to MHs. A bandwidth broker will manage the resource allocation over the DiffServ registration domain.

This article is structured as follows. First, we describe the system architecture. Then we investigate resource allocation and call admission control over the DiffServ registration domain. After that we present an adaptive assured service, in which multimedia applications experience bandwidth degradation and compensation, depending on the resource availability in wireless links. Finally, we present some numerical results and conclude this research.

System Architecture

The system under consideration is a registration-domain-based mobility management architecture as illustrated in Fig. 1, where DiffServ is used to provision QoS. Under the assumption that Cellular IP, HAWAII, or TeleMIP can be used to manage mobility and support fast handoff in this architecture, we focus on QoS provision and resource allocation in the following.

In the system, all the registration domains are DiffServ administrative domains in which all the routers are DiffServ IP routers. The gateway and base stations are edge routers, and are connected through core routers. The gateway is the interface connecting to the DiffServ Internet backbone, where an SLA is negotiated to specify the resources allocated by the Internet service provider to serve the aggregate traffic flowing from/into the gateway. We consider wireless links as bottleneck links in the domain, and the SLA is negotiated mainly based on the wireless resource availability. The gateway conditions the aggregate traffic for each service class according to the SLA resource commitments. The base stations provide MHs with access points to the Internet, and perform per-flow traffic conditioning and marking when data flow in the uplink direction. All base stations in the same registration domain are connected to the same gateway router. All DiffServ routers use three separate queues to provide premium service, assured service, and best effort service, respectively. The three buffers are served under priority scheduling or Weighted Fair Queue (WFQ) scheduling. The traffic classes provided by the next-generation wireless networks can be mapped to these three DiffServ classes. For example, in a Univer-

Once an MH is admitted to a registration domain, it can hand off to other cells within the domain without the involvement of further call admission control in the bandwidth broker.

sal Mobile Telecommunications System (UMTS) wireless network [6], the conversational and streaming classes can be mapped to the premium and assured services, respectively, while the interactive class or background traffic can be mapped to the best effort class. A bandwidth broker in the gateway router is responsible for resource allocation and call admission control over the DiffServ registration domain.

Domain-Based Call Admission Control

For simplicity, we assume that an *effective bandwidth* can be used to characterize both the traffic characteristics and QoS requirements, and the calls belonging to the same service class have homogeneous traffic characteristics, and thus the same effective bandwidth. For example, for a premium service call the peak rate can be used as effective bandwidth, and for an assured service call the effective bandwidth can be determined using the approach given in [7]. The resource commitments specified in the SLA can then be represented in terms of how many calls for each class are allowed in the registration domain. As a result, the proposed admission control procedure is straightforward: whenever a new MH requests admission to a registration domain, the bandwidth broker determines whether to admit or reject the new call, based on the number of calls currently in service and the SLA allocation for the service class to which the new call subscribes. The new call must be dropped if all the SLA allocation is occupied. This procedure requires very simple communications between the edge router (base station) and the bandwidth broker (in the gateway router), and can be executed very fast. Furthermore, once an MH is admitted to a registration domain, it can hand off to other cells within the domain without the involvement of further call admission control in the bandwidth broker.

Such a simple resource allocation scheme in fact implies a very complicated design problem. The number of base stations in a domain, the resources allocated to each service class in each base station, and the resource commitment in the SLA should be determined carefully so that the new call blocking and handoff dropping probabilities are reasonably low, while considering the traffic load in the registration domain, the mobility information, and the call duration statistics. Naghshineh and Acampora solved this design problem in situations where the interval between call arrivals, cell residence time, and call duration are independently and exponentially distributed [8]. An important conclusion is that a predetermined handoff call dropping probability can always be guaranteed by properly designing the admission controller. However, in the analysis, handoff calls and new calls are not differentiated. From the users' point of view, it is better to be blocked at the beginning of a call than to be dropped in the middle of one. As a result, handoff calls should be serviced with higher priority than new calls. To further decrease the handoff dropping probability, here we use the guard channel scheme [9] to reserve a fixed percentage of each base station's resources for handoff calls and extend the analysis given in [8] to include this situation.

