QoS Provisioning over GPRS Wireless Mobile Links

Oliver Yu and Shashank Khanvilkar
Department of ECE, University of Illinois at Chicago
851 S. Morgan Street, 1020 SEO, Chicago, IL 60607

***Abstract*-The General Packet Radio Service (GPRS) offers performance guaranteed packet data services to mobile users. This paper presents a 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 prioritized handoff call admission control (PH-CAC) protocol to maintain 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. The PH-CAC protocol provides prioritized admission by differentiating handoff requests with different higher admission priorities over new calls via a multiple guard channels scheme.**

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 mobile stations (MS's) contending for access to a base station (BS) in a radio cell, with traffic generated according to the negotiated quality-of-service (QoS) profiles, defined in terms of precedence, delay, reliability, mean and peak throughputs. In this paper, the QoS profiles are classified as streaming, conversational, interactive and background, and their attributive values are summarized in TABLE I.

The wireless link is characterized by a broadcast mode in the downlink (BS to MS) and a multiple access mode in the uplink (MS to BS). A medium access control (MAC) protocol distributes packet transmission over the shared medium among all users. The GPRS standard [1] specifies the FDD/TDMA multiple access with four radio access priorities, and reference guidelines on the resource sharing method.

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.

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 maximizes the number of admitted or in-session traffic sources supported over the wireless medium while guaranteeing their QoS requirements. Forced terminations of ongoing call sessions due to blocked-calls-dropped discipline are generally more objectionable than new calls blocking from the user's perspective. One common prioritized admission scheme is the guard channel scheme [5][6], which gives a higher access priority to handoff requests over new calls by assigning them a higher capacity limit.

TABLE I: QOS PROFILE

Traffic Classes	*Latency*	*Jitter*	*Loss*	*Throughput*	*Burstiness*
Streaming	Bounded (<500 msec)	Stringent	Tolerable ($<10^{-2}$)	Guaranteed	Low
Conversational	Bounded (<80 msec)	Stringent	Tolerable ($<10^{-2}$)	N/A	High
Interactive	Less than Conversational and Streaming	N/A	Loss sensitive ($<10^{-5}$)	Guaranteed	Higher than Conversational
Background	N/A	N/A	N/A	N/A	N/A

For the proposed GQ-MAC protocol, this paper proposes the Prioritized Handoff CAC (PH-CAC) protocol that differentiates handoff requests with different higher admission priorities over new calls via a multiple guard channels scheme. Handoff requests associated with per-session dedicated reservation have higher admission priority than those associated with on-demand reservation.

This paper is organized as follows. Section II and III describe the GQ-MAC and PH-CAC protocols in terms of their features, implementation, performance analysis and results. Section IV concludes the paper.

II. GQ-MAC PROTOCOL

A. Channel Access Procedures

The GPRS packet data channels (PDCH) are classified as follows: (1) *Packet Random Access Channel (PRACH)* is the

0-7803-7376-6/02/$17.00 (c) 2002 IEEE.

request access channel for uplink, consisting of two time-multiplexed channels of Signaling PRACH (S-PRACH) and User-data PRACH (U-PRACH); (2) *Packet Access Grant Channel (PAGCH)* is the request acknowledgement channel for downlink, which is used by the BS to broadcast the request status information; and (3) *Packet Data Traffic Channels (PDTCH)* are the remaining PDCH's to carry the payload on the uplink and downlink.

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	

The request access priorities of the S-PRACH and U-PRACH are illustrated in TABLE II. Using slotted Aloha, the S-PRACH is accessed by the signaling requests of new calls and handoff calls to gain admission, with handoff calls given higher access priority. Admitted streaming traffic sources enjoy per-session dedicated reservations, while admitted conversational and interactive traffic sources have to perform on-demand reservation by multi-accessing the U-PRACH. When collision occurs, only the conversational traffic sources are allowed to multi-access the U-PRACH via the tree limited-contention access protocol. Consequently, conversational class is given a higher U-PRACH access priority than interactive class. The background traffic sources are allocated with unused PDTCH's in a round robin fashion. The access protocols employed by the GQ-MAC are described as follows.

A.1. Slotted Aloha

This is used by the signaling requests of new calls and handoffs to access the S-PRACH. MS's 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.

