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

Hong Zhou
University of Southern Queensland
Toowoomba, Austraila
Email: hong.zhou@usq.edu.au

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

***Abstract*—A wide range of emerging real-time services (e.g. VoIP, video conferencing, video-on-demand) require different levels of Quality of Services (QoS) guarantees over wireless networks. Scheduling algorithms play a key role in meeting these QoS requirements. A distinction of QoS guarantees is made between deterministic and statistical guarantees. 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.**

I. Introduction

In recent years, there have been increasing demands for delivering a wide range of real-time multimedia applications (e.g. VoIP, video conferencing, video-on-demand) in broadband wireless access networks. IEEE 802.16 standard for broadband wireless access systems [18] 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. It requires the wireless access networks 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 [1], [2], [3]. Unlike non-real-time services, quality of real-time services is mainly reflected by their delay behaviors, namely, delay and delay variation. A distinction in QoS performance guarantees is made between *deterministic guarantees* and *statistical guarantees* [15]. In the deterministic case, guarantees provide a bound on the performance of all packets from a session. In other words, deterministic delay guarantees promise that no packet would be delayed more than D time units on an end-to-end basis. The value D is defined as the *Deterministic Delay Bound* (DDB). On the other hand, statistical guarantees promise that no more than a specified fraction, α, of packets would experience delay more than $D(\alpha)$ time units. $D(\alpha)$ is defined as the *Statistical Delay Bound* (SDB). As the fraction α becomes smaller, the statistical delay bound increases. In the special case of $\alpha = 0$, the statistical delay bound reaches the maximum value and is equal to the deterministic delay bound. That is, $D(\alpha) = D$.

Similarly, delay variation is defined as the difference between the best and worst case expectation of variable delays (i.e. mainly queueing delays). In statistical case, the best case is equal to zero and the worst case is a value likely to be exceeded with a probability less than α(for example 10^{-9}). It should be noted that, when the occasional exceptions are rare enough (e.g. $\alpha = 10^{-9}$), though the SDB may still be much smaller than DDB, the distinction between statistical guarantees and deterministic guarantees is negligible for most real-time services. Consequently, the QoS offered by statistical guarantees will be as good as those offered by deterministic guarantees for these real-time services. In general, the deterministic delay bound is much larger than the statistical delay bound and thus it is overly conservative. A statistical delay bound is sufficient for almost all real-time services.

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) [4], Self-Clocked Fair Queueing (SCFQ) [5], Latency-Rate (LR) Server [6], Start-time Fair Queueing (SFQ) [7], Wireless Packet Scheduling (WPS) [8] and Energy Efficient Weighted Fair Queueing (E^2 WFQ) [9]. 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.

In this paper, we examine the access delay of CBR real-time traffic in wireless access systems (e.g. IEEE 802.16). Future backbone networks have very high bandwidth and the delay experienced is very low. On the other hand, the access networks have relatively limited speed and the delay experienced by CBR traffic is very large. Therefore the analysis and design of the wireless access systems to support the QoS of real-time services is very important.

IEEE 802.16 standard for 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. UGS supports CBR traffic including VoIP streams while rtPS supports real-time VBR flows such as MPEG video [13], [14]. IEEE 802.16 standard left the scheduling algorithm for the uplink and downlink scheduling algorithm undefined. Wongthavarawat and Ganz in [17] proposed a combination of strict priority scheduling, Earliest Deadline First [19] and WFQ [4]. 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 solving class-based nD/D/1 queue.

The superposition of independent streams with periodic arrival patterns has been modelled by nD/D/1 queue in several past works [10], [11], [12], [16]. The problem of traffic delay analysis can be solved by finding the waiting time distribution of nD/D/1 Queue. Our study is different from the cited research as we differentiate CBR streams by priorities and analyse the delay for nD/D/1 queue with arbitrary number of service priorities in IEEE 802.16 broadband access networks.

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 mathematical 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.

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_x. 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.

