End-to-end Dynamic Adaptive QoS Provisioning over GPRS Wireless Mobile Network

Oliver Yu and Shashank Khanvilkar
Department Of Electrical and Computer Engineering
University of Illinois at Chicago
851 South Morgan Street, 1020 SEO
Chicago, Illinois 60607

Correspondence: oyu@ece.uic.edu; Phone: (312) 996-2308; Fax: (312) 413-0024

Abstract: **The General Packet Radio Service (GPRS) offers performance guaranteed packet data services to mobile users over wireless frequency-division duplex links with time division multiple access, and core packet data networks. This paper presents a dynamic adaptive guaranteed quality-of-service (QoS) provisioning scheme over GPRS wireless mobile links by proposing a guaranteed QoS media access control (GQ-MAC) protocol and an accompanying adaptive prioritized-handoff call admission control (AP-CAC) protocol to maintain GPRS QoS guarantees under the effect of mobile handoffs. The GQ-MAC protocol supports bounded channel access delay for delay-sensitive traffic, bounded packet loss probability for loss-sensitive traffic, and dynamic adaptive resource allocation for bursty traffic with peak bandwidth allocation adapted to the current queue length. The AP-CAC protocol provides dynamic adaptive prioritized admission by differentiating handoff requests with higher admission priorities over new calls via a dynamic multiple guard channels scheme, which dynamically adapts the capacity reserved for dealing with handoff requests based on the current traffic conditions in the neighboring radio cells. Integrated services (IntServ) QoS provisioning over the IP/ATM-based GPRS core network is realized over a multi-protocol label switching (MPLS) architecture, and mobility is supported over the core network via a novel mobile label-switching tree (MLST) architecture. End-to-end QoS provisioning over the GPRS wireless mobile network is realized by mapping between the IntServ and GPRS QoS requirements, and by extending the AP-CAC protocol from the wireless medium to the core network to provide a unified end-to-end admission control with dynamic adaptive admission priorities.**

I. INTRODUCTION

General Packet Radio Service (GPRS) is a Global System for Mobile communications (GSM) service that provides mobile subscribers with performance guaranteed packet data services over GSM radio channels and external packet data networks. The GPRS wireless subsystem consists of a number of mobile stations (MS's) generating traffic according to the negotiated quality-of-service (QoS) profiles, which compete for communication with a base station (BS) in a radio cell area via a wireless frequency-division duplex (FDD) link with time division multiple access (TDMA). The GPRS QoS profiles are defined in terms of the following QoS classes: precedence, delay, reliability, mean and peak throughputs. In this paper, the implemented QoS profiles are classified as streaming, conversational, interactive and background, and their mappings to GPRS QoS classes and attributive values are summarized in TABLE I.

TABLE I: QOS PROFILES DEFINITION

GPRS QoS Class	**QoS Attributes**	**Implemented QoS Profiles**			
		Streaming	**Conversational**	**Interactive**	**Background**
Precedence	Congestion Packet Discard Probability	Tolerable ($<10^{-2}$)	Tolerable ($<10^{-2}$)	Loss sensitive ($<10^{-5}$)	N/A
Delay	Latency	Bounded (<500 msec)	Bounded (<80 msec)	Less stringent than that of Conversational & Streaming	N/A
	Jitter	Stringent	Stringent	N/A	N/A
Reliability	Packet Loss Probability	Tolerable ($<10^{-2}$)	Tolerable ($<10^{-2}$)	Loss sensitive ($<10^{-5}$)	N/A
Mean and Peak Throughputs	Throughput	Guaranteed	N/A	Guaranteed	N/A
	Burstiness	Low	High	Higher than Conversational	N/A

A. QoS Provisioning over GPRS Wireless Medium

The wireless link is characterized by a broadcast mode in the downlink (BS to MS) direction and a multiple access mode in the uplink (MS to BS) direction. A medium access control (MAC) protocol is required to distribute packet transmission over the shared medium among all users. MAC protocols could be classified according to their method of resource sharing and the accompanying multiple access schemes. The GPRS standard [1] specifies the FDD/TDMA multiple access with four radio access priorities, and some reference guidelines on the resource sharing method but its overall design will be decided by the developers.

Resource sharing based on demand assignment can be employed to minimize wasted bandwidth due to under-utilization with dedicated assignment and to collision with random access. With Packet Reservation Multiple Access (PRMA) [2] protocol, voice source uses slotted ALOHA for reserving the same slot position in future frames, while data sources have to contend for a slot whenever they have packets to send. Enhanced PRMA protocols (such as Centralized-PRMA [3] and Integrated-PRMA [4]) improve channel efficiency and provide some kind of service fairness for data sources. Since all these protocols suffer variable packet access delay, QoS with bounded delay could not be guaranteed. In Distributed Queuing Request Update Multiple Access (DQ-RUMA) [5] protocol proposed for wireless ATM networks, all traffic sources can reserve slots in future frames; new traffic sources and admitted idle traffic sources will contend equally in the request access slots for reservation. Consequently, already admitted traffic sources suffer variable packet access delay. Most MAC protocols support average throughput resource reservation for bursty traffic since it optimizes resource utilization but it also increases packet loss probability due to buffer overflow, and packet delay and jitter. On the other hand, peak throughput reservation for bursty traffic support delay and jitter bounds but causes inefficient resource utilization.

This paper proposes a Guaranteed QoS Medium Access Control (GQ-MAC) protocol to enable performance guarantees for the four defined GPRS QoS classes. The protocol supports per-session dedicated reservation for streaming traffic class and prioritized on-demand reservation for conversational and interactive traffic classes. Traffic burstiness is counteracted with dynamic adaptive resource allocation with peak bandwidth allocation adapted to the current queue length.

In a mobile wireless system, call admission control (CAC) protocol aims to maximize the number of admitted or in-session traffic sources supported over the wireless medium while guaranteeing their QoS requirements. For MAC protocols that supports per-session dedicated reservation, the accompanying CAC protocol should give prioritized admission to handoff requests, which require lower blocking probability relative to new calls. In general, forced terminations of ongoing call sessions due to mobile handoff blocking are generally more objectionable than new calls blocking from the user's perspective. One common prioritized admission scheme is the static guard channel scheme [6][7] which gives a higher access priority to handoff requests over new calls by assigning them a higher capacity limit, while optimizing utilization of resources shared by handoff requests and new calls. In [8], the dynamic guard channel scheme is proposed to adapt the number of guard channels in a radio cell according to the current estimate of the handoff arrival rate derived from the current number of ongoing calls in neighboring cells and the mobility pattern, so as to keep the handoff block probability close to the targeted objective while constraining the new call blocking probability to be below a given level. In [9], the dynamic guard channel scheme is extended to provide admission control for multiple traffic classes; the handoff blocking probabilities are minimized at any cost without paying attention to the degradation of the new call blocking probabilities. In this paper, we propose a multiple dynamic guard channel scheme to provide admission control for multiple traffic classes; with the objective to minimize handoff blocking probabilities while minimizing the degradation to the new call blocking probabilities.

For the proposed GQ-MAC protocol that supports per-session dedicated reservation and on-demand reservation, this paper proposes the Adaptive Prioritized-handoff CAC (AP-CAC) protocol that differentiates handoff requests with different higher admission priorities over new calls via a dynamic multiple guard channels scheme, which dynamically adapts the capacity reserved for dealing with handoff requests based on the current number of ongoing calls in the neighboring radio cells and the mobility pattern. Handoff requests associated with per-session dedicated reservation have higher admission priority than those associated with on-demand reservation.

B. QoS Provisioning over GPRS Core Network

The GPRS core network employs GPRS supporting nodes (GSN's) to support packet data routing. As illustrated in Fig. 1, the gateway GSN (GGSN) acts as a logical interface to the external packet data networks (PDN's) and consequently to the global Internet. The supporting GSN (SGSN) is responsible for the delivery of packets to the MS's through one or more BS's within its service area. Intra-SGSN handoff requires switching between BS's and mobility support over the GPRS wireless medium. On the other hand, inter-SGSN handoff requires switching between edge SGSN's and mobility support over the GPRS core network via path rerouting. The protocol stacks of the GPRS component nodes are shown in Fig. 15. Between GGSN and SGSN, all user data and signaling packets are encapsulated by means of the GPRS Tunneling Protocol (GTP) to enable support of non-IP-based PDNs. Between SGSN and MS, further encapsulation is performed by the subnetwork-dependent convergence protocol (SNDCP) that maps network-level characteristics onto the characteristics of the underlying network.

