Differentiated Statistical QoS Guarantees for Real-Time CBR Services in Broadband Wireless Access Networks

Hong Zhou
University of Southern Queensland
Toowoomba, QLD 4350
Email: hzhou@usq.edu.au

Zhongwei Zhang
University of Southern Queensland
Toowoomba, QLD 4350
Email: zhongwei@usq.edu.au

***Abstract*—A wide range of emerging real-time services require different levels of Quality of Services (QoS) guarantees over wireless networks. Scheduling algorithms play a key role in meeting these QoS requirements. Most of research in this area have been focused on deterministic delay bounds and the statistical bounds of differentiated real-time services are not well known. This paper provides the mathematical analysis of the statistical delay bounds of different levels of Constant Bit Rate (CBR) traffic under First Come First Served with static priority (P-FCFS) scheduling. The mathematical results are supported by the simulation studies. The statistical delay bounds are also compared with the deterministic delay bounds of several popular rate-based scheduling algorithms. It is observed that the deterministic bounds of the scheduling algorithms are much larger than the statistical bounds and are overly conservative in the design and analysis of efficient QoS support in wireless access systems.**

***Index Terms*—statistical delay bound, deterministic delay bound, QoS, broadband wireless access**

I. Introduction

In recent years, there have been increasing demands for delivering a wide range of real-time multimedia applications in wireless networks. IEEE 802.16 broadband wireless access systems provide fixed-wireless access for individual homes and business offices through the base station instead of cable and DSL in wired networks. This creates great flexibility and convenience as well as challenges for the design and analysis of such networks. Multimedia communications require certain level of Quality of Services (QoS) guarantees and individual applications also have very diverse QoS requirements. The wireless networks are required to support real-time multimedia applications with different QoS guarantees.

QoS performance is characterized by a set of parameters in any packet-switched network, namely end-to-end delay, delay variation (i.e. jitter) and packet loss rate. Unlike non-real-time services, quality of real-time services is mainly reflected by their delay behaviors, namely, delay and delay variation.

Scheduling algorithms play a key role in satisfying these QoS requirements. In the past twenty years, a significant volume of research has been published in literature on scheduling algorithms such as Packet-by-packet Generalized Processor Sharing (PGPS) [1], Self-Clocked Fair Queueing (SCFQ) [2], Latency-Rate (LR) Server [3], Start-time Fair Queueing (SFQ) [4], Wireless Packet Scheduling (WPS) [5] and Energy Efficient Weighted Fair Queueing (E^2 WFQ) [6]. However, these research were basically focused on the deterministic delay bounds. The statistical delay bounds of scheduling algorithms meeting different QoS requirements have not been adequately studied.

IEEE 802.16 Broadband wireless access systems are designed to support a wide range of applications (data, video and audio) with different QoS requirements. IEEE 802.16 defines four types of service flows, namely Unsolicited Grant Service (UGS), real-time Polling Service (rtPS), non-real-time Polling Service (nrtPS) and Best effort service (BE). There are two types of service flows for real-time services, i.e. USG supports CBR traffic including VoIP streams while rtPS supports real-time VBR flows such as MPEG video. IEEE 802.16 standard left the scheduling algorithm for the uplink and downlink scheduling algorithm undefined. Wongthavarawat and Ganz in [7] proposed a combination of strict priority scheduling, Earliest Deadline First [8] and WFQ [1]. The CBR real-time traffic, (UGS) has preemptive priority over other type of flows. In this paper, we are concerned with real-time CBR traffic and we assume there are different levels of QoS requirements within CBR

traffic. For example, the emergence and remote medical CBR services should have higher QoS requirements than normal VoIP chats. We analyse the delay for different service levels of CBR traffic by sloving class-based nD/D/1 queue.

The rest of this paper is organised as follows. In Section II, a discrete-time P-FCFS queueing system model with Constant Bit Rate (CBR) inputs is defined and illustrated. In Section III, we analyse the model in general cases that there are arbitrary number of priority levels and there are arbitrary number of traffic sources at individual levels. The queueing delay distribution for each service level is derived. In Section IV, we provide the delay distributions of different priority classes obtained by both mathematical and simulation analysis. Section V concludes the paper.