For simplicity, we assume that the complete partitioning scheduling mechanism among the service classes is used, so the admission control of each class can be considered separately. Consider a registration domain including M cells, where the new call arrivals of a certain service class are Poisson with a mean rate of λ calls per cell per unit time, and the call duration is exponentially distributed with mean $1/\mu$. Each cell can serve up to C calls of the class under consideration, and a percentage (α) of the cell capacity, αC, is set as the guard channel to protect the handoff calls. The channel holding time in a base station (i.e., the time a call spends with any base station before handing off to another base station) is exponentially distributed with mean $1/h$ (i.e., the handoff rate is h). Assume that the handoff rate from any cell to any other cell is such that all cells experience the same rate of handoff call arrivals. The algorithms in [8] can be used to calculate the new call blocking and handoff call dropping probabilities in such a registration domain, with an extension to include the effect of the guard channel.

There are two levels of new call blocking. A new call is blocked by the admission controller if the total number of calls in the domain exceeds the SLA resource allocation N ($C < N < MC$), and/or if the serving cell runs out of resources for new calls and cannot accept additional ones. The handoff call dropping happens only at the cell level, when the serving cell runs out of all its resources (including the guard channel) and cannot accept the new handoff call. Based on [8], the handoff call dropping can be taken into account by considering the fact that the effective (actual) call departure rate μ_e is higher than the "natural" call departure rate μ, as $\mu_e = \mu + hP_H$ where P_H denotes the handoff call dropping probability. The probability of being blocked by the admission controller, denoted P_{Badm}, can be obtained by analyzing an $M/M/N$ queue with an Erlang load of $M\lambda/\mu_e$. With the domain admission controller, new calls arrive at a cell with rate $\lambda(1 - P_{Badm})$. The call blocking probability at the cell (denoted P_{Bcell}) and call dropping probability can be obtained by analyzing an $M/M/C$ queue with an Erlang load of $\lambda(1 - P_{Badm})/\mu_e$ and the guard channel αC . When the queue occupancy exceeds the guard threshold, the Erlang load reduces to $h/(\mu + h)$ of the total traffic load, since each call intends to hand off to a neighbor cell with a probability of $h/(\mu + h)$ in the steady state. We can see that the $M/M/N$ and $M/M/C$ queues are coupled via P_H. We present in [10] the details of how to solve the coupled queues to obtain P_{Badm}, P_{Bcell}, and P_H. The overall new call blocking probability in a registration domain is approximately given by $P_{Badm} + P_{Bcell}$.

The Adaptive Assured Service

The streaming class defined in third-generation wireless networks can be supported by the assured service in a DiffServ architecture.

Buffer configuration	Management rule	QoS level	Loss probability	Effective bandwidth
[0, B]	Any packet accepted, when $0 \le X < B$	High	$\in_1$ to L_1, L_2 and L_3	e_1
[0, B_1, B] ($0 < B_1 < B$)	L_1 packets accepted, when $0 \le X < B$; L_2 and L_3 packets accepted, when $0 \le X < B_1$	Medium	$\in_1$ to L_1, $\in_2$ to L_2 and L_3	e_2
[0, B'_1, B'_2, B] ($0 < B'_1 < B'_2 < B$)	L_1 packets accepted, when $0 \le X < B$; L_2 packets accepted, when $0 \le X < B'_2$; L_3 packets accepted, when $0 \le X < B'_1$;	Low	$\in_1$ to L_1, $\in_2$ to L_2, and $\in_3$ to L_3	e_3

