International Journal of Pure and Applied Mathematics
Volume 40 No. 2 2007, 165-183

APPROXIMATION OF THE DELAY DISTRIBUTION IN BATCH ARRIVAL M/G/1 PRIORITY QUEUES

Hideaki Takagi[1] §, Sang-Yong Kim[2]
[1,2]Graduate School of Systems and Information Engineering
University of Tsukuba
1-1-1 Tennoudai, Tsukuba-Shi, Ibaraki, 305-8573, JAPAN
[1]e-mail: takagi@sk.tsukuba.ac.jp
[2]e-mail: sykim@sk.tsukuba.ac.jp

Abstract: We propose new approximate formulas for the distribution functions of the delay in batch arrival M/G/1 nonpreemptive and preemptive resume priority queues. It is assumed that the delay consists of the waiting time and the service time of a whole batch. Our formulas are of the exponential distribution type with the coefficients being matched with the exact mean and second moment of the delay. By numerical examples, they are shown to be useful for evaluating the delay percentile, which may be used as a measure of the quality of service required by the user.

AMS Subject Classification: 60K25
Key Words: queue, batch arrival, M/G/1, priority

1. Introduction

Recently, there was a need to provide a quick method for calculating the percentile of the delay distribution (DF) in the M/G/1 priority queues. Here we mean by delay the time spent by a customer from the arrival to the departure, i.e., the sum of the waiting time and the service time. It arose in the spectrum estimation methodology for packet-switched traffic in the next generation wireless communication systems, [4, 5, 10, 11]. The M/G/1 nonpreemptive priority queue was used as a model for calculating the system capacity so as to satisfy the requirement on the quality of service (QoS) specified in terms of the mean delay as well as the delay percentile. This model may not reflect the

Received: July 6, 2007
§Correspondence author

precise transmission mechanism of the next generation wireless communication systems, but it was employed in the Draft New Recommendation of ITU-R Working Party 8F [5] because the explicit formula of the mean delay is available for a system of several classes of customers with different arrival rates and service time distributions. However, no explicit formula is available for the DF of the delay which is needed to calculate the delay percentile. Thus certain approximation methods are proposed for obtaining the delay distribution in [4] and [11].

In [4], the service time is assumed to be discrete as observed often in the Internet traffic. The approximate DF for the waiting time is given in the form of an exponential distribution with the coefficients determined by matching with the available exact mean and second moment. Then the delay distribution is obtained by convoluting the distributions of the waiting time and the service time. In [11], the approximate DF for the delay is directly given in the form of (12) below. However, neither of these studies shows validation of the approximation. Therefore, it is the purpose of this paper to present the validation by comparison with exact results when they are available in some cases of single class queues and the validation by simulation in other cases.

In this paper, we consider not only nonpreemptive priority queues as in [4, 11] but also preemptive resume priority queues in both of which explicit expressions are available for the mean and the second moment of the delay. Furthermore, we consider batch arrival queues, because they seem to model the real transmission mechanism more appropriately than single arrival ones. For example, a session of file transfer consists of a number of IP packets, each of which is decomposed into several radio frames when transmitted over the wireless channel, [1].

The rest of the paper is organized as follows. In Section 2, we first describe the queueing models, and then show the expressions for the mean and the second moment of the delay. In Section 3, we present an approximate form of the DF for the delay, and show how to determine its coefficients in terms of the exact mean and second moment. We compare the approximate and exact DFs for single class queues. We also show the comparison between our approximate DF and the result of simulation for priority queues with several classes. We conclude in Section 4 that the accuracy of our approximation is good enough to calculate the delay percentile at high values.

2. Batch Arrival M/G/1 Priority Queues

In this section, we describe the models of M/G/1 nonpreemptive priority and preemptive resume priority queues with batch arrivals. We then show the exact expressions for the mean and the second moment of the delay.