II. Discrete-Time Priority Queueing Model

The nD/D/1 model with several priority levels analyzed here has the following characteristics: (a) independent periodic sources with same period; (b) deterministic service/transmission time; (c) with priority levels; (d) discrete-time queueing system, or say slotted server.

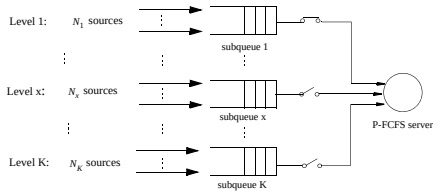

Fig. 1. Priority queueing model

As illustrated in Figure 1, we assume that there are totally N active real-time sources which are classified into K priority levels. For each priority level x ($1 \leq x \leq K$), the number of sources is N_y. Each source generates fix length cells periodically with same period T. To keep the system stable, period T has to be greater than the total number of sources N.

The discrete-time model assumes slotted transmission on the path. The time axis is divided into fixed length slots and the transmission is restricted to start at a slot boundary. As a result of this restriction, each packet has to wait at least until the start of the next time slot.

Time slot and cell length are assumed to be unity. Then the service time of each packet is also equal to unit time. In a P-FCFS queue, packets with higher priorities will be served first and those with lower priority will wait until all the higher priority packets have been served. The packets with the same priority will be served in FCFS principles. Thus, the delay experienced by a packet is equal to the number of packets found in the same and higher priority queues at the arrival time and packets with higher priority that have arrived between the arrival time and transmission time plus the remaining transmission time of the packet in service. Note that the packets from the lower priorities do not affect the delay experienced by a packet from higher priorities.

III. Mathematical Analysis

Let the source tested, say i, be the tagged source and all other sources be background sources. Suppose that the priority of the tagged source is x ($x = 1, 2, \cdots, K$). Because the sources with lower priorities than the tagged source do not affect the delay of the tagged source, we only consider the sources with the same or higher priorities than the tagged one. Let N_H be the total number of sources with higher priorities than the tagged one. Thus, $N_H = N_1 + \cdots + N_{x-1}$. Let the waiting time/queueing delay experienced by the packet from the tagged source q_i be the interval from the beginning of the first slot since the packet arrives to that of the slot at which it starts to be served. Note that the residual slot period until the start of the next time slot is omitted and the delay is always an integer. In general this simplification does not effect our results. In what follows, we calculate the probability when the queueing delay q_i is equal to d, namely $Pr\{q_i = d\}(d \geq 0)$.

Consider a period of time T from $t+d-T$ to $t+d$ and separate this interval into three sub-intervals. Suppose the arrival time of the tagged source i is uniformly distributed within the tth time slot ($[t-1, t]$). The arrivals of background sources are independent and uniformly distributed in the interval $[t+d-T, t+d]$. The numbers of sources arriving on the sub-intervals are defined as follows (See Figure 2).

- n_H is the number of sources with higher priorities than arriving during $(t, t+d]$;
- n_x is the number of sources with priority x arriving during $(t, t+d]$;
- $n_H^{'}$ is the number of sources with higher priorities arriving during $(t-1, t]$;
- $n_x^{'}$ is is the number of sources with the same priority arriving during $(t-1, t]$;
- A_τ is number of background sources with higher priorities arriving during the $t+\tau$th time slot;
- $n_H^{"}$ is the number of sources with higher priorities arriving during $(t+d-T, t-1]$, $n_H^{"} = N_H - n_H - n_H^{'}$

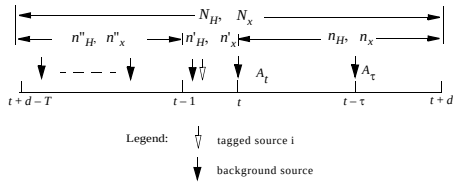

Fig. 2. The numbers of sources arriving in a period of time

Q_t is defined as the total length of packets waiting in higher priority sub-queues and packets in sub-queue x ahead of tagged packet at the end of tth time slot. When the queueing delay q_i is equal to d, the server will keep busy till $t+d$. The probability distribution of queueing delay for tagged source i can be obtained as shown in [2].

$$Pr\{q_i = d\} = T^{-(N_H+N_x)+1}[U(d, N_H, N_x) - V(d, N_H, N_x)]$$

where $U(d, N_H, N_x)$ and $V(d, N_H, N_x)$ represent question (3) and (4).

and Ψ, $\#\Omega$ are defined as equation (5) and (6).