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. 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 * $(2^{(\log_2 j + 1)} - 1)$ [7]; Where j is the number of conversational MS's, simultaneously trying to access the U-PRACH. Based on the on-off model for voice, with 40% voice activity, it can be shown that the probability of more than 5 MS's trying to access the U-PRACH simultaneously is very low, thus limiting the in-session channel access delay to 20 msec (assuming GSM frame size).

A.3. Modified Slotted Aloha

This is used for U-PRACH access by interactive traffic sources. It is similar to slotted Aloha protocol, except that when a collision occurs, a binary exponential backoff algorithm is used which reduces the transmission probability, "P_a", in the next slot by 0.5 (for 8 max. retries). When a contention resolution cycle is in progress, the value of "P_a" is reduced to "0", i.e. interactive MS's do not participate in the contention resolution cycle.

B. QoS Support

To initiate a new call, a Call Initiation (Call_Init) request is sent on the S-PRACH using slotted Aloha. 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 PH-CAC module determines if enough resources are available to support this new call, and if admissible, sets up two state variables for that session in the BS via 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 the corresponding 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. 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

Upon admission, a conversational traffic source 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. Since resource is reserved till it is explicitly released; the 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 a distributed scheduling algorithm for allocation of uplink PDTCH's, by modifying the algorithm in [8]. In the proposed 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: RDR which is sent in the Call-Init request packet; and ADR which 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 BS attempts to allocate as many PDTCH's requested as available. For this, it might even pre-empt other similar sources, which have achieved ADR >= RDR.

If the MS is allocated a rate, which is less than its packet generation rate, packets will start queuing at the output and the queue length increases. Depending on the queue length the MS will send another RRP, demanding higher rate, with access probability P(x) as:

$$P(x) = e^{\frac{(x/L_U)-1}{(1-\alpha)}} \quad \forall x \in \{1..L_U - 1\};\ and\ P(x) = 1 \quad \forall x > L_U$$

Here 'L_U' is a fixed upper threshold and 'α' is a factor that is directly proportional to the ratio ADR/RDR and changes the P(x) curve as shown in Fig. 1. This mechanism gives higher priority to MS's that have not been able to achieve the requested data rate to send a RRP.

B.4. Background Traffic

After getting admitted and allocated an identifier, the background traffic sources 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, using the OPNET network simulation tool. TABLE III lists the simulation parameters. Streaming and background traffic types were not simulated because streaming calls with dedicated reservations 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 the results. Simulation experiments were carried out for conversational and interactive traffic types using the following traffic models:

- *Conversational*

An on-off model is used to simulate the conversational source. Voice packets are dropped if they exceed a limit of 60 msec. Packet dropping probability is defined as the ratio of packets dropped, to the total number of packets generated during the call. An ideal wireless channel is assumed, hence packet loss occurs only because of deadline violation.

- *Interactive*

The interactive data users generate packets according to a Poisson distribution, with rates ranging from 22.4 to 56 Kbps. Since interactive traffic is not real-time, no packets are discarded due to excessive delay. We carried out simulations for two different cases as follows.

TABLE III: SIMULATION PARAMETERS

Parameter	Value
No. of Uplink/Downlink carrier pairs	1
TDMA frame duration	4.615 msec
No. of time slots in a TDMA frame	8
No. 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)
Avg. msg. inter-arrival time.	Varied as 2, 3, and 5 msec
Data Rate	56, 37.33, 22.4 Kbps
Average data message size	112 bits
Simulation time	10 minutes (13K frames)

C.1. Conversational traffic only

With only conversational traffic sources present in the cell, it was found that in order to get a good mean opinion score, the number of admitted users must be limited to 11, as 95% of the users experienced a packet dropping probability (PDP) of less than 2%. Thus a multiplexing gain of 1.6 (approx.) was obtained. Also the average channel access delay and average packet delay was found to be limited to 20 msec as predicted before.

C.2. Integrated Conversational and Interactive traffic

Fig. 2 shows the cumulative distribution function (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 no significant change in the voice PDP. 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. 3. This graph 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. Fig. 4 illustrates the CDF of the average packet delay faced by conversational mobiles. As observed, this is also limited to 20 msec and is also independent of number of interactive mobiles present in the cell.