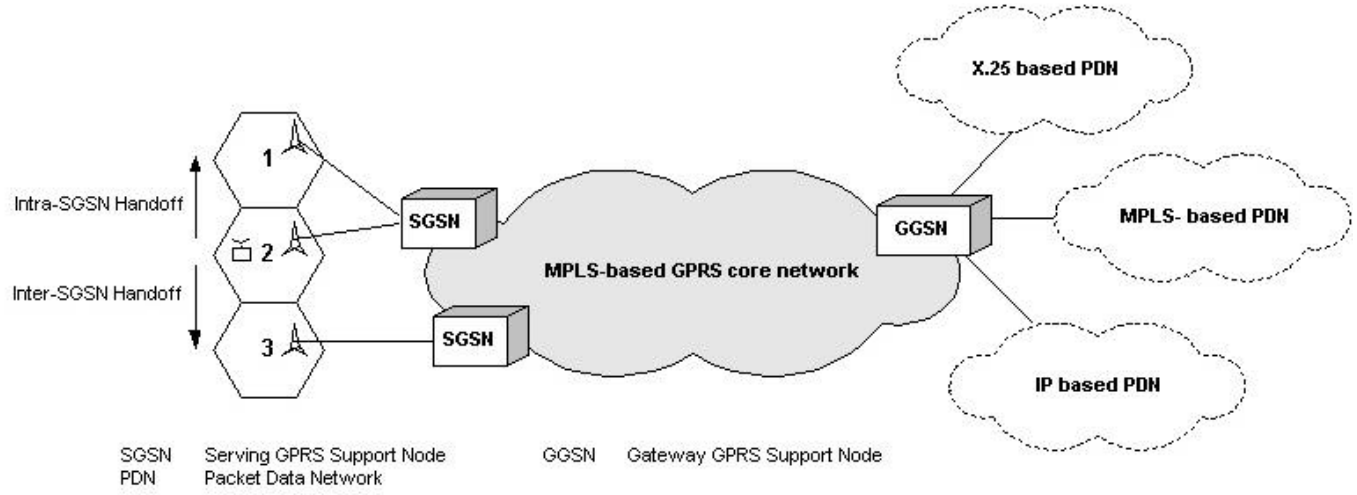

Fig. 1: MPLS-Based GPRS Architecture.

The QoS provisioning over the IP/ATM based GPRS core network will be realized with the integrated services (IntServ) QoS architecture [10][11]. The IntServ QoS architecture has been designed for the Next Generation Internet to provide performance guarantees through QoS control mechanism and per-flow resource allocation; and to provide fast packet forwarding through label switching (rather than IP destination address routing) by employing the multi-protocol label switching (MPLS) technology [12][13]. The IntServ QoS architecture provides guaranteed [14], controlled-load [15] and best-effort services. Guaranteed service provides an assured level of bandwidth and a firm end-to-end delay bound to jitter-intolerant real-time multimedia applications, e.g., two-way telephony and circuit emulation services. Controlled-load service provides soft quantitative guarantees on bandwidth or delay bound to jitter-tolerant near-real-time multimedia applications, e.g., various packetized audio or video streaming services. Best-effort service could be used to support non-real-time applications without any performance guarantees.

C. End-to-end QoS Provisioning over GPRS Wireless Mobile Network

End-to-end QoS provisioning over the GPRS wireless mobile network is realized by mapping the IntServ QoS to the GPRS QoS. The IntServ QoS architecture specifies the RSVP resource reservation protocol [16] for soft-state resource reservation along the traffic flow, through which an application could communicate its QoS requirements to the nodes along the traffic flow. In [17], extension of tunnel and trigger signaling to the RSVP is proposed to support end-to-end resource reservation. RSVP requires periodic refreshment of the soft-state reservation information maintained along the IP route of the traffic flow. We employ a hard-state RSVP-like reservation protocol [18] to reserve resources along the "pinned" MPLS path of the traffic flow.

New calls and handoff requests compete for bandwidth resources over a wired link in the core as well over a wireless link on the air interface. The proposed AP-CAC protocol over the wireless medium supports adaptive prioritized admissions for multiple traffic classes via the multiple dynamic guard channel scheme, which dynamically adapts the channel access capacities for handoff requests of different traffic classes based on the current number of ongoing calls in the neighboring radio cells and the mobility pattern. If a non-AP-CAC protocol is employed over the core network, then end-to-end adaptive prioritized admission priorities could not be maintained. This paper presents the extension of the AP-CAC scheme over the GPRS IP/ATM core network by employing the proposed mobile label switched tree (MLST) and the dynamic guard bandwidth scheme.

D. Organization of Paper

This paper is organized as follows. Section II describes the GQ-MAC protocol in terms of its features, implementation, performance analysis and results. Section III describes the AP-CAC protocol in terms of its features, implementation, performance analysis and results. Section IV describes the extension of AP-CAC protocol from the wireless medium to the GPRS core network. Section V concludes the paper.

II. GQ-MAC PROTOCOL

A. Channel Access Procedures

The GPRS packet data channels (PDCH) are classified as one of the following:

- Packet Random Access Channel (PRACH): This is the request access channel for uplink. It consists of two time-multiplexed channels: Signaling PRACH (S-PRACH) and User-data PRACH (U-PRACH).
- Packet Access Grant Channel (PAGCH): This is request acknowledgement channel for downlink. The BS uses PAGCH to broadcast the request status information.
- Packet Data Traffic Channels (PDTCH): The remaining PDCH's, on the uplink and downlink, act as PDTCH's. These are used for carrying the payload.

The request access priorities of the S-PRACH and U-PRACH are illustrated in TABLE II. Through the slotted Aloha random access protocol, the S-PRACH is multi-accessed by the signaling requests of new calls and handoffs to gain admission to the PDTCH. Handoff requests are assigned with higher access priority than that of new call requests. Admitted streaming traffic source enjoys per-session dedicated reservation of PDTCH resources. Admitted conversational and interactive traffic sources would have to perform on-demand reservation by multi-accessing the U-PRACH via slotted Aloha random access protocol. During the contention cycle, only the conversational traffic sources are allowed to multi-access the U-PRACH via the tree limited-contention access protocol. Consequently, conversational traffic class has a higher U-PRACH access priority than the interactive traffic class. On the other hand, the background traffic sources are allocated with unused PDTCH resources in a round robin fashion. The access protocols employed by the proposed GQ-MAC protocol are described as follows.

TABLE II: REQUEST ACCESS PRIORITIES

Request Access Type		**Access Priority**	**PRACH used**
Signaling	New Call	Low	S-PRACH
	Handoff	High	
User Data	Conversational Traffic	High	U-PRACH
	Interactive Traffic	Low	

A.1. Slotted Aloha

This random access protocol is employed by the signaling requests of new calls and handoff to access the S-PRACH. Mobile stations with signaling packets transmit immediately on the first available slot. If a collision occurs, they transmit on the next slot with probability "P_a". Handoff requests are given higher priority by having higher "P_a" value as compared to new call initiation requests.

A.2. Tree Protocol

This limited-contention access protocol is used for in-session channel access request for conversational traffic because it provides a deterministic channel access time [19]. A state transition diagram for the algorithm is illustrated in Fig. 2. Initially all the MS's in the cell can access the channel. If a collision occurs on any slot "k", the BS starts a contention cycle, which reserves slot "k+1" for MS's having MSB (Most Significant Bit) of their identifiers equal to "0". If a collision occurs again on slot "k+1", then slot "k+2" is reserved for MS's with first two MSB bits equal to "00". On the other hand if a request was successfully transmitted on slot "k+1" or if the slot were idle, the slot "k+2" would have been reserved for MS's having MSB bits equal to "1XX". This is continued according to the figure.

By allowing only conversational traffic sources to participate in the contention resolution cycle, it is possible to guarantee a bounded delay on channel access. The contention cycle can be showed to have a bounded length of: TDMA_FRAME_LENGTH * $(2^{(\log_2 j + 1)} - 1)$ [20]; Where j is the number of conversational MS's, simultaneously trying to access the U-PRACH. If voice is assumed to be the only conversational source in the system, then using the popular On-Off model with a voice activity factor of 0.4, it can be shown that the probability of more than 5 MS's trying to access the U-PRACH simultaneously is very low. Hence we can expect the channel access delay to be limited to 20 msec, assuming a GSM frame length of 4.615 msec.