Further, the tail distribution is

$$Pr\{q_i > d\} = 1 - \sum_{j=0}^{d} Pr\{q_j = n_x\}$$

IV. Numerical Results

A. Simulation Results

Besides mathematical analysis, the queueing system discussed in the previous sections was implemented and analyzed in simulations using OPNET. The parameters in the simulation were the same with those used in the mathematical calculations. The results shows that the delay distributions from mathematical analysis and simulation are identical. The tail distributions with different

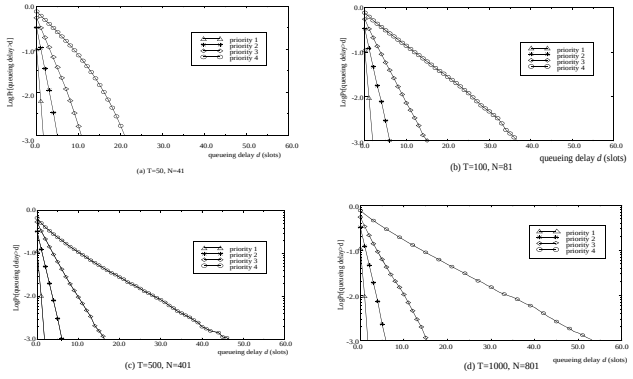

Fig. 3. Delay distributions of nD/D/1 queue with four service classes

numbers of sources are shown in Figure 3. Comparisons of Figure 3 (a), (b), (c) and (d) show that as the number of sources increases, the delay for each class increases. However, the increases in delay become slower as the number of sources increases, especially for higher priorities. Thus, the delay is statistically bounded even in the core network with a large number of competing sources.

For comparison, the delay distributions of a queueing system without any priority is given in Figure 4. Comparing the curves in Figure 3 with the curve of

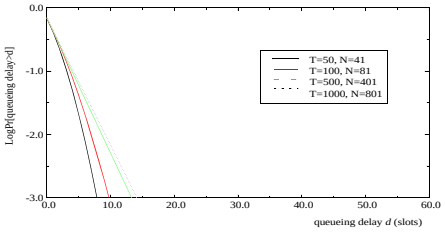

Fig. 4. Delay distributions of nD/D/1 queue w/o diff. services

$T = 50$ and case $N = 40$ in Figure 4, the first two higher priorities have better performance than the one without priority. The third priority has a slightly worse performance than the one without priority. The fourth priority has obvious worse performance than the one without priority. Thus, priority queueing can provide individual sources with a desirable QoS according to different requirements. The network resources can be efficiently utilized.

B. Comparison with Deterministic Delay Bounds

This section compares the statistical delay bounds with the deterministic delay bounds (i.e. queueing latencies) of several rate-based scheduling algorithms. In this example, we use the same scenario as defined in the first example of Section IV-A which results are shown in Figure 3. In Figure 3, the statistical delay bound of a session with priority level 4 (the lowest priority) when $LogPr[q_i > d] = -3$ is equal to 22 time slots.

The deterministic delay bounds of several scheduling algorithms are given in Table I. In Table I, N denotes the number of sessions sharing the outgoing link L_i, and ρ_i denote the maximum packet length and allocated rate for session t, L_{max} denotes the maximum packet length for all sessions except session i, r denotes the outgoing link capacity. For comparison, we assume time slot = 1 second. We also assume that the bandwidth of the outgoing link $r = 1kb/s$, the packet lengths $L_i = L_{max} = 1kb$, the number of sessions $N = 41$, and the allocated rate $\rho_i = r/N = 0.024kb/s$. As shown in Table I, the deterministic queueing delay is much

$$U(d,N_H,N_x)=\sum_{n_H}\sum_{n_x}\sum_{n_H^{'}}\Psi\sum_{n_x^{'}=1}^{N_x-n_x}\left(\begin{array}{c}N_x-n_x\\ n_x^{'}\end{array}\right)\#\Omega_{\leq m}[T-d-1,n_{x+H}^{''}] \tag{3}$$

