

A Call Admission Control Algorithm Based on Stochastic Performance Bound for Wireless Networks*

Wei-jen Hsu¹ and Zsehong Tsai²

¹ Chunghwa Telecom
wjhsu@cht.com.tw

² Department of Electrical Engineering, National Taiwan University
ztsai@cc.ee.ntu.edu.tw

Abstract. In this paper, we derive stochastic performance bounds under the assumption of exponentially bounded burstiness (EBB) traffic model and exponentially bounded fluctuation (EBF) channel model. Then we propose a measurement-based call admission algorithm providing statistical service level agreement (SLA) guarantee for accepted flows based on the QoS prediction equations for both single and multiple priority services. Our call admission control algorithm is characterized by tunable tradeoff between channel utilization and SLA violation probability.

1 Introduction

Wireless access technologies have become a competitive solution for the access network in recent years. Unfortunately, high packet error rate and sporadic service outage due to channel impairments have been a challenge for network engineers to deploy wireless network with QoS guarantee or to provide satisfactory streaming media services. Thus, systematic approaches to provide QoS guaranteed service on error-prone wireless channels have become an important research issue.

Most performance bounds currently available in literature can be classified into two broad categories, namely deterministic bound [2] and stochastic bound [1]. Directly providing deterministic QoS guarantee in the wireless environment is either infeasible or can be with extremely high cost. The performance bound we seek in wireless environment falls in the category of probabilistic forms or the so-called stochastic bound.

We believe that stochastic bounds fit better in terms of theoretical tightness and validity in the wireless environment. Its applicability in the call admission control is also better. (We use the term “call admission control” for the mechanism deciding whether we accept a new traffic flow.) When deterministic bounds are used as call admission criteria, the system utilization is usually lower than that if stochastic bounds are used. To provide acceptable quality of multimedia service to users, a small probability of SLA violation events may be tolerable, thus providing stochastic bound is sufficient.

Although measurement-based admission control has been largely available in the literature [4][5][6], we propose one characterized by low operation overhead. In addi-

* Most of this work is done when W. Hsu was with National Taiwan University.
This work is partially sponsored by MOE under grant 89E-FA06-2-4-7.

tion, not many of these previous works on measurement-based admission control include the discussion about wireless environment, as we emphasize in this work.

This paper is organized as follows: Stochastic performance bounds in wireless network under FCFS and prioritized access queueing disciplines are derived in section 2. Based on these bounds, we proposed a measurement-based call admission algorithm for wireless access networks in section 3. Simulation results are given in section 4. We conclude the paper in section 5.

2 Performance Bounds in Wireless Networks

2.1 Network Model

We consider the network environment in which end terminals access the Internet through a shared wireless channel, as illustrated in Fig 1. In such an environment, the shared wireless channel serves as a substitute for point-to-point wire link.

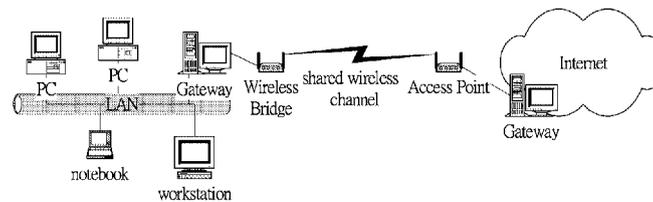


Fig. 1. Illustration of shared wireless channel

We consider two different scheduling algorithms: the FCFS (First Come First Serve) queueing discipline and prioritized access. We derive stochastic bounds for the corresponding queue size and queueing delay for each flow, under either service discipline.

In this paper, we adopt the EBB source model and EBF channel model. Please refer to [1] for its definition and notation. Most traffic models used in modeling data sources, such as IPP sources, MMPP sources, or on-off sources, can be substituted by EBB sources after choosing appropriate pre-factor and decay-factor. Detailed mathematical treatment of such transform can be found in [3].

As well as EBB sources can be used to substitute most source models, EBF channel can be used to substitute most channel models, such as on-off channel, finite state Markov channel, and channel with non-zero, time-varying packet loss probability. Hence, we use it to model the error prone wireless channel. Interested readers can refer [8] for detailed discussions.