Table 1. *The buffer configurations used to provide the assured service.*

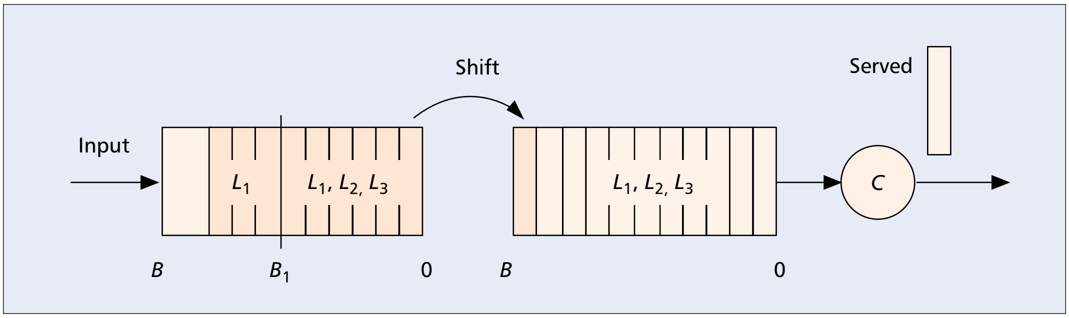
Streaming class traffic, such as a streaming video, normally does not require very strict timely delivery but does need a guaranteed minimum delivery rate. Adaptive coding can be applied to this type of traffic to improve sustain probability when the network congests. A good example is the MPEG video coding format, where a base layer contains basic and extension layer additional information. The video quality and bandwidth consumption can be scaled down to the bottom by only transmitting the base layer information. In a wireless network, because resource availability fluctuates frequently due to user mobility and channel quality variations, this type of adaptive service is very important to improve resource utilization efficiency. The adaptive framework to be discussed in this section only takes mobility into consideration. That is, the bandwidth allocated to an adaptive video changes only when there is a new call arrival, call completion, or handoff.

Here we propose to use a partitioned buffer [7] with size B to serve a layered video at each DiffServ router. Assume video traffic is coded to three layers, L_1, L_2, and L_3. The assured service buffer uses three configurations (1, 2, and 3) to provide three levels of QoS (high, medium, and low) to video traffic, as shown in Table 1, where X denotes the number of packets queued in the buffer, and $\in_1$, $\in_2$, and $\in_3$ ($\in_1 < \in_2 < \in_3$) denote different levels of loss probability provided by the different buffer configurations, respectively. For simplicity, we consider a homogeneous environment where all the video traffic for the assured service in the domain has homogeneous statistical characteristics and the buffer for the assured service in a DiffServ router is a homogeneous multiplexing system. We model each video traffic flow as a multiclass Markov-modulated fluid source (MMFS), where an underline Markov chain determines the traffic generation rate at each time instant. At each state of the Markov chain, three layers of traffic, L_1, L_2, and L_3, are generated. In such an environment, an *optimal effective bandwidth* can be calculated by using the technique developed in [7], which is the minimal channel capacity required to guarantee the loss requirements for all the layers of MMFS traffic when the buffer partition thresholds are optimally selected. For each QoS level listed in Table 1, the associated effective bandwidth can be calculated and is denoted e_1, e_2 and e_3 ($e_1 > e_2 > e_3$), respectively.