A.3. Modified Slotted Aloha

This is used for U-PRACH access by interactive traffic sources. It is similar to slotted Aloha protocol discussed above, with following modifications:

- If a collision occurs, a binary exponential back-off algorithm is used which reduces the transmission probability, "P(x)", in the next slot by 0.5 (up to a lower limit). P(x) is calculated as a function of output queue length (§ II-B.3) and fraction of QoS that was satisfied.
- When a contention resolution cycle is in progress, P(x) is reduced to "0", i.e. Interactive MS's do not participate in the contention resolution cycle.

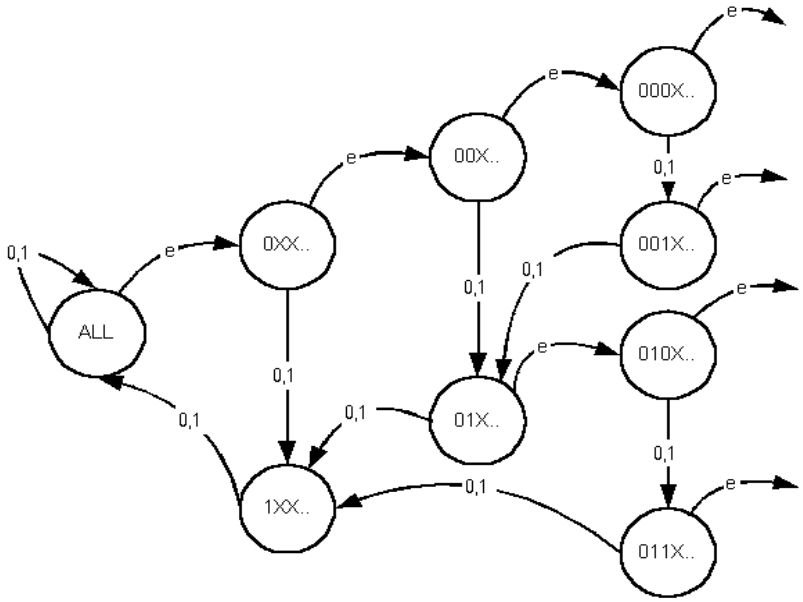

Fig. 2: State diagram for limited contention Tree Protocol.

B. QoS Support

To initiate a new call, a Call Initiation (Call_Init) request is sent on the S-PRACH using the slotted aloha protocol. The Call_Init request contains one or more of the following: Desired service type (Conversational, Streaming, Interactive or Background) and Requested Data Rate (RDR). On successful reception of this request, the AP-CAC module determines if enough resources are available to support this new call, without degrading the service guaranteed to others, and if admissible, sets up two state variables for that session in the BS: RDR and ADR (Achieved Data Rate). A temporary buffer is also set up to hold packets for that session and a suitable MS identifier is generated, which is transmitted to that MS on the PAGCH. We now discuss how each traffic type can be supported in the system.

B.1. Streaming Traffic

The system offers streaming as a dedicated service in multiples of quantized data rates, according to $C_i = i \times C/N$; where $i_{min} < i = i_{max}$, C_i is the quantized bandwidth requested, C is the total bandwidth that is available and N is total number of PDCH's. i_{min} and i_{max} depend on the lowest and highest streaming rate that is offered by the system. Thus for a streaming rate "X", one full PDTCH is allocated, while for rate "0.5X", only half of the PDTCH (a slot in every alternate frame) will be allocated and so on. Since resources are permanently allocated to a streaming call, it faces the problem of multi-access only during call set up. Thus a streaming call, once admitted, is guaranteed a bounded packet delay, constant inter-packet delay (i.e. minimal jitter) and a guaranteed throughput. By using a suitable FEC scheme, the packet loss due to corruption on the wireless link is limited.

B.2. Conversational Traffic

A conversational traffic source, after getting admitted, demands a channel resource only when it has data to send. The request is sent on the U-PRACH using the tree protocol. If no resources are available at that time, the BS will reallocate resources allocated to other sources (except streaming) to this conversational source. If all the resources are currently allocated to other such conversational sources, then the BS rejects the resource request, and the MS has to send another request, after discarding the first packet in the queue. When there is no data, an allocated resource is held for a channel holding time of 3

TDMA frames, following which an explicit release message is sent. Thus by using the tree protocol for channel access, packet delay for a conversational traffic is guaranteed to be bounded. Also since resource is reserved, till it is explicitly released, inter-packet delay is guaranteed to be constant.

B.3. Interactive Traffic

For interactive traffic sources, a scheduling algorithm is required which can guarantee the required throughput. We propose to use a distributed scheduling algorithm for allocation of uplink PDTCH's. This algorithm has been inspired from [21], with modifications. In this algorithm, MS takes active part in the scheduling process on the uplink. For every interactive stream that is admitted into the cell, the BS and the corresponding MS maintain the following state variables.

- Requested Data Rate (RDR): The MS sends the RDR value in the Call-Init request packet.
- Achieved Data Rate (ADR): The ADR value is continuously updated by MS/BS as it sends/gets data packets.

Every MS maintains a queue at its output interface having a finite length. After getting admitted into the cell, the MS sends a Rate Request Packet (RRP) on the U-PRACH requesting some number of PDTCH's, closely matching its RDR value. The access probability for this first RRP varies according to:

$$P(x) = e^{\frac{\frac{x}{L_U}-1}{(1-a)}} \qquad 1 < x \leq L_U$$

$$P(x) = 1 \qquad x > L_U$$

Where *P(x)* is the access probability with which the MS transmits the first RRP when there are *x* packets in the queue, L_U is a fixed upper threshold (100, say) and a is a factor that depends on the ADR/RDR value and changes the curve for *P(x)* as shown in Fig. 3.

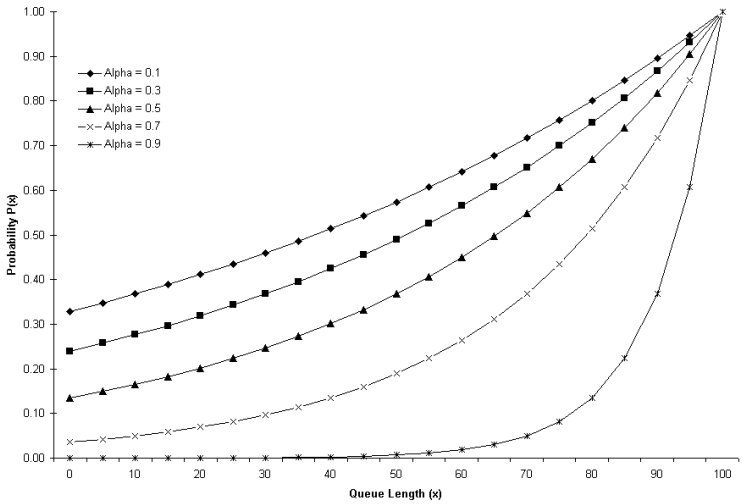

Fig. 3: Access Probability V/s Queue length for different values of a.

By having a directly proportional to the ratio of ADR/RDR, we have a mechanism that gives higher priority to MS's that have not been able to achieve the requested data rate, to send an RRP. If queue length increases beyond the threshold L_U, the MS will send an RRP with probability 1. Multiple thresholds can be defined in the same way to allow the MS to demand PDTCH's in addition to the minimum required. The BS attempts to allocate, as many PDTCH's requested, as available. For this, it might even pre-empt other like sources, which have achieved ADR >= RDR.

If a MS is allocated additional PDTCH's in response to its RRP packet, it will start transmitting on these PDTCH's. When there is no more data to be sent on the additional PDTCH, it is released. The last PDTCH is held by the MS for the channel holding time of 3 TDMA frames. The entire operation is illustrated in Fig. 4.

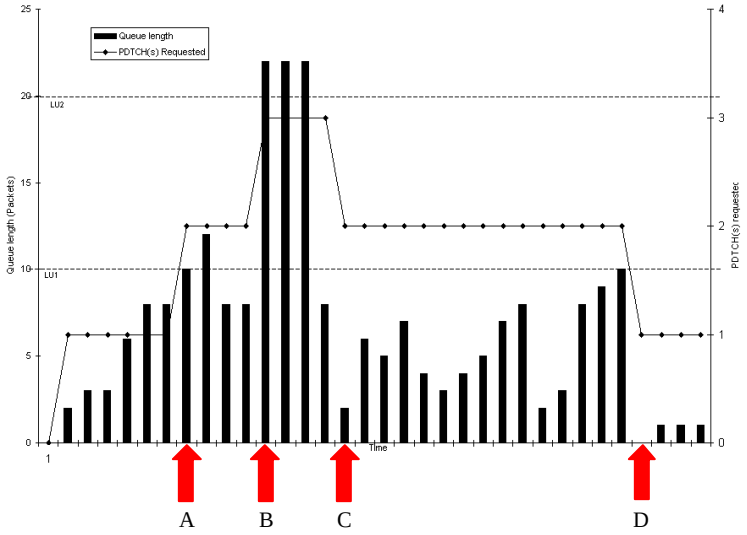

Fig. 4: Relation between queue length and PDTCH's requested.