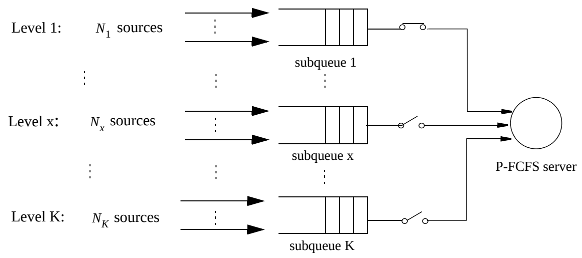


Figure 1. Priority queueing model

Without loss of generality, 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,

$$N_H = N_1 + \cdots + N_{x-1} \tag{1}$$

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 the 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 $(t+d-T, t-1]$, $n^{"}_H = N_H - n_H - n^{'}_H$
- $n^{"}_x$ is the number of sources with the same priority arriving during $(t+d-T, t-1]$, $n^{"}_x = N_x - n_x - n^{'}_x$.

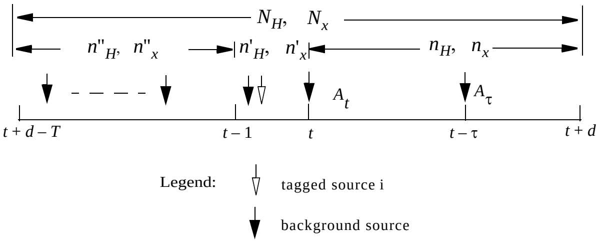


Figure 2. The numbers of sources arriving in a period of time

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 $t+d$. Thus, the following two necessary and sufficient conditions must be satisfied.

$$Q_{t+d} = \max(Q_{t+d-1} - 1, 0) + A_d = 0 \qquad (2)$$

$$\text{and} \qquad Q_{t+\tau} > 0 \text{ for all } \tau = 0, 1, \cdots, d-1 \qquad (3)$$

From [3],

$$\begin{aligned} Q_{t+\tau} &= max(Q_{t+\tau-1} - 1, 0) + A_\tau \\ &= Q_{t+\tau-1} - 1 + A_\tau \qquad (4) \end{aligned}$$

By iteration on τ, $Q_{t+\tau}$ is equal to

$$\begin{aligned} &Q_{t+\tau} \\ &= Q_t + A_1 + A_2 + \cdots + A_{\tau-1} + A_\tau - \tau \qquad (5) \\ &= Q_t + A(t, t+\tau) - \tau \qquad (6) \end{aligned}$$

where $A(t, t+\tau)$ represents the total number of arrivals with higher priorities in the interval $[t, t+\tau]$. Then (2) and (3) are equivalent to

$$Q_t = d - n_H \qquad (7)$$

$$\text{and} \qquad A(t, t+\tau) > \tau - Q_t \text{ for all } \tau = 0, 1, \cdots, d-1 \, (8)$$

respectively. Thus, the probability of queueing delay is

$$\begin{aligned} &Pr\{q_i = d\} \\ =&Pr\{Q_{t+\tau} > 0, \tau = 0, 1, \cdots, d-1, Q_{t+d} = 0\} \\ =&Pr\{A(t, t+\tau) > \tau - Q_t, Q_t = d - n_H\} \\ &\sum_{n_H=0}^{d-1} \sum_{n_x=0}^{N_x-1} \sum_{n^{'}_H=0}^{N_H-n_H} Pr\{A(t, t+\tau) > \tau - Q_t, Q_t = d - n_H | \\ &A_H(t, t+d) = n_H, A_x(t, t+d) = n_x, A_H(t-1, t) = n^{'}_H\} \\ &\cdot Pr\{A_H(t, t+d) = n_H, A_x(t, t+d) = n_x, A_H(t, t+d) = n^{'}_H\}. \end{aligned}$$

For the convenience of expression, in the following discussion, the term $Pr\{A(t, t+\tau) > \tau - Q_t, Q_t = d - n_H\}$ is used to represent $Pr\{A(t, t+\tau) > \tau - Q_t, Q_t = d - n_H | A_H(t, t+d) = n_H, A_x(t, t+d) = n_x, A_H(t-1, t) = n^{'}_H\}$ Moreover, $A_H(t, t+d)$, $A_x(t, t+d)$, and $A_x(t-1, t)$ are independent random variables and (8) can be written