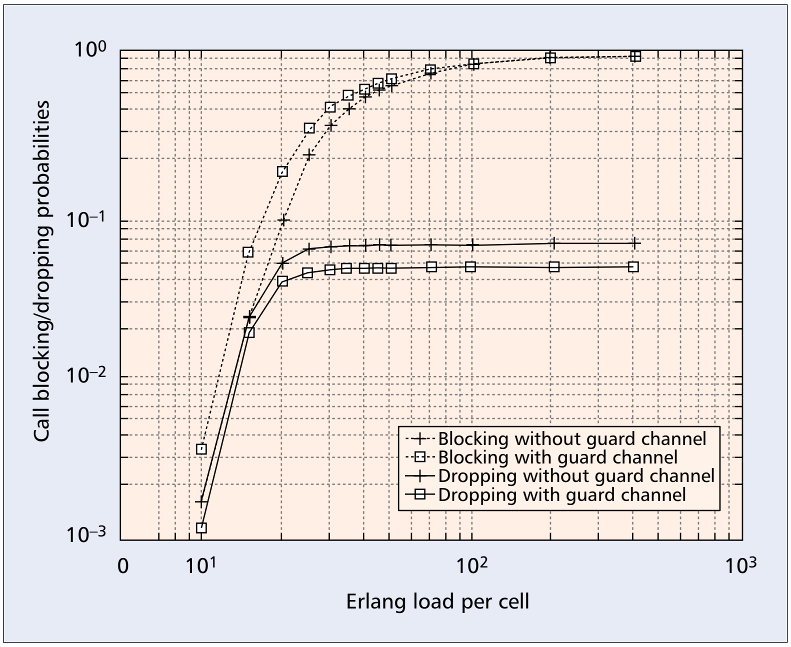
Under the assumption that the wireless link is always the bottleneck and the traffic will be served without loss in the wireline links, the adaptive mechanism is mainly used for the buffers in the MHs (when the traffic flows in the uplink direction) or the base stations (when the traffic flows in the downlink direction). Each base station has a local admission/rate controller (LARC) to manage admission control and adaptive bandwidth allocation in the cell. In the downlink direction, the LARC can directly adjust the buffer configuration in the base station when bandwidth adaptation happens; in the uplink direction, the LARC should send messages to MHs to control the buffer configuration adjustment for bandwidth adaptation. The adaptive algorithm based on effective bandwidth is straightforward. If we begin with the light load situation when sufficient resources are available, a new call or handoff call arrival to the cell is admitted with bandwidth allocation of e_1, meaning that the high level of QoS is provided to the traffic. As the traffic load increases to the level where a new call or handoff call cannot be admitted with bandwidth e_1, the LARC then reduces the bandwidth allocation to each already accepted video call from e_1 to e_2, and adjusts the assured service buffer in the base station or MHs from configuration 1 to configuration 2 to fit the adaptive bandwidth allocation. The new request is accepted also with bandwidth allocation of e_2. In this situation, the medium level of QoS is to be provisioned. If the traffic load further increases to a certain degree, the LARC reduces the bandwidth allocation of both existing traffic and new requests to e_3 and changes to buffer configuration 3, trying to accept as many calls as possible by degrading the QoS to the low level. On the other hand, when traffic load decreases, the buffer can shift back to the higher-level configuration, and the LARC can allocate more bandwidth to the calls for better QoS. Consider that at the current level each call is allocated an effective bandwidth of e_i ($i = 2, 3$). At the moment a call completes or hands off to a neighbor cell, the LARC will check whether the available resources are enough to improve the bandwidth allocation of each call to e_{i-1}. If the free resources can be used to accept two more calls with allocation of e_{i-1} after the bandwidth increase,[1] the LARC will implement the adaptation and the buffer will shift up to the one-level-higher configuration.

When the adaptation happens from one configuration to another providing higher QoS, the

[1] *Here the capacity for two calls with allocation of* e_{i-1} *serves as a threshold to detect whether the traffic load really decreases. If the threshold is set to 0, the fluctuation of available capacity due to the new arrival, handoff or call completion will lead to too frequent adaptation of the buffer configuration, but with little improvement of the QoS. How to optimally set the threshold needs further investigation.*



■ **Figure 2.** *The buffering mechanism for smooth QoS adaption.*



■ **Figure 3.** *A comparison of domain-based admission control systems with and without guard channel.*

buffer is lightly loaded at the moment and has free queuing space left. The higher-level configuration can use the space to allow more packets to be buffered to provide better QoS. The transition is always smooth. On the other hand, when the buffer needs to switch into a configuration providing lower QoS, more packets should be dropped. The following mechanism can be used to achieve a smooth transition from a higher QoS level to a lower one (e.g., from `high` to `medium`). When the LARC detects that the QoS level needs to scale down, it is reasonable to expect that the buffer is full or almost full. We then allocate a dynamic buffer of size B, which concatenates with the original buffer to accept the input traffic, as shown in Fig. 2. However, the dynamic buffer accepts packets according to the `medium` configuration. When a packet is pumped out from the original buffer, the dynamic buffer will shift its head-of-line packet to the free space. After B packets (all of which were accepted according to the `high` configuration) are served, the queue in the original buffer has been updated to the `medium` configuration. Then the dynamic buffer can be released.