For ease of explanation we assume that the MS transmits an RRP only when the threshold is reached. Also multiple thresholds are shown (L_{U1} = 10 and L_{U2} = 20) in this figure as an example. Initially the MS has been allocated a single PDTCH. At time "A", the queue length in the MS increases beyond the threshold L_{U1} (=10). This will make the MS send an RRP demanding an additional PDTCH, which we assume is granted. Thus now the MS will start transmitting on two PDTCH's. Again at "B", the length of the queue increases beyond the second threshold L_{U2} (=20), which makes the MS send another RRP demanding an additional PDTCH, which is again assumed to be granted. Now the MS starts transmitting with 3 PDTCH's. At time "C", the MS has only two packets in its output queue. Thus no packet is available to be transmitted on the 3rd PDTCH. As a result the MS has to relinquish the 3rd PDTCH and is left with just two PDTCH's. At time "D", the MS has 0 packets in its output queue. So it has no packets to be sent on both the PDTCH's. But the MS will relinquish one of them and hold the last PDTCH for the channel holding time.

4. Background Traffic

The background traffic sources, after getting admitted and allocated an identifier, camp on the PAGCH to see which slots are allocated to them in the uplink frames. The BS allocates unused PDTCH's to background traffic sources in a round robin fashion.

C. Performance Analysis and Results

We simulated a single cell containing one BS and a number of MS's. The simulation was carried out using OPNET, a network simulation tool. TABLE III lists the important parameters that were used in the simulation. We do not present the results for streaming and background traffic types due to the following reasons.

- A streaming call is the same as a circuit switched call. Hence its presence in the cell would only reduce the capacity of the system without affecting other results.
- Since no guarantees are given to background flows, their presence or absence do not affect results.

The simulation was, however, carried out only for conversational and interactive traffic types with the following models:

Conversational

We used the popular on-off model for voice, to simulate a conversational source. Since voice packets must be delivered in real time, there is a limit on the maximum transmission delay; any voice packet that has not been transmitted within 60 ms of its time of generation is discarded at the source. Packet dropping probability is defined as the ratio of number of packets dropped by an MS, to the total number of packets generated during the call. Voice packets may be dropped if they exceed the packet deadline mentioned above, or if they are corrupted. Since we have assumed an ideal wireless channel with no errors, packet loss occurs only because of deadline violation.

TABLE III: SIMULATION PARAMETERS

Parameter	Value
Number of Base Stations simulated	1
Number of Uplink/Downlink carrier pairs	1
TDMA frame duration	4.615 msec
Number of time slots in a TDMA frame	8
Number of traffic channels/carrier	7
Channel Data rate	270 kbps (approx.)
Average length of talkspurt	1 sec (216 frames)
Average length of silent periods	1.35 sec (292 frames)
Average inter-arrival time between data message	Varied as 2, 3, and 5msec to get data rates of 56, 37.33, 22.4 Kbps respectively.
Average data message size	112 bits
Simulation time	130,000 frames (10 min. approx.)

Interactive

The interactive data users in the system were modeled to generate packets according to a Poisson process, with rates ranging from 22.4 to 56 Kbps. We used a Poisson model, instead of simulating a telnet, email or Web client, as we wanted to verify the efficacy of our scheme to satisfy bandwidth guarantees offered to interactive traffic. Applications like telnet, email, and web clients generate very low and bursty traffic. Since interactive traffic is not real-time, no packets are discarded due to excessive delay. We define a ratio of the Achieved Data Rate (ADR) to the Requested Data Rate (RDR) as ADR/RDR. When this ratio is "1", it indicates that the traffic source has been able to achieve the data rate that it had requested. Simulations were carried out for the following two cases:

Case 1. Only conversational traffic sources are present in the cell

Fig. 5 shows the Cumulative Distributive Function (CDF) of the average voice packet dropping probability (PDP). In order to get a good mean opinion score (MoS) for voice, this should be limited between 1and 2 %. From the figure it is clear that to achieve this goal the number of voice users must be limited to 11, as 95% of the users will experience a PDP less than 2%. Thus a multiplexing gain of 1.6 (approx.) can be obtained. Fig. 6 plots the CDF of average channel access delay faced by conversational MS's. It can be seen that for the limiting case (11 MS's), 95% of all the in-session PDTCH access attempts are satisfied within 20 msec as we had said earlier.

Fig. 7 plots the CDF of the average packet delay. Again for the limiting case, it can be observed that 95% of the packets face an average delay of less than 20 msec.

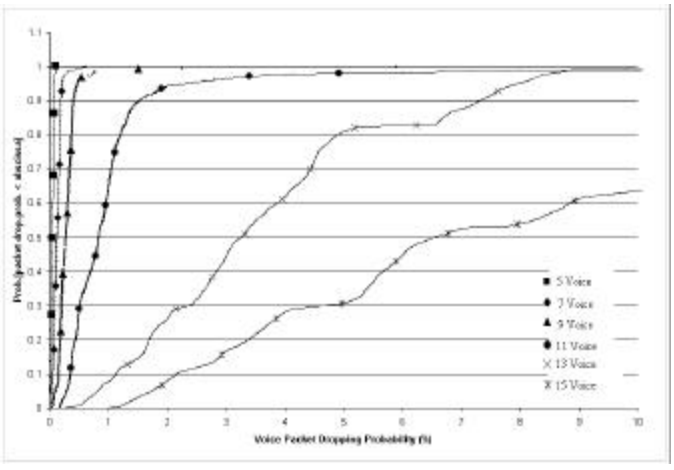

Fig. 5: CDF for voice packet dropping probability for a cell containing only conversational traffic.

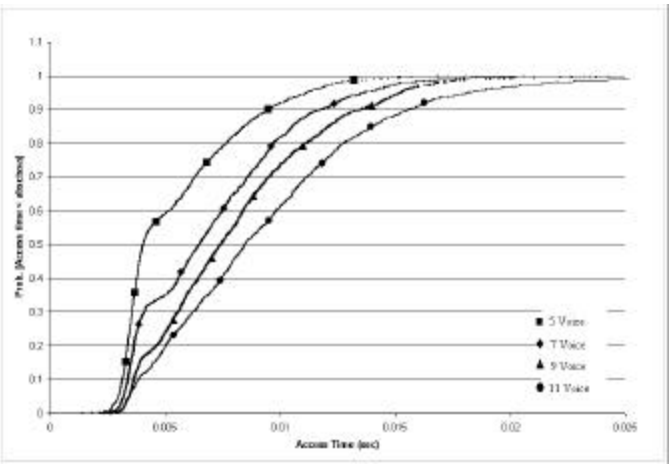

Fig. 6: CDF of average channel access time for a cell containing only conversational traffic.

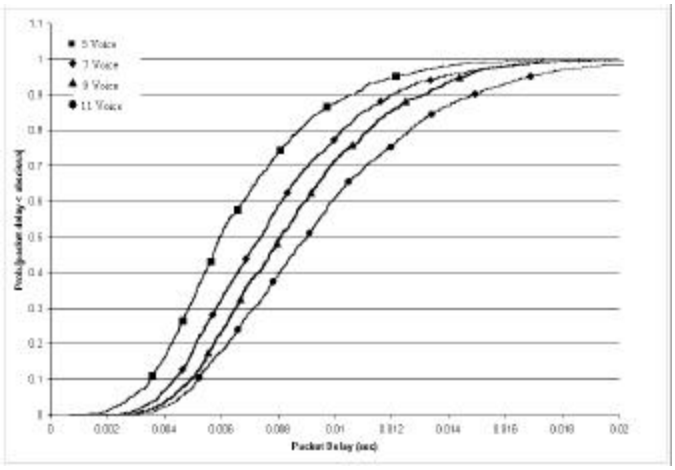

Fig. 7: CDF of average packet delay for a cell containing only conversational traffic.