$$\begin{aligned} &Pr\{q_i = d\} \\ =&\sum_{n_H=0}^{d-1} \sum_{n_x=0}^{N_x-1} \sum_{n^{'}_H=0}^{N_H-n_H} Pr\{A(t, t+\tau) > \tau - Q_t, Q_t = d - n_H\} \\ &\cdot Pr\{A_H(t, t+d) = n_H\} Pr\{A_x(t, t+d) = n_x\} Pr\{A_H(t-1, t) = n^{'}_H\}. \\ =&\sum_{n_H=0}^{d-1} \sum_{n_x=0}^{N_x-1} \sum_{n^{'}_H=0}^{N_H-n_H} Pr\{A(t, t+\tau) > \tau - Q_t | Q_t = d - n_H\} \cdot Pr\{Q_t = d - \\ &\cdot Pr\{A_H(t, t+d) = n_H\} \cdot Pr\{A_x(t, t+d) = n_x\} \cdot Pr\{A_H(t-1, t) = n^{'}_H \end{aligned}$$

Let $\tau = d - \tau^{'}$. Then,

$$A_H(t, t+d) = n_H - \sum_{j=1}^{\tau^{'}} A_{d-j} > d - \tau^{'} - d + n_H$$

which leads to

$$\sum_{j=1}^{\tau^{'}} A_{d-j} < \tau^{'} \;\; \tau^{'} = 1, 2, \cdots, d. \qquad (9)$$

By the Ballot Theorem 4 in [12],

$$Pr\{\sum_{j=1}^{\tau^{'}} A_{d-j} < \tau^{'}, \tau^{'} = 1, 2, \cdots, d\} = 1 - \frac{n_H}{d}. \qquad (10)$$

Substituting (11) int (9),

$$Pr\{q_i = d\}$$

$$= \sum_{n_H=0}^{d-1} \sum_{n_x=0}^{N_x-1} \sum_{n_H'=0}^{N_H-n_H} Pr\{Q_t = d - n_H\} Pr\{A_H(t, t+d) = n_H\}$$

$$\cdot Pr\{A_x(t, t+d) = n_x\} Pr\{A_H(t-1, t) = n_H'\}.$$

The next step is the evaluation of each term in (12) separately. The last three terms in (12) can be easily obtained from the *Probability Mass Function*(pmf) of Binomial random variables [17]. That is,

$$Pr\{A_H(t, t+d) = n_H\} = \left(\begin{array}{c} N_H \\ n_H \end{array} \right) \left(\frac{d}{T}\right)^{n_H} \left(1 - \frac{d}{T}\right)^{N_H - n_H}$$

$$Pr\{A_x(t, t+d) = n_x\} = \left(\begin{array}{c} N_x - 1 \\ n_x \end{array} \right) \left(\frac{d}{T}\right)^{n_x} \left(1 - \frac{d}{T}\right)^{N_x - n_x - 1}$$

$$Pr\{A_H(t-1, t) = n_H'\} = \left(\begin{array}{c} N_H - n_H \\ n_H' \end{array} \right) \left(\frac{1}{T-d}\right)^{n_H'} \left(1 - \frac{1}{T-d}\right)^{N_H - n_H' - n_H}$$

In the following, the second term in (12) is calculated. Firstly, define Y as the rank of the tagged source i within the n_x' sources. Then

$$Q_t = \max(Q_{t-1} - 1, 0) + Y - 1 + n_H'. \tag{11}$$

For convenience of expression, let $L = max(Q_{t-1} - 1, 0)$. Thus,

$$Pr\{Q_t = d - n_H\} = Pr\{L + Y + n_H' - 1 = d - n_H\}$$

$$= \sum_{n_x'=1}^{N_x - n_x} Pr\{Y = d - n_H - n_H' - L + 1 | A_x(t-1, t) = n_x'\} Pr\{A_x(t-1, t) = n_x'\}$$

$$= \sum_{n_x'} \sum_{l} Pr\{Y = d - n_H - n_H' - L + 1 | L = l, A_x(t-1, t) = n_x'\} Pr\{L = l\}$$

$$\cdot Pr\{A_x(t-1, t) = n_x'\}. \tag{12}$$

Making use of the result shown in Appendix E of [10], the probability of the rank $Y = y$ among n_x' sources is simply the reciprocal of n_x'. That is

$$Pr\{Y = d - n_H - n_H' - L + 1 | L = l, A_x(t-1, t) = n_x'\} = \frac{1}{n_x'} \tag{13}$$