III. PH-CAC PROTOCOL

A. Operations and procedure

Handoff requests compete with new call requests to gain admission. In general, call admission control (CAC) protocol aims to maximize the number of admitted traffic sources while guaranteeing their QoS requirements. For MAC protocols that support per-session dedicated reservation, the accompanying CAC protocol should give prioritized admission to handoff requests. For the proposed GQ-MAC protocol that supports per-session dedicated reservation and on-demand reservation, this paper proposes the Prioritized Handoff CAC (PH-CAC) protocol that differentiates handoff requests with different higher admission priorities over new calls. 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).

PH-CAC extends the guard channel concept to support multiple admission priorities for handoff requests associated with different traffic classes. Higher admission priority is given to a handoff request class as compared to new calls by dynamically adapting a number of guard channels for that handoff class. Here two guard channels (N_{G1} and N_{G2}) are used to support a three-priority level admission scheme, with two handoff priority classes over the *new call* base priority class. 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 calls involved in interactive, conversational or background session, which are admitted only when the number of free channels exceeds N_{G2}; (3) class 'H2' is associated with hand-off calls involved in streaming, which are admitted whenever a free channel is available. The relation between N_{G1}, N_{G2} and N_T (total number channels) is: $0 \le N_{G2} = \beta_2 N_T \le N_{G2max}$; $N_{G2} \le N_{G1} = \beta_1 N_T \le N_{G1max}$; $\beta_2 \le \beta_1 \le 1$.

To derive the blocking probability for each of the three admission priority classes (H1, H2 and N), we assume the arrival process for H1 handoff's, H2 handoff's and new calls to be Poisson with rates λ_{H1}, λ_{H2} and λ_N respectively. The call departure process is also assumed to be Poisson with a rate of μ. The blocking probabilities for H1 (B_{H1}), H2 (B_{H2}) and new calls (B_N) can thus be calculated as

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

$$let\ \lambda_\Delta = \lambda_{H1} + \lambda_{H2} + \lambda_N ;\ \lambda_\Omega = \lambda_{H1} + \lambda_{H2} ;\ N_\Delta = N_T - N_{G1} ;\ N_\Omega = N_T - N_{G2} ;\ then$$

$$\begin{aligned} P_J &= {\lambda_\Delta}^j P_0 / j!\mu^j & \forall j \in \{1\ to\ N_\Delta\} \\ &= {\lambda_\Delta}^{N_\Delta} {\lambda_\Omega}^{j-N_\Delta} P_0 / j!\mu^j & \forall j \in \{N_\Delta + 1\ to\ N_\Omega\} \\ &= {\lambda_\Delta}^{N_\Delta} {\lambda_\Omega}^{N_{G1}-N_{G2}} {\lambda_{H2}}^{j-N_\Omega} P_0 / j!\mu^j & \forall j \in \{N_\Omega + 1\ to\ N_T\} \end{aligned}$$

$$where\ P_0 = \left[\sum_{i=0}^{N_\Delta} \frac{{\lambda_\Delta}^i}{i!\mu^i} + \sum_{i=N_\Delta+1}^{N_\Omega} \frac{{\lambda_\Delta}^{N_\Delta} {\lambda_\Omega}^{i-N_\Delta}}{i!\mu^i} + \sum_{i=N_\Omega+1}^{N_T} \frac{{\lambda_\Delta}^{N_\Delta} {\lambda_\Omega}^{N_{G1}-N_{G2}} {\lambda_{H2}}^{i-N_\Omega}}{i!\mu^i} \right]^{-1}$$

B. Simulation and Results

Simulations were carried out to test how the guard channels (N_{G1} and N_{G2}) are adaptively varied to satisfy targeted values of B_{H1} and B_{H2} under stationary traffic conditions with following system parameters: new call arrival rate is exponential with a mean λ_N (0.05, 0.1 calls/sec); H1 and H2 handoff arrival rates are exponential with means λ_{H1} ($0 \le \lambda_{H1} \le 0.25$) and λ_{H2} ($0 \le \lambda_{H2} \le 0.25$) respectively; channel holding time is also exponential with mean T_H = 120s ($\approx 1/\mu$); total number of channels (N_T) is 50 and the adapted number of guard channels at any estimation interval k, is set within the range $0 \le N_{G1} \le N_{Gmax}$ ($= N_T - 1$), $0 \le N_{G2} \le N_{G1}$.