Case 2. Both conversational and interactive traffic sources are present in the cell

Fig. 8 shows the CDF of the average Voice PDP, when there are 11 conversational MS's and the interactive MS's are varied from 0 to 9. It can be seen that there is not much significant change in the voice packet dropping probability. This is due to the tree protocol used by conversational mobiles for in-session channel access on the U-PRACH, which gives them sufficient isolation from others.

The CDF for the average access time is also shown in Fig. 9. This graph also shows that the in-session channel access time for conversational MS's is relatively independent of the number of interactive mobiles present in the cell.

We now demonstrate the effect of overloading on the interactive MS's. Fig. 10 shows the CDF for the ADR/RDR ratio for 11 conversational and varying interactive MS's. To guarantee throughput to interactive MS's, it is necessary that ADR/RDR ratio is unity. From Fig. 11 it can be seen that for lower loading, there is a higher probability that this condition is achieved for all interactive MS's. By admitting lesser conversational MS's, improvements are observed as illustrated in Fig. 11.

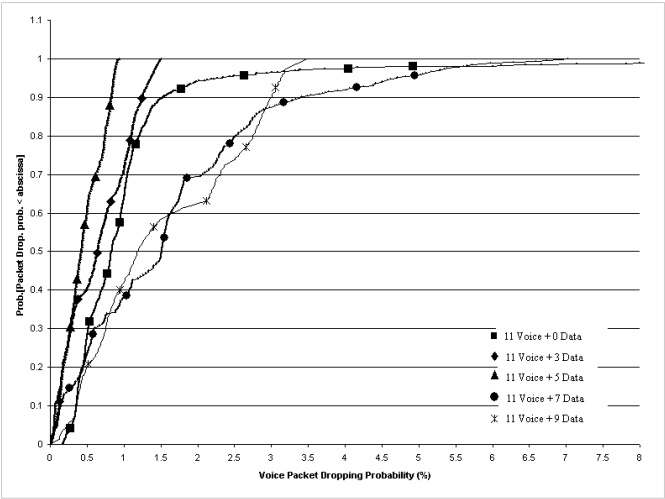

Fig. 8: CDF for voice packet dropping probability for a cell containing conversational and interactive traffic.

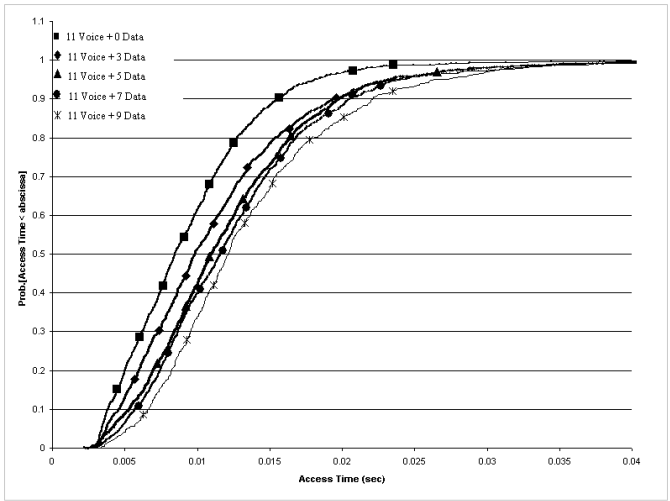

Fig. 9: CDF of average channel access time for a cell containing conversational and interactive traffic.

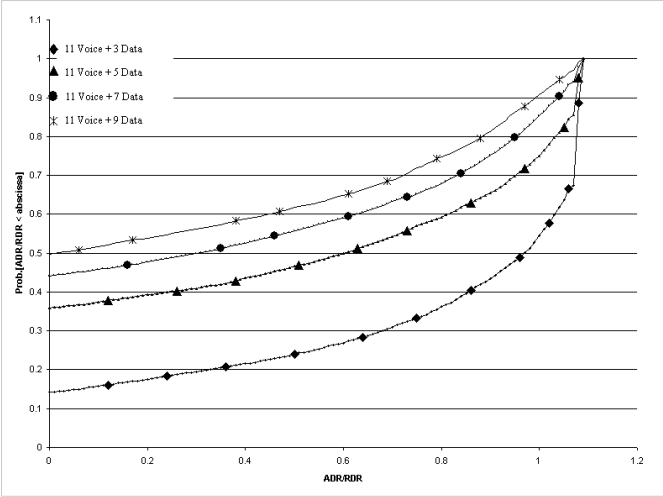

Fig. 10: CDF for ADR/RDR ratio for a cell containing conversational and interactive traffic.

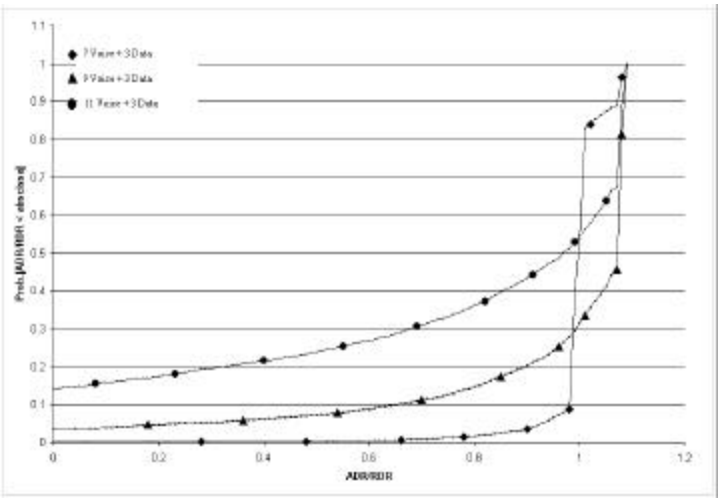

Fig. 11: CDF for ADR/RDR ratio for a cell containing lower conversational and interactive traffic.

III. AP-CAC PROTOCOL

A. Operations and procedures

Support for mobile handoff requires rerouting of traffic to a new BS through a different wireless link. Depending on the handoff scheme employed, ongoing calls could be subjected to various degrees of potential degradation, such as transient data loss, delay variability and data sequencing error. Handoff requests compete with new call requests, to gain admission into a target cell. In general, call admission control (CAC) protocol aims to maximize the number of admitted or in-session traffic sources supported, while guaranteeing their QoS requirements.

For MAC protocols that supports per-session dedicated reservation, the accompanying CAC protocol should give prioritized admission to handoff requests since disruptions of ongoing call sessions are considered more objectionable than new call blocking from the user's point of view.

For the proposed GQ-MAC protocol that supports multiple traffic classes, we propose the Adaptive Prioritized Handoff CAC (AP-CAC) protocol that differentiates handoff requests with different higher admission priorities over new calls via a multiple guard channel scheme. The proposed protocol dictates that handoff requests associated with per-session dedicated reservation (e.g., streaming traffic sources) have higher admission priority than those associated with on-demand reservation (e.g. conversational, interactive and background traffic sources).

A single guard channel scheme has been proposed and analyzed in [6][8]. According to this scheme higher priority is given to handoff requests as compared to new calls by reserving a fixed number of guard channels (N_G) for handoff calls. In other words if the number of available free channels is less than N_G, no new calls are admitted. It is shown that for stationary traffic, even for small value of N_G, handoff call blocking probability is substantially reduced at the expense of only a slight increase in the new call blocking probability. The AP-CAC protocol extends the above concept to support multiple dynamic admission priorities for handoff requests of multiple traffic QoS classes over new calls. Two levels of guard channels (N_{G1} and N_{G2}) are used to support a three-priority level admission scheme, with two premium priority levels for handoff requests over the base priority level for new calls. The three admission priority classes in ascending priority order are as follows: (1) class 'N' is associated with new calls, which are admitted only when the number of free channels exceeds N_{G1}; (2) class 'H1' is associated with hand-off requests of interactive, conversational or background traffic, which are admitted only when the number of free channels exceeds N_{G2}; (3) class 'H2' is associated with hand-off requests of streaming traffic, which are admitted whenever a free channel is available.

Each level of guard channels is continuously adapted according to the instantaneous estimate of the handoff request arrival rate of the corresponding traffic class, which depends on the number of active MS's with ongoing calls in the neighboring cells, the mobility patterns of the active MS's in terms of speed and direction during the estimation interval, the size of the cells currently resided by the active MS's and the remaining call duration of the ongoing calls.