2.1. Queueing Models

In a batch arrival M/G/1 priority queue, there are N classes of customers indexed $i = 1, 2, \ldots, N$, where class i has priority over class j if $i < j$. Batches of customers of class i arrive in a Poisson process at rate λ_i [batches/sec], each batch containing $\mathcal{G}_i$ customers. The factorial moments of $\mathcal{G}_i$ are given by

$$g_i := E[\mathcal{G}_i], \quad g_i^{(2)} := E[\mathcal{G}_i(\mathcal{G}_i - 1)], \quad g_i^{(3)} := E[\mathcal{G}_i(\mathcal{G}_i - 1)(\mathcal{G}_i - 2)], \ \ldots \qquad 1 \le i \le N.$$

The service time $\mathcal{B}_i$ [sec] of each customer of class i has moments given by

$$b_i := E[\mathcal{B}_i], \quad b_i^{(2)} := E[\mathcal{B}_i^2], \quad b_i^{(3)} := E[\mathcal{B}_i^2], \quad \ldots \qquad 1 \le i \le N.$$

Let the traffic intensity of customers of classes $i \le n$ be

$$\rho_{\le n} := \sum_{i=1}^{n} \lambda_i g_i b_i \qquad 1 \le n \le N.$$

It is assumed that $\rho_{\le N} < 1$ for the stability of the queueing system. Regarding the service discipline, we assume either nonpreemptive or preemptive resume priority discipline. This queue is analyzed by Takagi and Takahashi [9]; see also [8, Section 3.5, p. 373].

For a batch of customers of class n, we denote by $\mathcal{W}_n$ the waiting time of the first customer in a batch, by $\mathcal{T}_n$ the time interval from the start of service for the first customer to the end of service for the last customer in a batch, and by $\mathcal{D}_n$ the delay of the last customer in a batch, i.e., the delay of the whole batch (see Figure 1). Then we have

$$\mathcal{D}_n = \mathcal{W}_n + \mathcal{T}_n, \tag{1}$$

where $\mathcal{W}_n$ and $\mathcal{T}_n$ are independent. Let us denote the mean and the second moment of these random variables as

$$D_n := E[\mathcal{D}_n], \quad W_n := E[\mathcal{W}_n], \quad T_n := E[\mathcal{T}_n]$$

and

$$D_n^{(2)} := E[\mathcal{D}_n^2], \quad W_n^{(2)} := E[\mathcal{W}_n^2], \quad T_n^{(2)} := E[\mathcal{T}_n^2].$$

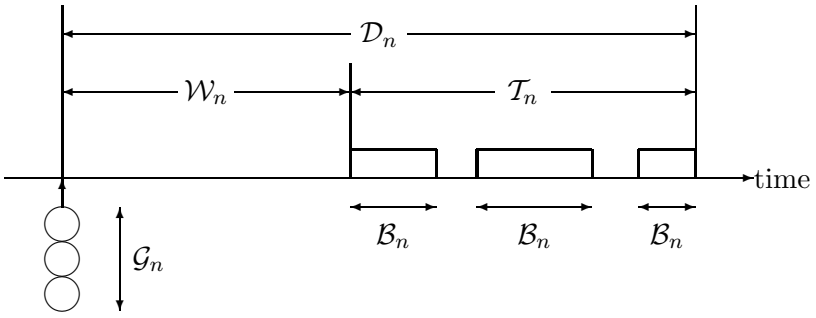


Figure 1: Delay of a customomer of class n in a batch arrival priority queue

Then we have

$$D_n = W_n + T_n \tag{2}$$

and

$$D_n^{(2)} = W_n^{(2)} + 2W_nT_n + T_n^{(2)}. \tag{3}$$

In [9], the waiting time $\mathcal{W}_n$ of the first customer in a batch is studied as well as the waiting time of an *arbitrary* customer in the batch. However, the delay $\mathcal{D}_n$ of the *last* customer in the batch is not considered. Therefore, in the Appendix A of this paper, we give the Laplace-Stieltjes transforms (LST) of the DF for $\mathcal{T}_n$ in (20) for a nonpreemptive priority queue and in (21) for a preemptive resume priority queue.