Thus

$$Pr\{Q_t = d - n_H\}$$

$$= \sum_{n_x'} \sum_{l} \frac{1}{n_x'} Pr\{L = l\} Pr\{A_x(t-1, t) = n_x'\}.$$

Note that $L = d - n_H - n_H' - Y + 1$ and $1 \le Y \le n_x'$. If set $m = d - n_H - n_H'$, then,

$$max(m - n_x' + 1, 0) \le l \le m.$$

Thus,

$$\begin{array}{rl} m - n_x' + 1 \le l \le m & \text{when } 1 \le n_x' < m + 1 \\ 0 \le l \le m & \text{when } m + 1 \le n_x' \le N_x - n_x. \end{array} \tag{14}$$

Equation (19) becomes

$$Pr\{Q_t = d - n_H\} =$$

$$\sum_{n_x'=1}^{m} Pr\{A_x(t-1, t) = n_x'\} \sum_{l=m-n_x'+1}^{m} Pr\{L = l\}$$

$$+ \sum_{j_x=m+1}^{N_x - n_x} \frac{1}{n_x'} Pr\{L = l\} Pr\{A_x(t-1, t) = n_x'\}.$$

Noting that

$$\sum_{l=a}^{b} Pr\{L = l\} = Pr\{L > a - 1\} - Pr\{L > b\} \tag{15}$$

Applying (22) in (21) and combining the same terms, (21) becomes

$$Pr\{Q_t = d - n_H\} =$$

$$\sum_{n_x'=1}^{m} \frac{1}{n_x'} Pr\{A_x(t-1, t) = n_x'\} Pr\{L > m - n_x'\}$$

$$+ \sum_{j_x=m+1}^{N_x - n_x} \frac{1}{n_x'} Pr\{A_x(t-1, t) = n_x'\}$$

$$- \sum_{n_x=1}^{N_x - n_x} \frac{1}{n_x'} Pr\{A_x(t-1, t) = n_x'\} Pr\{L > m\}.$$

As shown in Appendix, the following can be obtained.

$$Pr\{Q_t = d - n_H\} = \frac{T - d}{N_x - n_x} (T - d)^{-(N_x - n_x)} (T - d - 1)^{-n''_H}$$

$$\left\{ \sum_{n_x'=1}^{N_x - n_x} \left(\begin{array}{c} N_x - n_x \\ n_x' \end{array} \right) \#\Omega_{\le m} \left[T - d - 1, n''_{x+H}\right] - \sum_{n_x'=1}^{m} \left(\begin{array}{c} N_x - \\ n_x' \end{array} \right. \right.$$

where $m = d - n_H - n_H'$, $n''_{x+H} = n''_x + n''_H = N_H + N_x - n_H - n_H' - n_x - n_x'$, and

$$\#\Omega_{\le c}[a, b] = (a - b + c + 1) \sum_{j=0}^{c} \left(\begin{array}{c} b \\ j \end{array} \right) (-1 - c + j)^j (a + c - j + 1)^{b - j - 1}. \tag{17}$$

Substituting (13), (14), (15) and (24) into (12), and canceling the like terms gives

$$Pr\{q_i = d\} = T^{-(N_H + N_x) + 1} \left[U(d, N_H, N_x) - V(d, N_H, N_x)\right] \tag{18}$$

where $U(d, N_H, N_x)$ and $V(d, N_H, N_x)$ represent

$$U(d, N_H, N_x) = \sum_{n_H} \sum_{n_x} \sum_{n_H'} \Psi \sum_{n_x'=1}^{N_x - n_x} \left(\begin{array}{c} N_x - n_x \\ n_x' \end{array} \right) \#\Omega_{\le m}[T - d -$$

$$V(d, N_H, N_x) = \sum_{n_H} \sum_{n_x} \sum_{n_H} \Psi \sum_{n_x'=1}^{m} \left(\begin{array}{c} N_x - n_x \\ n_x' \end{array} \right) \#\Omega_{\le m - n'_{n_x}}[T -$$