For a given total number of channels N_T, the ranges of the time-varying $N_{G1}(t)$ and $N_{G2}(t)$ are given as: $0 = N_{G2}(t) = N_{G2max} = ß_2 N_T$; $N_{G2}(t) = N_{G1}(t) = N_{G1max} = ß_1 N_T$; where, $0 < ß_2 = ß_1 = 1$. In deriving the blocking probability for each of the three proposed admission priority classes, the arrival processes for H2 handoffs, H1 handoffs and new calls are assumed to be Poisson with time varying rates of $?_{H1}(t)$, $?_{H2}(t)$ and $?_N(t)$ respectively. The departure processes are also assumed to be Poisson with a constant rate of µ. The blocking probabilities $B_{H2}(t)$ and $B_{H1}(t)$ for H2 and H1 handoff calls, and $B_N(t)$ for new calls are given as follows:

$$B_N(t) = \sum_{j=N_T-N_{G1}(t)}^{N_T} P_j(t); \; B_{H1}(t) = \sum_{j=N_T-N_{G2}(t)}^{N_T} P_j(t); \; B_{H2}(t) = P_{N_T}(t)$$

$$\text{let } l_\Delta(t) = l_{H1}(t) + l_{H2}(t) + l_N(t); \; l_\Omega(t) = l_{H1}(t) + l_{H2}(t);$$
$$N_\Delta(t) = N_T - N_{G1}(t); \; N_\Omega(t) = N_T - N_{G2}(t);$$

$$\text{then } P_j(t)$$
$$= l_\Delta(t)^j P_0(t) / j! m^j \qquad \forall j \in \{1 \text{ to } N_\Delta(t)\}$$
$$= l_\Delta(t)^{N_\Delta(t)} l_\Omega(t)^{j-N_\Delta(t)} P_0(t) / j! m^j \qquad \forall j \in \{N_\Delta(t)+1 \text{ to } N_\Omega(t)\}$$
$$= l_\Delta(t)^{N_\Delta(t)} l_\Omega(t)^{N_{G1}(t)-N_{G2}(t)} l_{H2}(t)^{j-N_\Omega(t)} P_0(t) / j! m^j \qquad \forall j \in \{N_\Omega(t)+1 \text{ to } N_T\}$$

$$\text{with } P_0(t) = \left[\sum_{i=0}^{N_\Delta(t)} \frac{l_\Delta(t)^i}{i! m^i} + \sum_{i=N_\Delta(t)+1}^{N_\Omega(t)} \frac{l_\Delta(t)^{N_\Delta(t)} l_\Omega(t)^{i-N_\Delta(t)}}{i! m^i} + \sum_{i=N_\Omega(t)+1}^{N_T(t)} \frac{l_\Delta(t)^{N_\Delta(t)} l_\Omega(t)^{N_{G1}(t)-N_{G2}(t)} l_{H2}(t)^{i-N_\Omega(t)}}{i! m^i} \right]^{-1}$$

B. Simulation and Results

We simulated a cellular network model with concentric layers of rural, suburb and city cells over the central downtown core cell. The transit probability P_X of an active MS from one cell to another depends on the time of the day. For example, during morning rush hours, P_X is set as follows: downtown-to-city 0.167, city-to-city 0.1, city-to-downtown 0.7, city-to-suburb 0.1, suburb-to-suburb 0.167, suburb-to-city 0.5, suburb-to-rural 0.167, rural-to-rural 0.5, and rural-to-suburb 0.167.

The calls of H1 and H2 classes are assumed to have 0.8 and 0.2 occurrence probabilities respectively. We assume their unencumbered call durations to be exponential with respective means of 150s and 300s, and the call holding times to be exponential with respective means of 120s and 240s. New call arrival rates under stationary and non-stationary traffic conditions have the same long-term nominal rate of 0.475 call/second. Under non-stationary traffic condition, the average new call arrival rate during morning/afternoon rush hours increases to 0.8 call/second, and varies among 0.2, 0.3 and 0.6 call/second during other hours. The target objectives of $B_{H2} = 0.003$, $B_{H1} = 0.05$, and $B_N = 0.1$ are set given that $N_T = 100$.

The performances of the reference (multiple static guard channel) scheme and the AP-CAC (multiple dynamic guard channel) scheme are compared under non-stationary traffic condition. The reference scheme employs static guard channels of $N_{G1} = 2$ and $N_{G2} = 1$. The AP-CAC scheme would try to track the target objectives (allow 50% tolerance for the B_{H1} and B_N soft targets, but no tolerance for the B_{H2} hard target) by adapting the guard channels dynamically. The running averages of blocking probabilities for new calls (Fig. 12), H1 handoffs (Fig. 13) and H2 handoffs (Fig. 14) are measured for a 24-hours duration in a city test cell.

With the reference scheme, B_N deviates from the target objective by 40% to 0.14, B_{H1} deviates from the target objective by 25% to 0.06, and B_{H2} deviates from the target objective by 300% to 0.05. With the AP-CAC scheme, B_N deviates from the target objective by 70% to 0.17, but B_{H1} and B_{H2} meet the target objectives.

Under the non-stationary traffic condition, the results show that the AP-CAC scheme is able to maintain the handoff blocking probabilities at the target objectives, while the reference scheme fails to accomplish that. The AP-CAC scheme does cause the new call blocking probability to deviate more from the target objective as compared to the reference scheme. The reference scheme fails to achieve all the target objectives, while the AP-CAC scheme is able to meet the target objectives of the prioritized handoff blocking probabilities at the reasonable expense of the new call blocking probability.

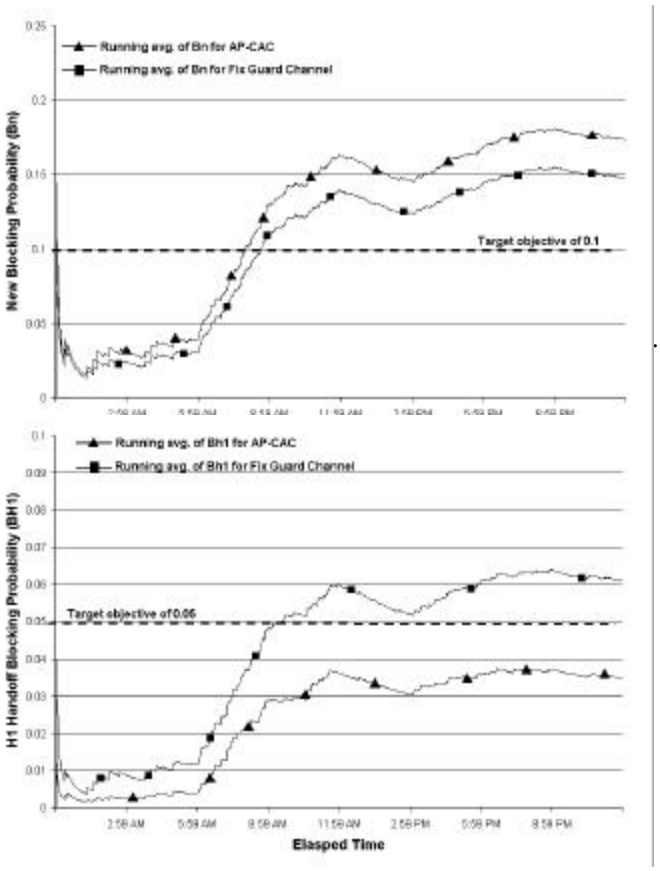

Fig. 13: Running average of H1 Handoff Blocking Probability (B_{H1}).

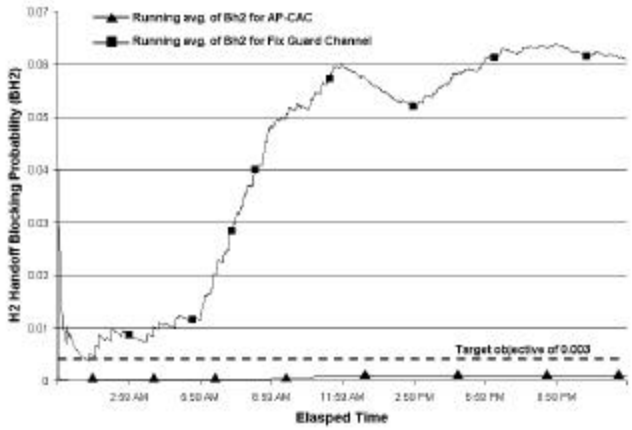

Fig. 14: Running average of H2 Handoff Blocking Probability (B_{H2}).