2.2. Delay Moments in a Nonpreemptive Priority Queue

For a batch arrival M/G/1 nonpreemptive priority queue, from [8, p. 378] and [9], we have

$$W_n = \frac{\displaystyle\sum_{i=1}^{n} \lambda_i g_i^{(2)} b_i^2 + \sum_{i=1}^{N} \lambda_i g_i b_i^{(2)}}{2(1-\rho_{\le n-1})(1-\rho_{\le n})} \tag{4}$$

and

$$W_n^{(2)} = \frac{\displaystyle\sum_{i=1}^{n} \lambda_i \left(g_i^{(3)} b_i^3 + 3 g_i^{(2)} b_i b_i^{(2)} \right) + \sum_{i=1}^{N} \lambda_i g_i b_i^{(3)}}{3(1-\rho_{\le n-1})^2 (1-\rho_{\le n})}$$

$$+ \left[\frac{\sum_{i=1}^{n-1} \lambda_i \left(g_i^{(2)} b_i^2 + g_i b_i^{(2)} \right)}{(1-\rho_{\le n-1})^2} + \frac{\sum_{i=1}^{n} \lambda_i \left(g_i^{(2)} b_i^2 + g_i b_i^{(2)} \right)}{(1-\rho_{\le n-1})(1-\rho_{\le n})} \right] W_n. \quad (5)$$

From (20), we can derive

$$T_n = \frac{(g_n - 1) b_n}{1 - \rho_{\le n-1}} + b_n \quad (6)$$

and

$$T_n^{(2)} = \frac{(g_n - 1) b_n \sum_{i=1}^{n-1} \lambda_i \left(g_i^{(2)} b_i^2 + g_i b_i^{(2)} \right)}{(1-\rho_{\le n-1})^3} + \frac{(g_n^{(2)} - 2g_n + 2) b_n^2 + (g_n - 1) b_n^{(2)}}{(1-\rho_{\le n-1})^2} + \frac{2(g_n - 1) b_n^2}{1 - \rho_{\le n-1}} + b_n^{(2)}. \quad (7)$$

Then we can obtain D_n and $D_n^{(2)}$ by substituting these expressions into (2) and (3), respectively.

2.3. Delay Moments in a Preemptive Resume Priority Queue

For a batch arrival M/G/1 preemptive resume priority queue, from [8, p. 381] and [9], we have

$$W_n = \frac{\sum_{i=1}^{n} \lambda_i \left(g_i^{(2)} b_i^2 + g_i b_i^{(2)} \right)}{2(1-\rho_{\le n-1})(1-\rho_{\le n})} \quad (8)$$

and

$$W_n^{(2)} = \frac{\sum_{i=1}^{n} \lambda_i \left(g_i^{(3)} b_i^3 + 3 g_i^{(2)} b_i b_i^{(2)} + g_i b_i^{(3)} \right)}{3(1-\rho_{\le n-1})^2 (1-\rho_{\le n})} + \left[\frac{\sum_{i=1}^{n-1} \lambda_i \left(g_i^{(2)} b_i^2 + g_i b_i^{(2)} \right)}{(1-\rho_{\le n-1})^2} + \frac{\sum_{i=1}^{n} \lambda_i \left(g_i^{(2)} b_i^2 + g_i b_i^{(2)} \right)}{(1-\rho_{\le n-1})(1-\rho_{\le n})} \right] W_n. \quad (9)$$

From (21), we can derive

$$T_n = \frac{g_n b_n}{1 - \rho_{\le n-1}} \tag{10}$$

and

$$T_n^{(2)} = \frac{g_n b_n \sum_{i=1}^{n-1} \lambda_i \left(g_i^{(2)} b_i^2 + g_i b_i^{(2)} \right)}{(1 - \rho_{\le n-1})^3} + \frac{g_n^{(2)} b_n^2 + g_n b_n^{(2)}}{(1 - \rho_{\le n-1})^2}. \tag{11}$$

Then we can obtain D_n and $D_n^{(2)}$ by substituting these expressions into (2) and (3), respectively.