2.2 Performance Bounds under FCFS Queueing Discipline

We now consider the FCFS queueing model with traffic sources modeled by $(\lambda_i, A_i, \alpha_i)$ -EBB processes, where i is the index of source. The channel is modeled as $(\mu - \varepsilon, B, \beta)$ -EBF, where μ is the ideal channel rate, ε is the error rate.

Theorem1. Queue size process $Q(t)$ of the shared FIFO queue is upper bounded by an exponentially bounded process (EB process)

$$Q(t) \sim \left(\frac{\sum_{\text{for all } i} A_i + B}{1 - e^{-\zeta(\mu - \varepsilon - \sum_{\text{for all } i} \lambda_i)}}, \zeta \right)\text{-EB where } \frac{1}{\zeta} = \sum_{\text{for all } i} \frac{1}{\alpha_i} + \frac{1}{\beta} . \tag{1}$$

Theorem2. Queueing delay process $D(t)$ of each EBB flow at the shared FIFO queue is upper bounded by an EB process as follows.

$$D(t) \sim \left(\frac{\sum_{\text{for all } i} A_i + B}{1 - e^{-\zeta(\mu - \varepsilon - \sum_{\text{for all } i} \lambda_i)}}, \zeta(\mu - \varepsilon) \right)\text{-EB} . \tag{2}$$

where ζ is the same as in Eq.(1).

The proofs of theorems are similar to those in [1] and are omitted for sake of limited space. Interested readers can refer [8] for details.

2.3 Performance Bounds under Prioritized Service Queueing Discipline

In this section, we consider the prioritized service queueing discipline. The traffic source i of class n is modeled by $(\lambda_{ni}, A_{ni}, \alpha_{ni})$ -EBB process. The channel is again modeled as $(\mu - \varepsilon, B, \beta)$ -EBF. Packets belong to each class is put into a separated queue. Under this service discipline, whenever the server is ready to provide service, it serves the backlogged queue with the highest priority. Each class receives service only when higher priority classes have no backlog in queues at all. By the duality between data traffic and error in channel in [1], the equivalent channel model seen by a non-highest priority class is a channel with higher error rate that combines the actual channel error process and the traffic processes of higher priority classes.

Theorem3. Under prioritized access queueing discipline, the stochastic bounds in Theorem1~Theorem2 still apply, with the modification in channel parameters depending on incoming flow’s priority class as follows.

(i).The queue size of class n is an EB process and its parameters are abbreviated as

$$Q_n(t) \sim \left(\frac{A'_n + B}{1 - e^{-\zeta(\mu - \varepsilon - \Lambda_n)}}, \zeta \right) . \tag{3}$$

where

$$\Lambda_n = \sum_{\substack{\text{for all flow } k \text{ with same} \\ \text{or higher priority than class } n}} \lambda_k, \quad A'_n = \sum_{\substack{\text{for all flow } k \text{ with same} \\ \text{or higher priority than class } n}} A_k, \quad \frac{1}{\zeta} = \sum_{\substack{\text{for all flow } k \text{ with same} \\ \text{or higher priority than class } n}} \frac{1}{\alpha_k} + \frac{1}{\beta}. \quad (4)$$

(ii). The queueing delay of each EBB flow of class n is an EB process a satisfying

$$D(t) \sim \left(\frac{A'_n + B}{1 - e^{-\zeta(\mu - \varepsilon - \Lambda_n)}} \right), \zeta(\mu - \varepsilon - \Lambda_n). \quad (5)$$

where Λ_n, A'_n, ζ are defined as above.

3 Measurement-Based Call Admission Control under Wireless Channel

In this section, we present an algorithm that makes call admission decision based on queue size statistics at the entrance node of wireless channel. The objective of our algorithm is maintaining statistical guarantee on queueing delay of incoming packets.