IV. EXTENSION OF AP-CAC OVER MPLS-BASED GPRS CORE NETWORK

New calls and handoff requests compete for bandwidth resources over a wired link in the core as well over a wireless link in the air interface. The proposed AP-CAC protocol over the wireless medium supports adaptive prioritized admissions for multiple traffic classes via the multiple dynamic guard channel scheme, which dynamically adapts the channel access capacities for handoff requests of different traffic classes based on the current number of ongoing calls in the neighboring radio cells and the mobility pattern. To extend AP-CAC to the GPRS core network, the instantaneous handoff call arrival rates

estimation should dynamically change the guard bandwidths reserved on every link within the core to support prioritized resource reservation for different traffic types. This reservation process is based on the accounting of all potential handoff connections (associated with active MS's with ongoing calls) passing through a core link, and the determination of the rerouting probabilities from the ongoing connections to the potential handoff connection. In the wireless medium, handoffs into a radio cell can only originate from the neighboring cells, and the accounting of all potential handoffs at a radio cell can be performed by monitoring the state information of the neighboring cells.

This paper presents the extension of the AP-CAC scheme over the GPRS core network by employing the proposed mobile label switched tree (MLST) to enable the equivalent accounting of potential handoffs at a core link, and multiple dynamic guard bandwidth scheme with quantized guard bandwidths reserved on individual links over the GPRS core network.

A. Mobile Label Switched Tree (MLST)

The GPRS core network is based on IP over ATM or frame relay. To unify the control planes of IP and ATM traffic engineering, MPLS path architecture and its associated traffic control functions are employed. The MPLS-based GPRS core network may form a separate MPLS domain or may be a part of a larger global Internet domain. In this paper, we assume that the GPRS core network, as illustrated in Fig. 1, forms an independent domain and the network-wide mobility support is limited to the core. The GGSN will function as an ingress label switching router (LSR), pushing labels into incoming packets, and as an egress LSR, popping labels from packets leaving this domain. The interface to the external PDN's remains the same with the MPLS deployment. The GPRS protocol stack with MPLS adaptation to layer 2 is illustrated in Fig. 15.

Similar to the mobile virtual circuit (MVC) [18], the mobile label switched tree (MLST) is an MPLS-based path architecture to support terminal mobility. It is a dynamic connection tree which supports network-wide terminal mobility across radio cells connected to any points of the core network while allowing sharing of network resources between mobile and non-mobile traffic, maximizing connection reuse, and minimizing label switching table updates.

As illustrated in Fig. 16, for an MS accessing a BS in a given radio cell, the neighboring cells terminate the endpoints of the potential handoff mobile paths. The group of current-in-use and potential-handoff BS's for a MT, and the corresponding peer terminal forms an extended multipoint-to-point MLST, with the current-in-use and the potential-handoff connections converged at the mobile merged point (MMP). After a handoff, the MMP may shift to another node as the MLST is reconfigured.

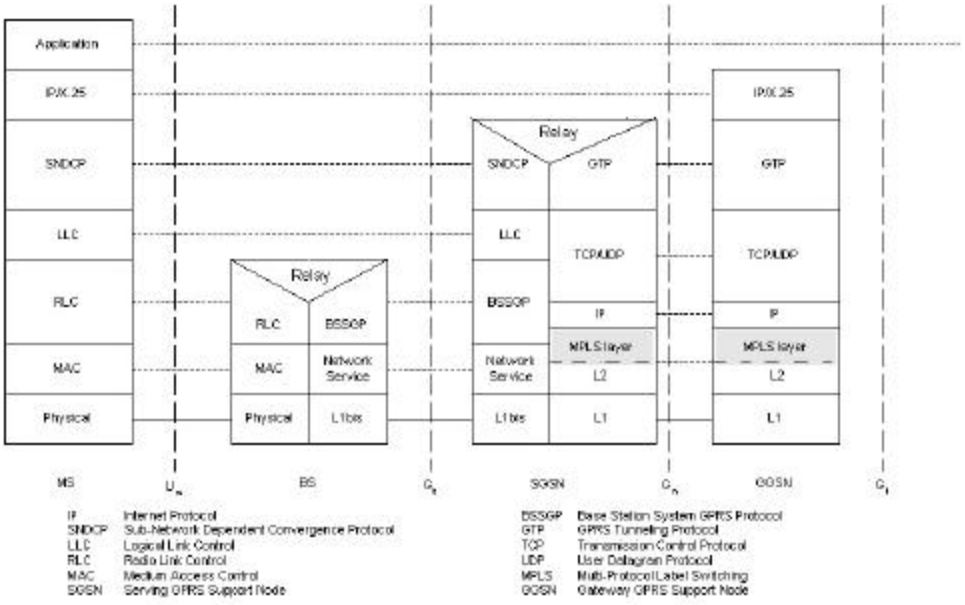

Fig. 15: Modified GPRS protocol stack with MPLS.

The MLST allows a compromise to minimize handoff delay while avoiding unnecessary resource allocation. The routes for potential handoff connections are determined at call setup time but resources are not reserved along those routes, a fast resource reservation scheme would be employed for completing the establishment of the selected handoff connection during actual call handoff. The dynamic guard channel scheme over the wireless medium is extended over the GPRS core network (as

the equivalent dynamic guard bandwidth scheme) by enhancing the MLST connection control services. At the time of MLST connection establishment, disconnection or reconfiguration, the potential handoff connection setup or release at each mobile-specific link is interpreted respectively as an increase or decrease in the number of potential handoff calls for the respective link, and propagating the estimated probability of handoff call arrival as derived from the MS mobility pattern over each link set during setup. Consequently, the dynamic guard bandwidth controller in each node would update the estimated instantaneous handoff call arrival rates at the core link as discussed below.

B. Dynamic Guard Bandwidth Controller

An active MS, situated at the cell terminating the current ongoing connection of an MLST, may generate a handoff request to the neighboring cells terminating the potential handoff paths of the MLST. Consequently, the probability of handoff request initiated by the active MS to the core links along a potential handoff path is determined by the corresponding handoff probability between the respective cells. The probability of a handoff request arrival at a cell (generated by an active MS in a neighboring cell during an estimation interval) depends on its mobility pattern, the cell size, the remaining call duration, and the length of the estimation interval.

By employing MLST to support connection rerouting during call handoff, the number of active MS's that can initiate handoff requests to a GPRS core link is limited by the number of potential handoff paths of the MLST's passing through it. For example in Fig. 16, the current in-use connection of MLST-2 and the two potential handoff connections of MLST's 1 and 3 pass through the "link X-Y", through which bandwidth is reserved for MLST-2 only. The MS's associated with MLST's 1 and 3 have the potential to generate handoff requests to the "link X-Y", thus determine the instantaneous handoff call arrival rate at the link. Consequently, the guard bandwidths for handoff requests and new calls will be dynamically adapted to the instantaneous handoff call arrival rate.

The continuous update of the instantaneous handoff call arrival rate at a core link is enabled by the signaling of changes in the number of ongoing calls that may hand off to the core link. The signaling should occur during potential handoff path setup and release, associated respectively with the establishment and release/reconfiguration of a MLST. We now discuss the strategy of integrating the dynamic guard bandwidth management functions into the potential handoff path control services associated with the MLST.

In a general connection-oriented packet-switched network, connection establishment involves the following steps: (1) reservation of a logical link identifier at each switch associated with the connection; (2) establishment of routing information for translating incoming logical link identifiers into outgoing logical link identifiers; and (3) reservation of communication resources (buffer and physical link bandwidth) at each switch. Similar steps are employed in MPLS path setup, with logical links identified by labels, which are allocated and reserved via the label distribution protocol (LDP).

The MLST establishment scheme is based on reserving logical link identifiers via LDP for potential handoff paths, but delaying resource reservation until handoff processing, when a fast hard-state resource reservation scheme is invoked to complete the setup of the handoff connection. The establishment of the MLST during initial call processing is decomposed into the following tasks:

1. Establishment of the fixed common connecting links shared by the original path and the potential handoff paths:
   - Standard Label Switched Path (LSP) is established along the fixed common connecting links. Incoming and outgoing labels are updated in the Label Information Base (LIB) in each intermediate LSR along the fixed common connecting links.
   - Communication resources are reserved to satisfy path QoS performance requirements.

2. Establishment of the mobile link sets between the multiple BS's and the MMP:
   - Standard LSP establishment between each BS and the MMP.
   - Bandwidth is reserved only for the current in-use path from the MMP to the BS.
   - For those LSP's associated with the mobile link sets specific to the potential handoff connections, actual bandwidth is not reserved immediately; instead, the guard bandwidth is increased according to the probability that the associated ongoing call will hand off to the core link. A fast bandwidth reservation scheme will be employed to reserve actual bandwidth if the anticipated handoff does occur.