3. Approximation of the Delay Distribution

In this section, we present an approximation method for the DF of the delay in M/G/1 nonpreemptive priority and preemptive resume priority queues with batch arrivals. Our method is based on the matching of two coefficients in the approximate DF of an exponential distribution type with the exact mean and second moment of the delay for each class of customers. The accuracy of our approximation is validated by comparison with exact results when they are available in some cases of single class queues and with simulation results in the cases of multiple class priority queues.

3.1. Approximate Distribution Function

Let us propose a new approximate form for the DF of the delay (waiting time plus service time) $\mathcal{D}$ with three parameters:

$$D(t) := P\{\mathcal{D} \le t\} = \begin{cases} 0\,, & 0 \le t < \beta\,, \\ 1 - qe^{-\gamma(t-\beta)}\,, & t \ge \beta\,, \end{cases} \tag{12}$$

where β is the minimum service time. We determine the parameters q and γ in (12) so that:

(a) It has the specified mean:

$$D := E[\mathcal{D}] = \int_0^\infty [1 - D(t)]dt = \beta + \frac{q}{\gamma}.$$

(b) It has the specified second moment:

$$D^{(2)} := E[\mathcal{D}^2] = 2\int_0^\infty t[1 - D(t)]dt = \beta^2 + \frac{2q\beta}{\gamma} + \frac{2q}{\gamma^2}.$$

Hence we get

$$q = \frac{2(D-\beta)^2}{D^{(2)} - 2D\beta + \beta^2} \quad \text{and} \quad \gamma = \frac{2(D-\beta)}{D^{(2)} - 2D\beta + \beta^2}. \tag{13}$$

Since $D^{(2)} > D^2$ and $D > \beta$, it follows that $q > 0$ and $\gamma > 0$. However, it does not always hold that $q < 1$.

The percentile of the delay distribution can be obtained as follows. Let $\pi(r)$ denote the rth percentile of the DF of $\mathcal{D}$ in (12). Then, from the relation

$$1 - qe^{-\gamma[\pi(r)-\beta]} = \frac{r}{100}$$

we get

$$\pi(r) = \beta + \frac{1}{\gamma} \log_e \left(\frac{100q}{100 - r} \right). \tag{14}$$

We note that the DF for the delay $\mathcal{D}$ in (12) is exact for the M/M/1 queue. For an M/M/1 queue with arrival rate λ and service rate μ, the DF for $\mathcal{D}$ is given by

$$D(t) = 1 - e^{-(\mu-\lambda)t}, \quad t \geq 0, \tag{15}$$

from which we have $D = 1/(\mu - \lambda)$ and $D^{(2)} = 2/(\mu - \lambda)^2$. Since $\beta = 0$, it follows from (13) that $q = 1$ and $\gamma = \mu - \lambda$. Hence, $D(t)$ in (12) is identical to (15). This is not surprising because $D(t)$ in (15) has only two parameters λ and μ so that the matching of two moments D and $D^{(2)}$ should lead to the exact result.

A comment may be in order on the simplistic approximation in (12). There are many sophisticated approximations for the DF of the waiting time in the M/G/1 queue without and with batch arrivals. For example, see [12, Sections 4.2 and 4.3]. Such an approximate DF could be used for the DF of $\mathcal{W}_n$ in our problem. However, we also need the explicit DF of $\mathcal{T}_n$, which is not trivial from (20) and (21) below for priority queues. Then the numerical convolution of the DF for $\mathcal{W}_n$ and that for $\mathcal{T}_n$ would produce the DF for $\mathcal{D}_n$. This method could bring more accurate DF for $\mathcal{D}_n$, but it requires significant computational time. Thus it is not suitable for quick evaluation which is mandatory for our purpose mentioned in Section 1. Hence we have chosen the simple approximation given in (12).