and

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

In addition, it is known that

$$Pr\{q_i > d\} = 1 - \sum_{j=0}^{d} Pr\{q_j = n_x\} \qquad (19)$$

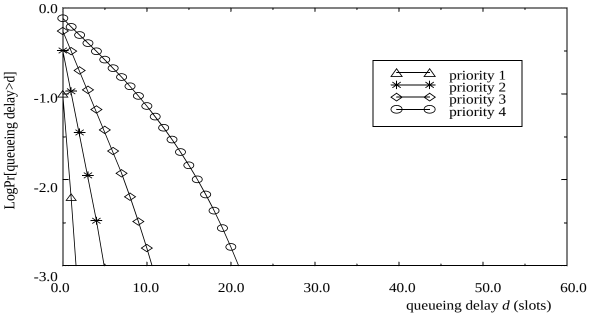
Substituting (26) into (27), the tail distribution of queueing delay will be obtained.

IV. Numerical Results

A. Mathematical Results

Based on the analysis in Section III, logarithmic functions of the tail distributions of the queueing system at a load of 0.8, $LogPr\{q_i > d\}$, are calculated and plotted in Figure 3 using Mathematica. In Figure 3, the period is equal to $T = 50$ and there are 41 background sources (i.e. $N = 41$). The sources are classified into four priorities and each priority has an equal number of sources (excluding the tagged source) in the queueing model, namely, $N_x^{'} = (N-1)/4, x \in [1,4]$. From Figure 3, it can be seen that delays are differentiated by their priorities and when the priority becomes lower, the delays become significantly larger.

For comparison, the tail distributions for the queueing delay of a queue with distinct priorities and a queue without priority are also plotted in Figure 4 and Figure 5 [15]. Figure 4 shows the delay distributions of sources of several priority levels when each source has a distinct priority level. The delay distribution of sources with similar priorities are very close to each other. For example, the delay difference between priority 9, 10 and 11, are negligible. Comparing Figure **??** with Figure 4, the gap between any different priority level is obvious. Thus, it is not necessary for every single source to have a distinct priority and three or four classes is enough for providing differentiated services.



(a) T=50, N=41

Figure 3. Delay distributions of nD/D/1 queue with four service classes(T=50,N=41)

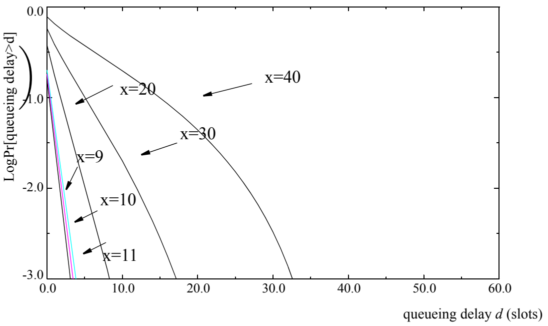


Figure 4. Delay distributions of nD/D/1 queue with distinct priorities(T=50,N=41)

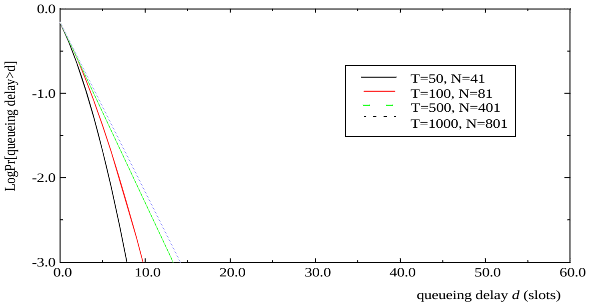


Figure 5. Delay distributions of nD/D/1 queue without differentiated services

The delay distributions of a queueing system without any priority is given in Figure 5. Comparing the curves in Figure 3 with the curve of $T = 50$ and $N = 41$ case in Figure 5, 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 their requirements. The network resources are efficiently and scientifically used.

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 i, L_{max} denotes the maximum packet length for all sessions except session i, r denotes the outgoing link capacity. For comparison, we assume 1 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 $\rho_i = r/N = 0.024kb/s$.