Performance Evaluation

In this section we present two numerical examples to show the performance of the proposed resource allocation techniques. The first example shows that handoff dropping probability can be decreased by setting a fixed guard channel in each cell for handoff calls. The second shows that the proposed adaptive scheme can decrease the new call blocking and handoff call dropping probabilities from the nonadaptive scheme.

Call Admission with Guard Channel

Here, we use the parameter configuration given in [8] and make comparisons between call admission with and without guard channel in terms of the call blocking and dropping probabilities. With the complete partitioning scheduling mechanism, we focus on the admission control of assured service traffic. New calls arrive at each cell according to a Poisson process, and the exponentially distributed call duration has an average of $1/\mu = 0.5$ units of time. The channel holding time in each base station is also exponentially distributed with an average of $1/h = 0.1$ units of time. Each registration domain has $M = 20$ cells. Each cell can support up to $C = 20$ assured service calls. Bandwidth adaptation is not considered here. New call requests are rejected by the domain admission controller if there are $N = 320$ (80 percent of the total capacity of the domain) calls currently in service in the domain. When guard channel is used, $\alpha = 10$ percent of cell capacity (2 calls) is reserved for handoff calls.

Figure 3 shows the new call blocking and handoff call dropping probabilities vs. the Erlang load of new calls per cell, ρ. For comparison, the performance curves of the admission control without guard channel [8] are also included. From the figure, we can see that the introduction of guard channel can decrease handoff call dropping probability at the cost of an increase in new call blocking probability. In the light load situation, it is highly possible that the number of ongoing calls in the registration domain is less than 320, and no call is rejected by the domain admission controller. The mechanism used in each cell determines the call blocking and dropping probabilities. The trade-off between the decrease of handoff call dropping and increase of new call blocking is clearly observed. As the Erlang load increases, both the call dropping and blocking probabilities increase. When the traffic load is extremely high, most of the new

Erlang load	New call blocking probability			Handoff call dropping probability		
	`High` configuration	`Medium` configuration	`Low` configuration	`High` configuration	`Medium` configuration	`Low` configuration
15	0.0563	0.0246	0.0102	0.0185	0.0062	0.0026
50	0.6614	0.6286	0.5960	0.0458	0.0336	0.0303
400	0.9576	0.9535	0.9494	0.0462	0.0339	0.0307

Table 2. *The new call blocking and handoff call dropping probabilities achieved by the adaptive assured service.*

requests will be rejected by the domain admission controller, which limits the Erlang load in each cell to approximately N/M [8], and therefore limits the handoff call dropping probability to 0.0644 without guard channel and to 0.0463 with guard channel. We can see that the handoff call dropping probability can be guaranteed to a predetermined level by properly selecting N when other parameters are fixed. This analysis should be helpful in determining the resource requirement in the SLA negotiation between the registration domain and the Internet service provider.

Adaptive Bandwidth Allocation

Consider a registration domain with the same configuration as in the above numerical analysis example with the guard channel. To show how new call blocking and handoff call dropping probabilities can be reduced by implementing the adaptive assured service, the bandwidth allocation to each call can now be adjusted between e_1, e_2, and e_3. The capacity allocated to the assured service in each cell is Ce_1 packets/s. If at some time the LARC adjusts the QoS to `medium` and reaches the stable state, each cell can admit $\lfloor Ce_1/e_2 \rfloor$ calls, and the domain admission controller will allow $0.8M\lfloor Ce_1/e_2 \rfloor$ calls to be accepted. The guard channel in each cell will be set to $\lceil \alpha Ce_1/e_2 \rceil$. If the LARC adjusts the QoS to `low`, then the cell capacity, admission controller capacity, and guard channel will be adjusted to $\lfloor Ce_1/e_3 \rfloor$, $0.8M \lfloor Ce_1/e_3 \rfloor$ and $\lceil \alpha Ce_1/e_3 \rceil$ calls, respectively.