3.2. Numerical Test for Batch Arrival M/G/1 Queues

The LST of the DF for the delay $\mathcal{D}$ in a single class batch arrival M/G/1 queue is given by

$$D^*(s) := \int_0^\infty e^{-st} dD(t) = \frac{(1-\rho)sG[B(s)]}{s - \lambda + \lambda G[B(s)]}, \tag{16}$$

where $G(z)$ is the probability generating function (PGF) for the batch size, $B(s)$ is the LST of the DF for the service time of each customer, and $\rho = \lambda g b$. As far as the delay of the whole batch is concerned, this queue is equivalent with a single arrival M/G/1 queue where each customer corresponds to the whole batch in which the LST of the DF for the service time is given by $G[B(s)]$. Equation (16) yields

$$D = \frac{\lambda\left(g^{(2)}b^2 + gb^{(2)}\right)}{2(1-\rho)} + gb$$

and

$$D^{(2)} = \frac{\lambda\left(g^{(3)}b^3 + 3g^{(2)}bb^{(2)} + gb^{(3)}\right)}{3(1-\rho)} + \frac{\lambda^2\left(g^{(2)}b^2 + gb^{(2)}\right)^2}{2(1-\rho)^2} + \frac{g^{(2)}b^2 + gb^{(2)}}{1-\rho}.$$

It is not always possible to obtain the explicit inverse transform of (16). In the following, we compare the approximate $D(t)$ in (12) with the exact $D(t)$ obtained by inverting (16) in some cases in which such inversion is possible.

(a) M/D/1. For an M/D/1 queue with fixed service time b, we have $G(z) = 1$ and $B(s) = e^{-bs}$ so that

$$D(t) = (1-\rho)e^{\lambda t}\sum_{i=1}^{\lfloor t/b \rfloor} \frac{(i\rho - \lambda t)^{i-1}}{(i-1)!} e^{-i\rho}, \quad t \geq b, \tag{17}$$

and $D(t) = 0$ for $t < b$, where $\rho = \lambda b$ and $\lfloor x \rfloor$ denotes the largest integer not exceeding x (floor function). This expression was obtained by Agner Krarup Erlang in 1909, as noted in [3] and [7, p. 220]. In Appendix B of this paper, we show a different derivation. Since

$$\frac{D}{b} = \frac{2-\rho}{2(1-\rho)}, \quad \frac{D^{(2)}}{b^2} = \frac{6 - 4\rho + \rho^2}{6(1-\rho)^2}, \quad \text{and} \quad \beta = b,$$

it follows that the parameters in the approximate DF $D(t)$ in (12) are given by

$$q = \frac{3\rho}{2+\rho} < 1 \quad \text{and} \quad \frac{\gamma}{b} = \frac{6(1-\rho)}{2+\rho} > 0$$

when $\rho < 1$.

(b) M^k/M/1 and M/E_k/1. For an M^k/M/1 queue, where the batch size k is a fixed positive integer, we have $G(z) = z^k$ and $B(s) = \mu/(s+\mu)$. For an

$M/E_k/1$ queue, we have $G(z) = 1$ and $B(s) = [\mu/(s+\mu)]^k$. Thus we get the same expression

$$D^*(s) = \frac{(1-\rho)\mu[s/(s+\mu)]^k}{s - \lambda + \lambda[\mu/(s+\mu)]^k}, \tag{18}$$

where $\rho = k\lambda/\mu$. This is a rational function in s, which can be inverted explicitly in principle for any integer value of k. Note that $M/E_k/1$ reduces to $M/M/1$ for $k = 1$ and that it approaches to $M/D/1$ when k becomes infinity while keeping the mean service time $b = k/\mu$ at a fixed value. In the $M/E_k/1$ queue, we have $b = k/\mu$, $b^{(2)} = k(k+1)/\mu^2$, and $b^{(3)} = k(k+1)(k+2)/\mu^3$ so that

$$\mu D = \frac{(k+1)\rho}{2(1-\rho)} + k, \quad \mu^2 D^{(2)} = \frac{(k+1)(k+2)\rho}{3(1-\rho)} + \frac{(k+1)^2\rho^2}{2(1-\rho)^2} + \frac{k(k+1)}{1-\rho},$$

and $\beta = 0$. Parameter q in the approximate DF $D(t)$ in (12) is then given by

$$q = \frac{2D^2}{D^{(2)}} = \frac{3[2k-(k-1)\rho]^2}{(k+1)[6k-4(k-1)\rho+(k-1)\rho^2]}.$$

It can be shown that $q > 1$ if $\rho < 1$ for $k \geq 2$. In fact, the function q in ρ decreases monotonously from $q = 2k/(k+1) > 1$ at $\rho = 0$ to $q \to 1$ as $\rho \to 1-0$.