We set a hard target for one of the handoff blocking probabilities ($B_{H1} \le 0.003$ say), and calculated the other (referred to as soft target), as well as B_N, for different values of λ_{H1}, λ_{H2} and λ_N. We further comment on the deviation of the handoff blocking probability from its soft target and degradation in the new call blocking probability.

For the given system configuration, B_{H2} (Fig. 5) is maintained below its target value by increasing the number of guard channels (N_{G2}). For example, consider the 'λ_N = 0.1, λ_{H1} = 0.05' curve (represented by upward facing triangles) in Fig. 5, Fig. 6 and Fig. 7 at the points 'A' and 'B' as marked. It can be seen that for an increase in λ_{H2} from 0.16 at 'A' to 0.17 at 'B', the AP-CAC module adapts to this change by increasing the guard threshold, N_{G2} (not shown), from 1 to 2, thereby reducing B_{H2} (Fig. 5) from 2.75×10^{-3} to 2.3×10^{-3}. This increase in N_{G2} has only minimal effect on B_N and B_{H1}, which increase slightly (Fig. 6 and Fig. 7) from their original values.

Degradations are observed in B_N and B_{H1} towards higher values of λ_{H2}, as the system approaches saturation. TABLE IV summarizes the range of λ_{H2}, for which the degradations in B_N and deviation of B_{H1} from its target value is insignificant.

TABLE IV: RANGE FOR λ_{H2} TO AVOID SYSTEM SATURATION.

λ_N	λ_{H1}	λ_{H2} range for which degradation in B_N is insignificant	λ_{H2} range for which deviation in B_{H1} from its soft target is insignificant
0.05	0.02	0-0.24	0-0.23
	0.05	0-0.23	0-0.2
	0.08	0-0.19	0-0.17
0.1	0.02	0-0.21	0-0.17
	0.05	0-0.17	0-0.16
	0.08	0-0.14	0-0.14

IV. CONCLUSION

The GQ-MAC protocol was analyzed and its performance was evaluated. The presented results show that this protocol is able to support guaranteed QoS for streaming, conversational, interactive and background traffic. Since these traffic classes are similar to those supported in UMTS, we are studying its application to future 3G cellular systems employing WCDMA/TDD mode on the wireless links. The PH-CAC protocol, for dynamic adaptation of call blocking probability for different traffic types, appears to be a promising approach, which can be further explored. Simulations, testing the feasibility of this approach, show that the PH-CAC scheme does provides a differentiated admission control mechanism for different traffic types.

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. Tech.*, 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 Tranmission", *IEEE Proceedings – I*, vol. 139(6) Dec'92.

[5] 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. Tech., Aug'86, vol. VT- 35(3), pp. 77-92.

[6] 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(7), pp: 1208-1224, Sept'1997.

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

[8] D. Dyson and Z. Haas, "A dynamic packet reservation multiple access scheme for wireless ATM", *MONET*, vol. 4, pp: 97-99, 1999.

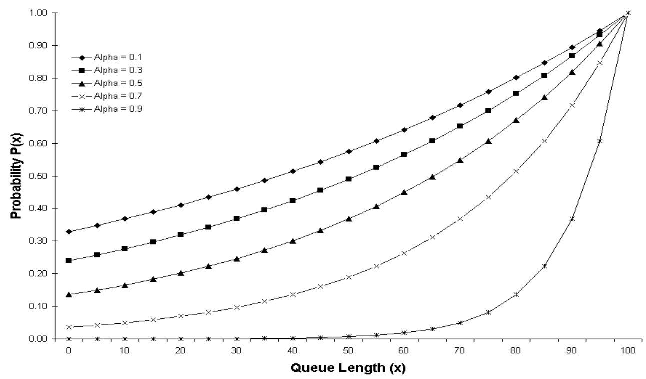

Fig. 1. Access Probability V/s Queue length for different values of "α".

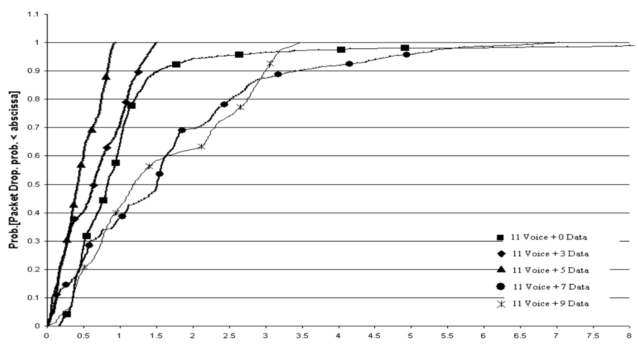

Fig. 2. CDF for Voice Packet Dropping Probability.