3.1 Call Admission Control for Single-Priority Class

To achieve the goal of maintaining statistical delay bound, an intuitive way is book-keeping statistics about queueing delay of each packet and using it for call admission decision. But doing this may introduce serious packet manipulation overhead. Thus, we propose a procedure to make call admission decisions based on queue size statistics, which can be gathered easier, while maintaining the QoS target specified in terms of delay.

First, we have a target QoS requirement specified in terms of delay (for example, more than 99% of packets encounter delay less than 50ms) and name it as *Target Point (TP)*. Once we specify the *TP* in delay domain, we can find the corresponding *TP* in queue size domain according to the Eq.(1) and Eq.(2). Namely, the decay-factor of queue size and decay-factor of delay are directly related by a proportional factor $\mu - \varepsilon$, which is the average service rate of the wireless channel.

We make call admission decision according to a simple guideline: Bookkeeping statistics about queue size and summarize it into a *System state Line (SL)*, as illustrated in Fig. 2. The leftmost point on *SL*, which is the probability of having non-zero queue size, is called *Starting Point (SP)*. As a new flow requests for service, we estimate the increment by which *SL* will shift upward according to the new flow's characteristics. If the shifted *SL* still remains below *TP*, we accept the new flow; otherwise, we reject it. A similar approach can be found in [4].

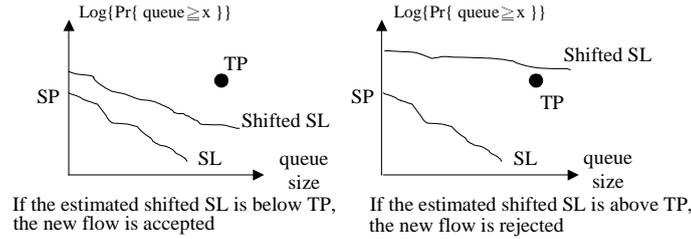


Fig. 2. Making call admission decision based on shifted SL and TP

When network is a time-varying system, statistics about earlier history provides less valuable information for making current decision. As a result, we have to make earlier data stand less weight in queue size statistics. A generally used technique in such situation is auto-regressive weighted average. Based on this method, we update the SL periodically.

The next step in making call admission decision is estimating SL shift. SL shift can be divided into two parts: The shift of SP and change of SL slope, corresponding to changes in pre-factor and decay-factor in Eq.(1), respectively. The change of slope is simpler to estimate, since decay-factor of queue size is only related to decay-factor of input flows and channel (See Eq.(1)). The decay-factors of existing flows and the channel are aggregated in decay-factor of current SL . We denote the estimated slope of current SL as m_{cnt} , the decay-factor of the new flow requesting service as α_{new} , and the estimated slope of shifted SL as m_{sfd} . Then we use the relation in Eq.(6) to estimate the decay-factor after the new flow joins the system, which is also the slope of the shifted SL .

$$1/m_{sfd} = 1/m_{cnt} + 1/\alpha_{new} . \tag{6}$$

The actual shift of SP is somewhat complicated to estimate. Thus, we use approximation technique to find upper bound of shift of SP . The SP , corresponding to $\text{Prob}\{\text{queue size} \geq 0\}$, also indicates the system utilization. If a new flow joins in, the system utilization increases by λ/μ , where λ is the average rate of new flow, μ is the ideal channel rate. Thus we can estimate the shift of SP by Eq.(7).

$$SP_{new} = SP_{old} + \lambda/\mu . \tag{7}$$

An important feature of our call admission control algorithm is that we adjust TP according to system utilization. Specifically, we choose a *Warning Level* of system utilization. If the estimated system utilization (SP_{new}) is under this level, we use original TP as call admission threshold. But if shifted SP is higher than this *Warning Level*, TP is moved down along the probability axis by multiplying a *Protection Factor*, which is less than 1. Different choices of *Warning Level* and *Protection Factor* can be made to achieve tradeoff between channel utilization and SLA violation probability.

3.2 Call Admission Algorithm for Multi-priority Classes

In order to provide better protection to delay sensitive traffic flows or to flows considered important by network operators, a widely used technique is creating multiple priority classes and assigning these flows to the high priority class.