We note that making $k \to \infty$ in the above q does not lead to the expression for q in (a), because β remains null for any finite value of k. However, the DF for the $M/E_k/1$ approaches to that for the $M/D/1$ as shown in Figure 2. In Figure 2, we plot the exact $D(t)$ for $M/E_k/1$ queues with $b = 1$ and $\lambda = 0.2$ for $k = 1, 3, 10, 50$ and ∞.

(c) $M/H_2/1$. We may also consider an $M/H_2/1$ queue for which the service time has hyperexponential distribution with balanced means [12, p. 359] such that

$$B(s) = p_1\frac{\mu_1}{s+\mu_1} + p_2\frac{\mu_2}{s+\mu_2}$$

with

$$p_1 = \frac{1}{2}\left(1 + \sqrt{\frac{c^2-1}{c^2+1}}\right), \quad p_2 = 1 - p_1, \quad \text{and} \quad \frac{p_1}{\mu_1} = \frac{p_2}{\mu_2} = \frac{b}{2}.$$

Here c^2 (≥ 1) denotes the squared coefficient of variation for the service time. Since $G(z) = 1$, we get

$$D^*(s) = \frac{(1-\rho)s[(p_1\mu_1 + p_2\mu_2)s + \mu_1\mu_2]}{(s-\lambda)(s+\mu_1)(s+\mu_2) + \lambda[(p_1\mu_1 + p_2\mu_2)s + \mu_1\mu_2]}, \tag{19}$$

where $\rho = \lambda b$. This is also a rational function in s which can be inverted easily. Note that $M/H_2/1$ reduces to $M/M/1$ for $c^2 = 1$. In the $M/H_2/1$ queue, we

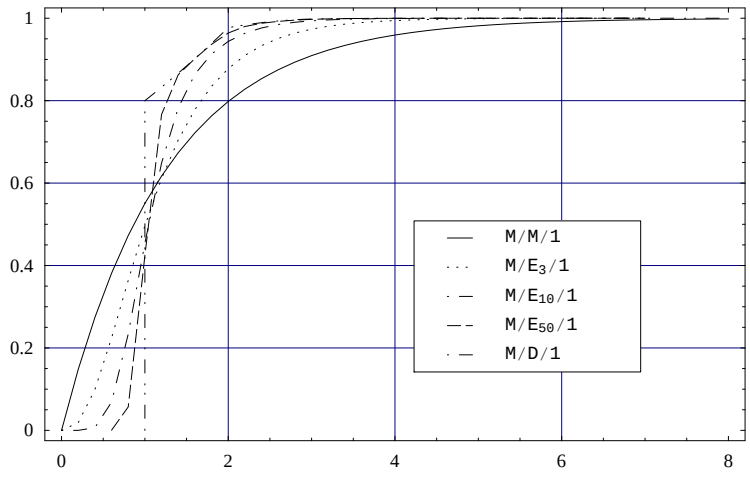


Figure 2: Delay distribution $D(t)$ in M/E$_k$/1 queues ($\lambda = 0.2, b = 1$)