The dynamic guard bandwidth management functions can be incorporated as follows. During MLST setup, the MMP would signal an increase in the number of potential handoff calls along the predetermined LSP's to enable updates of instantaneous handoff call arrival rate at the core links along the predetermined routes, so that the guard bandwidths can be increased accordingly by the dynamic guard bandwidth controllers of the core links. During MLST release, the MMP would signal a decrease in the number of potential handoff calls along the predetermined routes to enable updates of the instantaneous handoff call arrival rate at the wired links along the predetermined routes, so that the guard bandwidths can be decreased accordingly by the dynamic guard bandwidth controllers of the core links. To support successive handoffs so as to allow an MS to have an

unrestricted range of movement or network wide terminal mobility, the MLST connection tree is reconfigured after each handoff to account for the new set of immediate neighboring cells into which the MS can potentially enter.

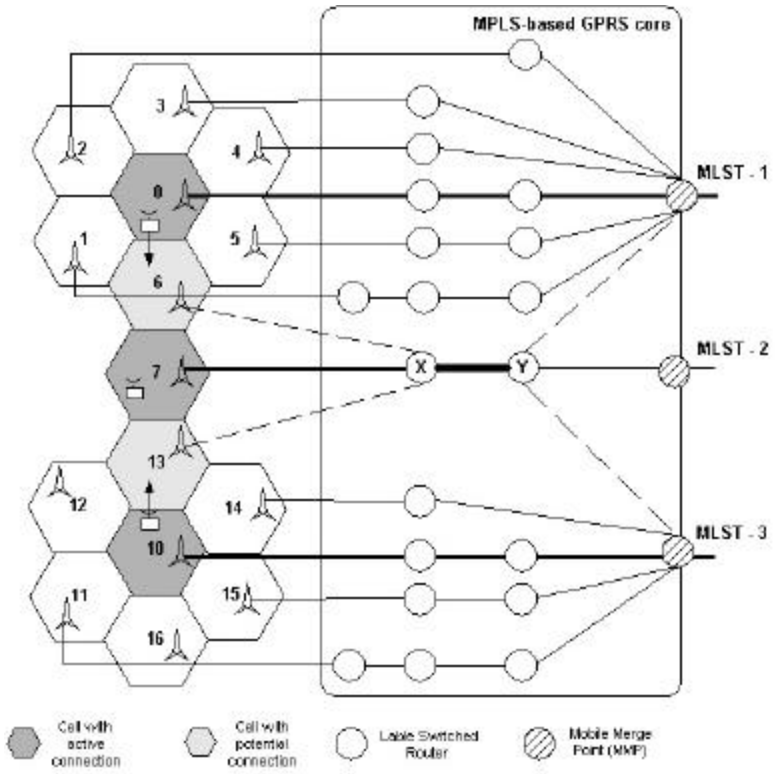

Fig. 16: Estimation of Instantaneous handoff call arrival rate at a wired link with MLST.

V. CONCLUSION

Performance analysis of the GQ-MAC protocol shows that it is capable of providing guaranteed QoS performances for the specified QoS profiles (streaming, conversational, interactive and background) over GPRS wireless links while optimizing channel resource utilization. Performance analysis of the AP-CAC protocol shows that it is capable of maintaining QoS performance guarantees under the effect of mobile handoffs by providing adaptive prioritized admission control for multiple traffic classes via the multiple dynamic guard channel scheme, which dynamically adapts the capacity reserved for dealing with handoff requests based on the current number of ongoing calls in the neighboring radio cells and the mobility pattern. The results show that hard target handoff blocking probability for a high admission priority traffic class can be maintained while minimizing the degradation to the soft target blocking probabilities of lower admission priority traffic classes. QoS provisioning over the MPLS-based GPRS core network could be realized via the IntServ QoS architecture, and mobility support over the core network could by realized by the mobile label switching tree (MLST) to optimize handoff processing delay and resource utilization. This paper shows that it is feasible to extend the AP-CAC protocol from the wireless medium to the GPRS core network by employing the MLST connection architecture and the dynamic guard bandwidth scheme. This paper presents a strategy to realize end-to-end dynamic adaptive QoS provisioning over the GPRS wireless mobile network. The strategy is based on mapping between the IntServ QoS over the core and the GPRS QoS over the wireless medium maintained by the GQ-MAC protocol, and by extending the AP-CAC protocol from the wireless medium to the core network to provide a unified end-to-end admission control with dynamic adaptive admission priorities.

REFERENCES

[1] GSM 04.60: "Digital cellular telecommunications system (Phase 2+); General Packet Radio Service (GPRS); Mobile Station (MS) – Base Station System (BSS) interface; Radio Link Control / Medium Access Control (RLC/MAC) protocol," version 7.4.1, Release 1998.

[2] D. J. Goodman, R.A. Valenzuela, K.T. Gayliard and B. Ramamurthi, "Packet Reservation Multiple Access for local wireless Communication," *IEEE Tran. on Comm.*, Aug'89.

[3] G. Bianchi, F. Borgonovo, L. Fratta, L. Musumeci and M. Zorzi, "C-PRAMA: A Centralized Packet Reservation Multiple Access for Local Wireless Communications," *IEEE Trans. on Veh. Technol.*, vol. 46(2), pp. 422-436, May 97.

[4] W. C. Wong and D. J. Goodman, "A Packet Reservation Multiple Access Protocol for Integrated Speech and Data Transmission," *IEEE Proceedings – I*, vol. 139(6) Dec. 92.

[5] M. J. Karol, Z. Liu and K. Y. Eng, "Distributed-Queuing Request Update Multiple Access (DQRUMA) for wirelss packet (ATM) networks," *Proc. ICC'95*, pp.1224-1231, June 95.

[6] D. Hong and S.S. Rappaport, "Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff Procedures," *IEEE Trans. on Veh. Technol.*, vol. 3, pp. 77-92, Aug. 1986.

[7] C. H. Yoon and C. K. Un, "Performance of Personal Portable Radio Telephone Systems with and without Guard Channels," *IEEE J. Select. Areas Commun.*, vol. 11, pp. 911-917, Aug. 1993.

[8] O. T. W. Yu and V. C. M. Leung, "Adaptive Resource Allocation for prioritized call admission over an ATM-based Wireless PCN," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 1208-1224, Sept. 1997.

[9] P. Ramanathan, K. M. Sivalingam, P. Agrawal and S. Kishore, "Dynamic Resource Allocation Schemes During Handoff for Mobile Multimedia Wireless Networks," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1270-1283, July 1999.

[10] R. Braden, C. Clark, S. Shenker, "Integrated Services in the Internet Architecture: An Overview," *RFC 1633*, June 1994.

[11] S. Shenker, J. Wroclawski, "General Characterization Parameters for Integrated Service Network Elements," *RFC 2215*, Sep. 1997.

[12] R. Callon, P. Doolan, N. Feldman, A. Fredette, G. Swallow, "A Framework for Multi-protocol Label Switching," *IETF Internet Draft*, November 1997.

[13] E. Rosen, A. Viswannthan, R. Callon, "Multi-protocol Label Switching Architecture," *IETF Internet Draft*, July 1997.

[14] S. Shenker, C. Patridge, R. Guerin, "Specification of Guaranteed Quality of Service," *RFC 2212*.

[15] J. Wroclawski, "Specification of the Controlled-Load Network Element Service," *RFC 2211*.

[16] R. Braden, L. Zhang, S. Berson, S. Herzog, S.Jamin, "Resource ReSerVation Protocol (RSVP) - Version 1 Functional Specification," *IETF RFC2205*, September 1997.

[17] G. Priggouris, S. Hadjiefthymiades, L. Merakos, "Support IP QoS in the General Packet Radio Service," *IEEE Network Magazine*, pp.9- 17, September/October 2000.

[18] O. T. W. Yu and V. C. M. Leung, "Connection architecture and protocols to support efficient handoffs over an ATM/B-ISDN personal communications network," *ACM/Baltzer J. Mobile Networks & Appl.*, vol. 1, pp. 123-139, Oct. 1996.

[19] J. Capetanakis, "Tree Algorithms for Packet Broadcast Channel," *IEEE Trans. on Info. Theory*, Sept. 79.

[20] D. Bertsekas and R. Gallager, *Data Networks, 3rd ed.*, Prentice Hall, 1989.

[21] D. Dyson and Z. Haas, "A dynamic packet reservation multiple access scheme for wireless ATM," *ACM/Baltzer J. Mobile Networks & Appl.*, vol. 4, pp. 87-99, 1999.