In such an environment, our call admission algorithm needs to be modified to check whether we can accept a new flow under current network condition, without violating the SLA of *each* priority class.

From Theorem3, we see that the admission of a flow has no impact on performance of flows with higher priority, but influences the performance of flows with the same or lower priority. When admitting a flow of a specific class, we should check whether SLA for each of the lower priority classes can be sustained.

To enable such a check, we maintain separate *SL* curve for each priority class. The *SL* of highest priority class collects the statistic of highest priority queue size, which is the only visible queue to the highest priority flows. The *SL* of second highest priority class collects the statistic of the sum of the highest priority queue size and the second highest priority queue size, which is the equivalent queue size for second highest priority flows, and so on. In this case, the SLA and *TP* can be different in each priority class. When a new flow requests to join, we must check all the *SLs* it influences and make sure that after the estimated shift, each of these *SLs* remains below the corresponding *TP*. If any of these checks fail, we conclude that SLA guarantee for some priority class may fail with the admission of the new flow and we reject the new flow.

4 Simulation Results

In this section, we present some simulation results indicating that our call admission algorithm provides an effective mechanism leading to high system utilization while keeping SLA violation probability low. The *Warning Level* and *Protection Factor* introduced in section 3 make the call admission algorithm adjustable to match different requirements in various operation environments.

4.1 Simulation Environment

The simulation environment is illustrated in Fig. 3. During simulation, the users make random selections among 14 video clips stored in the VoD server. The video streams are modeled by EBB processes, and its parameters are known. The video traces are packet patterns of movie previews encoded in Real Media format. We assume the queueing delay at the shared FIFO queue is the major part of end-to-end delay and neglect other factors, i.e. Congestion occurs only at the wireless channel. A rejected request is assumed to leave the system without changing the future request arrival pattern.

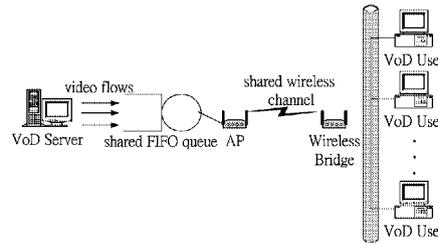


Fig. 3. The simulation environment

4.2 Adjustable Policy of the Call Admission Algorithm

In all following simulation cases, the target QoS guarantee in SLA for video flows is stated as “Less than 1% of packets encounter queueing delay more than 50ms.” The parameters used in simulation cases are summarized in Table 1(a). We test the call admission algorithm under the on-off channel model with alternating, exponentially distributed channel-on and channel-off periods with mean 0.99 and 0.01 second, respectively. We adapt this model to characterize the sporadic service outage in wireless channels. Simulation results are listed in Table 1(b).

Table 1. Simulation parameters and results using Real Media format video traces

(a) Simulation Parameters			(b) Results of on-off channel		
Simulation case	Warning Level	Protection Factor	Simulation case	SLA violation prob.	Average channel utilization
<i>Loose</i>	1.0	N/A	<i>Loose</i>	99.09%	0.875
<i>Step-1</i>	0.7	0.01	<i>Step-1</i>	29.12%	0.850
<i>Step-2</i>	0.6	0.01	<i>Step-2</i>	5.07%	0.803
<i>Step-3</i>	0.5	0.01	<i>Step-3</i>	2.29%	0.759

From the results above, we see a tradeoff between channel utilization and SLA violation probability for accepted flows. The SLA violation probability can be effectively reduced at the cost of lower channel utilization. The most desirable policy depends on the operator’s considerations and is different case by case. However, if we do not adjust *TP* according to utilization (the *Loose* case in simulation), the resulting SLA violation probability is unacceptable. This shows the need of using *Protection Factor*.

4.3 Mixing Video Traffic with Data Traffic

Next, we validate the applicability of our call admission algorithm for multiple priority classes. When delay sensitive video traffic flows and TCP flows are multiplexed in a single queue, the bursty nature of TCP flows causes performance degradation of

video traffic flows, as shown in Table 2(a). A solution for this problem is assigning delay insensitive TCP flows as low priority and video traffic flows as high priority.