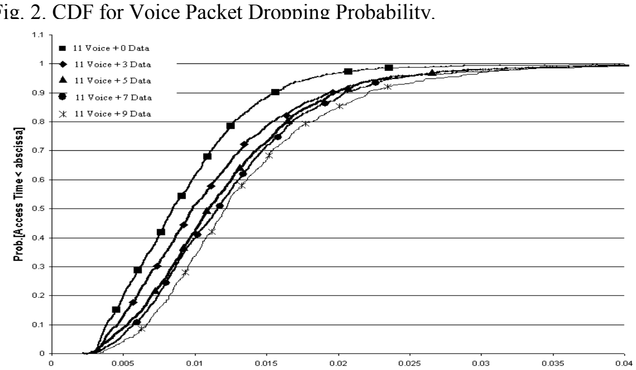

Fig. 3. CDF of average channel access time.

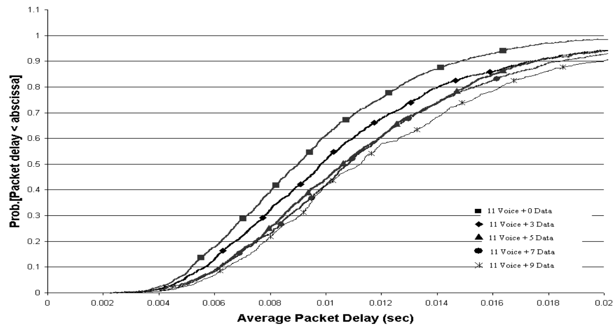

Fig. 4. CDF of average packet delay.

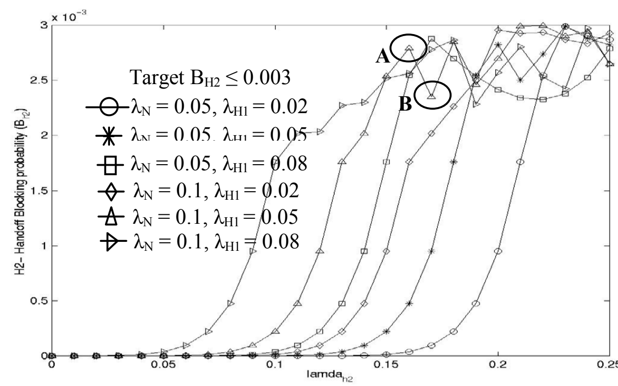

Fig. 5. H2 handoff blocking probability (B_{H2}).

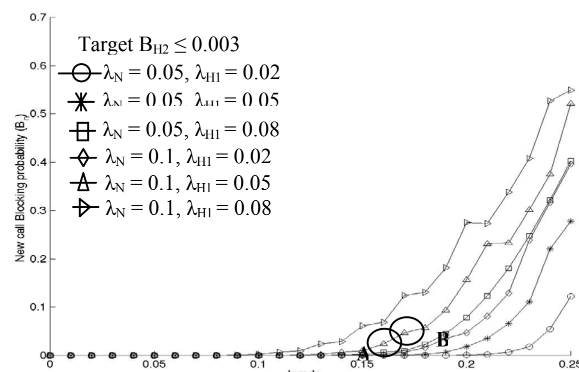

Fig. 6. New call blocking probability (B_N).

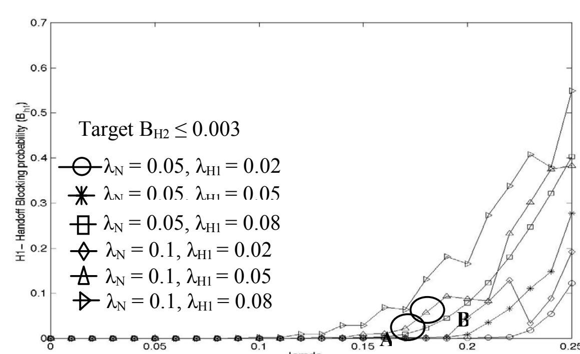

Fig. 7. H1 handoff blocking probability (B_{H1}).