For simplicity, we model a video flow as an on-off source as in [7]. The traffic is used to illustrate the efficiency of the adaptive scheme, even though the model may not be practical for video sources. Each video source in the on state generates traffic with a total rate R_p = 170.21 packets/s, which can be coded into three layers L_1, L_2, and L_3 at rates of $R_p/8$, $R_p/8$, and 3 $R_p/4$, respectively. The buffer size in an MH or a base station is set to B = 250 packets. The three levels of loss probabilities used in the adaptive system, $\in_1$, $\in_2$, and $\in_3$, are set to 10^{-10}, 10^{-4}, and 10^{-1}, respectively. For the `high` configuration listed in Table 1, a first-in-first-out buffer is used and the effective bandwidth (e_1) is 142.45 packets/s; for the `medium` configuration, the technique in [7] is used to calculate the effective bandwidth (e_2), equal to 119.82 packets/s, while the optimal partition threshold B_1 is 210 packets; for the `low` configuration, the effective bandwidth (e_3) is 112.65 packets/s and the optimal thresholds B_1' and B_2' are 64 and 171 packets, respectively. By scaling down the QoS requirement, the bandwidth allocated to the ongoing calls can be reduced. The resources accumulated by bandwidth reduction are then used to accept more new calls and handoff calls. Figure 4 shows the new call blocking and handoff call dropping probabilities for the three configurations. It is observed that both the new call blocking and handoff dropping probabilities decrease obviously in the adaptive scheme as the QoS level is reduced. For clarity, Table 2 gives the new call blocking and handoff call dropping probabilities for the medium load (λ = 15), high load (λ = 50), and extremely high load (λ = 400) conditions, respectively.

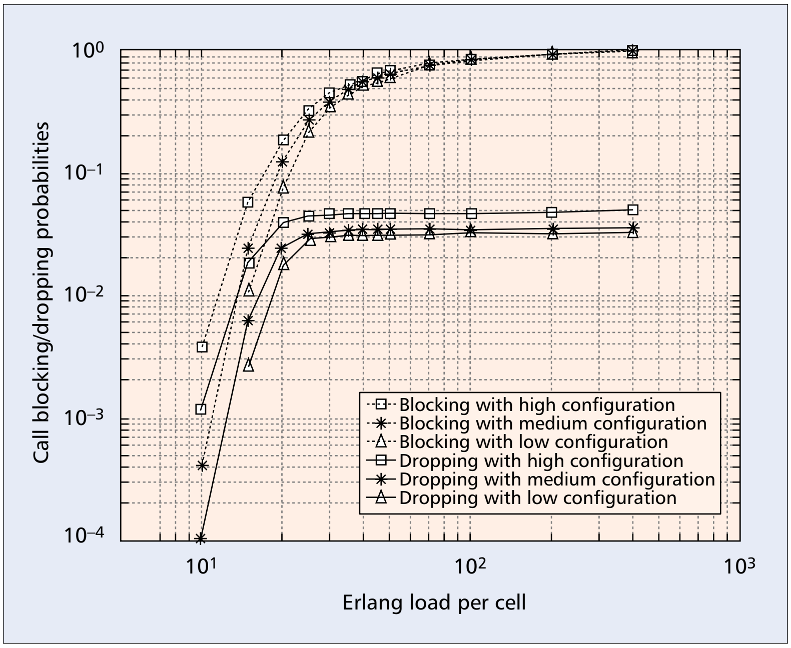


Figure 4. *The performance improvement achieved by the adaptive framework.*

Conclusion

This article presents the registration-domain-based architecture and adaptive assured service for the next-generation wireless mobile Internet. The architecture can support fast handoff and facilitate QoS provision for differentiated services. For the statistical environment with exponential distributions, the analysis demonstrates that:

- The handoff call dropping probability can be guaranteed to a predetermined level by properly allocating the resources to a certain class of traffic in a registration domain.
- The guard channel scheme can be used to further reduce handoff call dropping probability.