Table 2. Simulation results with LAN trace injected

(a) Single priority class				(b) 2 priority classes			
Simulation case	SLA violation prob.	Average channel utilization	Mean delay of LAN trace packets	Simulation case	SLA violation prob.	Average channel utilization	Mean delay of LAN trace packets
<i>Loose</i>	99.29%	0.892	465.3ms	<i>Loose</i>	27.96%	0.862	980.8ms
<i>Step-1</i>	42.78%	0.866	302.2ms	<i>Step-1</i>	2.67%	0.836	336.8ms
<i>Step-2</i>	22.79%	0.824	110.6ms	<i>Step-2</i>	2.16%	0.800	406.2ms
<i>Step-3</i>	7.17%	0.767	22.9ms	<i>Step-3</i>	0%	0.744	75.1ms

In the simulation case, we choose one of the LAN traces available at [7] as representative of data traffic flows from the Internet. Target QoS guarantee for video flows are the same as that in section 4.2. No call admission control is used for the data traffic and no QoS guarantee is provided to it. It is a “background traffic” that always exists during the simulation. We use the on-off channel model and simulation results are summarized in Table 2(b).

We see that the SLA violation probability of video flows is not adversely influenced by data traffic if prioritized access queueing discipline is used, but the mean delay of LAN trace packets is obviously larger. The average channel utilizations in these simulation cases are similar in comparison to those in Table 1. If the TCP flows can tolerate higher queueing delay, setting them as low priority can be a viable solution toward providing QoS for delay sensitive video flows in a general-purpose network environment.

5 Conclusions

In this paper, we first derive stochastic performance bounds for key performance metrics under FCFS and prioritized access queueing disciplines. Then, based on the bound equations, we propose a call admission algorithm, which performs on-line measurement of current network condition. With the call admission algorithm, the network operator can provide statistical SLA guarantee to accepted users. The call admission algorithm can be modified for multi-priority queueing discipline, in which important or delay sensitive flows are better protected by assigning them as high priority flows.

Simulation studies show that there is a tradeoff between system utilization and SLA violation probability. The parameters of our call admission algorithm can be adjusted to match different operator requirements. If video traffic and data traffic are multiplexed in single FIFO queue, some additional mechanism, such as prioritized access of channel at the data link or MAC layer, is required if one wants to provide statistical SLA guarantee to video flows in a general-purpose network.

We conclude that the stochastic bound approach for QoS control is suitable to be used for loss tolerant multi-media traffic or other Internet applications in the wireless access networks.

References

1. K. Lee, "Performance Bounds in Communication Networks with Variable-rate Links," *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, pp. 126 – 136, 1995.
2. R. Cruz, "A Calculus for Network Delay, part I: Network Elements in Isolation," *IEEE Trans. Information Theory*, vol. 37, no. 1, pp.114-131, Jan. 1991.
3. W. Fischer and K. Meier-Hellstern, "The Markov-Modulated Poisson Process (MMPP) Cookbook," *Performance Evaluation*, vol. 18, pp.149-171, 1992.
4. M. Venkatraman, N. Nasrabadi, "An Admission Control Framework to Support Media-Streaming over Packet-Switched Networks," *ICC 1999*, vol. 2, pp. 1357-1361, 1999.
5. T. Lee, M. Zukerman and R. Addie, "Admission Control Schemes for Bursty Multimedia Traffic," *INFOCOM 2001*, vol. 1, pp.478-487, 2001.
6. Y. Bao and A. Sethi, "Performance-driven Adaptive Admission Control for Multimedia Applications," *ICC 1999*, vol. 1, pp. 199-203, 1999.
7. <http://www.acm.org/sigs/sigcomm/ITA/>, The Internet Traffic Archive.
8. Wei-jen Hsu, *Performance Bounds and Call Admission Control Algorithm in Wireless Access Networks*, Master Thesis, National Taiwan University, 2001.