$$V(d,N_H,N_x)=\sum_{n_H}\sum_{n_x}\sum_{n_H}\Psi\sum_{n_x^{'}=1}^{m}\left(\begin{array}{c}N_x-n_x\\ n_x^{'}\end{array}\right)\#\Omega_{\leq m-n_{n_x}^{'}}[T-d-1,n_{x+H}^{''}] \tag{4}$$

$$\Psi=\frac{d^{n_H+n_x-1}}{N_x-n_x}(d-n_H)\left(\begin{array}{c}N_H\\ n_H+n_H^{'}\end{array}\right)\left(\begin{array}{c}N_x-1\\ n_x\end{array}\right)\left(\begin{array}{c}n_H+n_{H'}\\ n_H\end{array}\right) \tag{5}$$

$$\#\Omega_{\leq m-n_{n_x}^{'}}[T-d-1,n_{x+H}^{''}]=(T-d-n_{x+H}^{''})\sum_{j=0}^{m}\left(\begin{array}{c}n_{x+H}^{''}\\ j\end{array}\right)(-1-m+j)^j(T-d+m-j)^{n_{x+H}^{''}-j-1} \tag{6}$$

TABLE I
THE DELAY BOUNDS OF SCHEDULING ALGORITHMS

Scheduling Algorithm	Queueing Latency	Deterministic Delay Bound
WFQ [1]	$\frac{L_i}{\rho_i}-\frac{L_i}{r}+\frac{L_{max}}{r}$	41s
$SCFQ$ [2]	$\frac{L_i}{\rho_i}-\frac{L_i}{r}+(N-1)\frac{L_{max}}{r}$	80s
SFQ [4]	$(N-1)\frac{L_{max}}{r}$	40s

larger than the statistical delay bounds. The statistical delay guarantees do not care about a small fraction of packets (e.g. one in a million packets) which experience the delay exceed the bounds. The real-time services generally can tolerate a small number of packet losses, therefore statistical delay guarantees are sufficient and suitable for these applications.

V. CONCLUSION

In this paper, we analyse the statistical access delay of different classes of real-time CBR services in wireless networks. Numerical results from both mathematical and simulation studies are provided. Identical results are shown from the two different methods. The results also show that the performance can be effectively differentiated by the simple priority scheduling algorithm. The deterministic delay bounds/latencies are generally much larger than the statistical delay bounds. As real-time services generally tolerate a small number of packet losses, the statistical delay guarantees are sufficient and thus more important for real-time services. The analysis not only can provide accurate QoS performance for multiple-class real-time services but also can be used to design efficient admission control and upstream scheduling mechanisms in wireless access systems.

REFERENCES

[1] A. K. Parekh and R. G. Gallerger, "A generalized processor sharing approach to flow control in integrated services networks: The single node case," *IEEE/ACM Trans. Networking*, vol. 1, 1993.

[2] S. J. Golestani, "A self-clocked fair queueing scheme for broadband applications," in *Proceedings of IEEE INFOCOM*, 1994.

[3] D. Stiliadis and A. Varma, "Latency-rate servers: A general model for analysis of traffic scheduling algorithms," *IEEE/ACM Trans. Networking*, vol. 6, no. 5, 1998.

[4] P. Goyal, H. M. Vin, and H. Cheng, "Start-time fair queueing: A scheduling algorithm for integrated services packet switching networks," Department of Computer Sciences, The University of Texas at Austin, Tech. Rep. TR-96-02, Jan. 1996.

[5] S. Lu, V. Raghunathan, and R. Srikant, "Fair scheduling in wireless packet networks," *IEEE/ACM Trans. Networking*, vol. 7, no. 4, 1999.

[6] V. Raghunathan, S. Ganeriwal, M. Srivastava, and C. Schurgers, "Energy efficient wireless packet scheduling and fair queuing," *ACM Trans. Embedded Comput. Sys.*, vol. 3, no. 1, 2004.

[7] K. Wongthavarawat and A. Ganz, "Packet scheduling for qos support in ieee 802.16 broadband wireless access systems," *International Journal of Communication Systems*, 2003.

[8] L. Georgiadis, R. Guerin, and A. Parekh, "Optimal multiplexing on a single link: Delay and buffer requirements," in *Proceedings of IEEE INFOCOM94*, vol. 2, 1994.