Table I
THE DELAY BOUNDS OF SCHEDULING ALGORITHMS

Scheduling Algorithm	Queueing Latency	Deterministic Delay Bound
WFQ [4]	$\frac{L_i}{\rho_i} - \frac{L_i}{r} + \frac{L_{max}}{r}$	41s
$SCFQ$ [5]	$\frac{L_i}{\rho_i} - \frac{L_i}{r} + (N-1)\frac{L_{max}}{r}$	80s
SFQ [7]	$(N-1)\frac{L_{max}}{r}$	40s

As shown in Table I, the deterministic queueing delay is much 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 mathematical studies are provided. The results also show that the performance can be effectively differentiated by P-FCFS 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] ITU Study Group 13, *Traffic Control and Congestion Control in B-ISDN*, Draft Recommendation I.371, May 1996.

[2] ATM Forum, *ATM User-network Interface Specification*, Version 3.1, September 1994.

[3] ATM Forum, *Traffic Management Specification*, Version 4.0, February 1996.

[4] A. K. Parekh and R. G. Gallerger, *A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single Node Casde*, IEEE/ACM Transaction on Networking, Vol. 1 No.3, 1993.

[5] S. J. Golestani, *A Self-Clocked Fair Queueing Scheme for Broadband Applications*, Proceedings of IEEE INFOCOM 1994.

[6] D. Stiliadis and A. Varma, *Latency-Rate Servers: A General Model for Analysis of Traffic Scheduling Algorithms*, IEEE/ACM Transactions on Networking, Vol.6 No.5, 1998.

[7] P. Goyal, H. M. Vin and H. Cheng, *Start-Time Fair Queueing: A Scheduling Algorithm for Integrated Services Packet Switching Networks* Technical Report TR-96-02, Department of Computer Sciences, The University of Texas at Austin, January 1996.

[8] S. Lu, V. Raghunathan, and R. Srikant, *Fair Scheduling in Wireless Packet Networks*, IEEE/ACM Trans. Network, 7(4), 473-489, 1999.

[9] V. Raghunathan, S. Ganeriwal, M. Srivastava and C. Schurgers, *Energy Efficient Wireless Packet Scheduling and Fair Queueing*, ACM Trans. Embedded Comput. Sys. 391), 3-23, 2004.

[10] Guven Mercankosk, *Mathematical Preliminaries for Rate Enforced ATM Access*, Technical Memorandum of Australian Telecommunications Research Institute (ATRI), NRL-TM-071, July, 1995.

[11] J. W. Roberts, *Performance evaluation and Design of Multiserverce Networks*, Final Report of Management Committee of COST 224 Project, 1992.

[12] P. Humblet, A. Bhargava and M. G. Hluchyj, *Ballot Theorems Applied to the Transient Analysis of nD/D/1 Queues* IEEE/ACM Transactions on Networking, Vol.1 No.1, 1993.

[13] G. Mercankosk and Z. L. Budrikis, *Establishing a Real-Time VBR Connection over an ATM Network*, IEEE GLOBECOM 96, pp1710-1714, 1996.

[14] G. Mercankosk, *Access Delay Analysis for Extended Distributed Queueing* Technical Memorandum of Australian Telecommunications Research Institute (ATRI), NRL-TM-022, July, 1995.

[15] H. Zhou, *Real-Time Services over High Speed Networks* Ph.D thesis, Curtain University of Technology, 2002.

[16] Katsuyoshi Iida, Tetsuya Takine, Hideki Sunahara and Yuji Oie, *Delay Analysis for CBR Traffic Under Static-Priority Scheduling*, IEEE/ACM Transactions on Networking, Vol.9 No.2, April 2001.

[17] Kitti Wongthavarawat and Aura Ganz, *Packet Scheduling for QoS Support in IEEE 802.16 Broadband Wireless Access Systems*, International Journal of Communication Systems, 2003.

[18] IEEE 802.16 Standard, *Local and Metropolitan Area Networks*, Part 16 IEEE Draft P802.16-14

[19] L. Georgiadis, R. Guerin, and A. Parekh, *Optimal Multiplexing on A Single Link: Delay and Buffer Requirements*, Proceedings of IEEE INFOCOM'94, Vol.2, 1994.