- The adaptive service can improve resource utilization while guaranteeing call-level QoS (i.e., call blocking and dropping probabilities).

This analysis can help to determine the resource requirement in the SLA negotiation between the registration domain and the Internet service provider. However, the resource allocation over a registration domain in a nonexponentially distributed statistical environment needs further investigation.

The resource allocation over a registration domain in a nonexponentially distributed statistical environment needs further investigation.

Acknowledgments

This work is supported by research grants from Bell University Labs (BUL) at the University of Waterloo and the Canadian Institute for Telecommunications Research (CITR). This article is based on our published material from 3Gwireless 2002 by the Delson Group at http://www.delson.org.

References

[1] L. Bos and S. Leroy, "Toward an All-IP-Based UMTS System Architecture," *IEEE Network*, vol. 15, no. 1, Jan.–Feb. 2001, pp. 36–45.

[2] A. T. Campbell *et al.*, "Design, Implementation and Evaluation of Cellular IP," *IEEE Pers. Commun.*, vol. 7, no. 4, Aug. 2000, pp. 42–49.

[3] R. Ramjee *et al.*, "HAWAII: A Domain-based Approach for Supporting Mobility in Wide-area Wireless Networks," *Proc. IEEE Int'l. Conf. Network Protocols*, 1999, pp. 283–92.

[4] S. Das *et al.*, "TeleMIP: Telecommunications-Enhanced Mobile IP Architecture for Fast Intradomain Mobility," *IEEE Pers. Commun.*, vol. 7, no. 4, Aug. 2000, pp. 50–58.

[5] S. Blake *et al.*, "An Architecture for Differentiated Services," IETF RFC 2475, Dec. 1998.

[6] S. Dixit, Y. Guo, and Z. Antoniou, "Resource Management and Quality of Service in Third-Generation Wireless Networks," *IEEE Commun. Mag.*, vol. 39, no. 2, Feb. 2002, pp. 125–33.

[7] Y. Cheng and W. Zhuang, "Optimal Buffer Partitioning for Multiclass Markovian Traffic Sources," *Proc. IEEE GLOBECOM '01*, vol. 3, 2001, pp. 1852–56.

[8] M. Naghshineh and A. S. Acampora, "Design and Control of Micro-Cellular Networks with QoS Provisioning for Real-Time Traffic," *Proc. IEEE 3rd Int'l Conf. Univ'l. Pers. Commun.*, 1994, pp. 376–81.

[9] O.T.W. Yu and V.C.M. Leung, "Adaptive Resource Allocation for Prioritized Call Admission over an ATM-Based Wireless PCN," *IEEE JSAC*, vol. 15, no. 7, Sept. 1997, pp. 1208–25.

[10] Y. Cheng and W. Zhuang "A DiffServ Resource Allocation Scheme Supporting Fast Handoff in IP-Based Wireless Networks," *Proc. 3Gwireless 2002*, San Francisco, CA, May 2002.

Biographies

YU CHENG [S] (ycheng@bbcr.uwaterloo.ca) received his B.S. and M.S. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1995 and 1998, respectively. He is currently working toward a Ph.D. degree at the University of Waterloo, Canada. His current research interests are QoS provisioning and resource management in IP-based networks.

WEIHUA ZHUANG [SM] (wzhuang@bbcr.uwaterloo.ca) received a Ph.D. degree in electrical engineering in 1993 from the University of New Brunswick, Canada. Since 1993 she has been a faculty member at the University of Waterloo, Canada, where she is an associate professor in the Department of Electrical and Computer Engineering. Her current research interests include resource management and wireless networking for multimedia personal communications. She received the Premier's Research Excellence Award (PREA) from the Ontario Government in 2001.