have $b^{(2)} = (c^2+1)b^2$ and $b^{(3)} = 3c^2(c^2+1)b^3$ so that

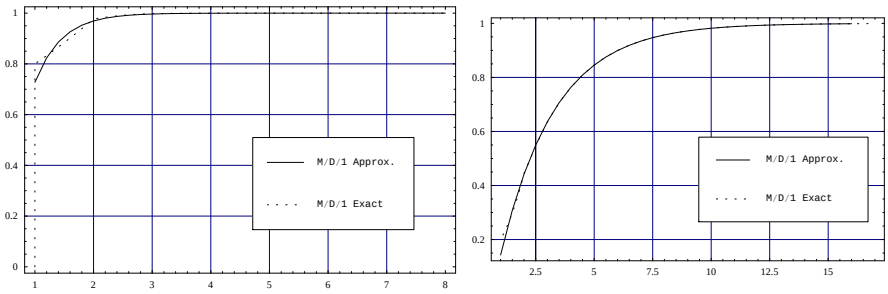
$$\frac{D}{b} = \frac{(c^2+1)\rho}{2(1-\rho)} + 1, \quad \frac{D^{(2)}}{b^2} = \frac{c^2(c^2+1)\rho}{1-\rho} + \frac{(c^2+1)^2\rho^2}{2(1-\rho)^2} + \frac{c^2+1}{1-\rho},$$

and $\beta = 0$. Parameter q in the approximate DF $D(t)$ in (12) is then given by

$$q = \frac{2D^2}{D^{(2)}} = \frac{[2+(c^2-1)\rho]^2}{(c^2+1)[2+2(c^2-1)\rho-(c^2-1)\rho^2]}.$$

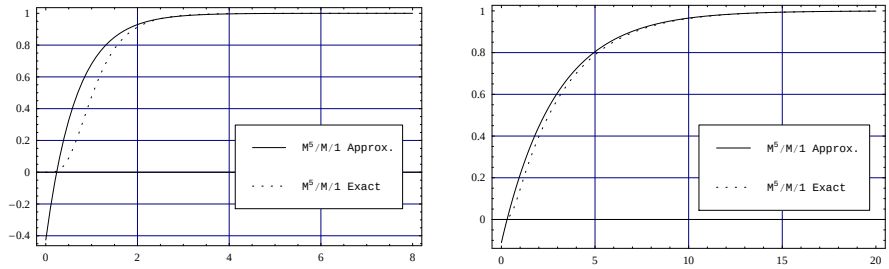
It can be shown that $q < 1$ if $\rho < 1$ for $c^2 > 1$. In fact, the function q in ρ increases monotonously from $q = 2/(c^2+1) < 1$ at $\rho = 0$ to $q \to 1$ as $\rho \to 1-0$.

Let us examine the accuracy of our approximation numerically in these cases. In Figure 3, we compare the exact $D(t)$ given in (17) and the approximate $D(t)$ in (12) for M/D/1 queues with $b = 1$ and $\lambda = 0.2$ and 0.8. The accuracy of approximation is good in this case. In Figure 4, we compare the exact $D(t)$ inverted from (18) and the approximate $D(t)$ in (12) for M^k/M/1 queues with $k = 5$, $\mu = 5$, and $\lambda = 0.2$ and 0.8. In this case, the approximate $D(t)$ becomes negative for small t owing to $q > 1$. However, the accuracy is reasonable for large t when $D(t)$ is close to 1. Thus our approximation can be used for obtaining, say, the 95 percentile of the delay distribution. In Figure 5, we plot the exact $D(t)$ inverted from (19) and the approximate $D(t)$ in (12) for M/H$_2$/1 queues with $b = 1$ and $\lambda = 0.2$ and 0.8 for $c^2 = 3$ and 5. In this case, the accuracy of approximation seems to be good for the most domain of $t \geq 0$.



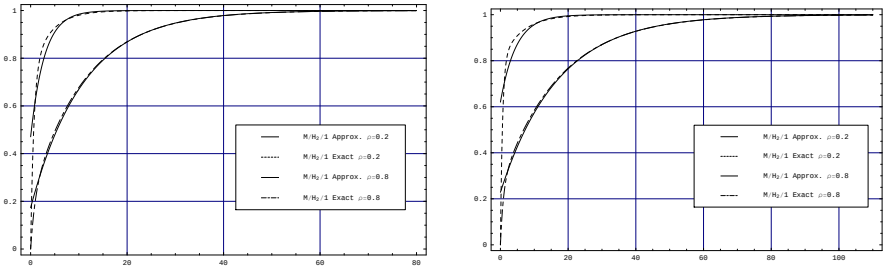
(a) $\rho = 0.2$ (b) $\rho = 0.8$

Figure 3: Delay distribution $D(t)$ in M/D/1 queues



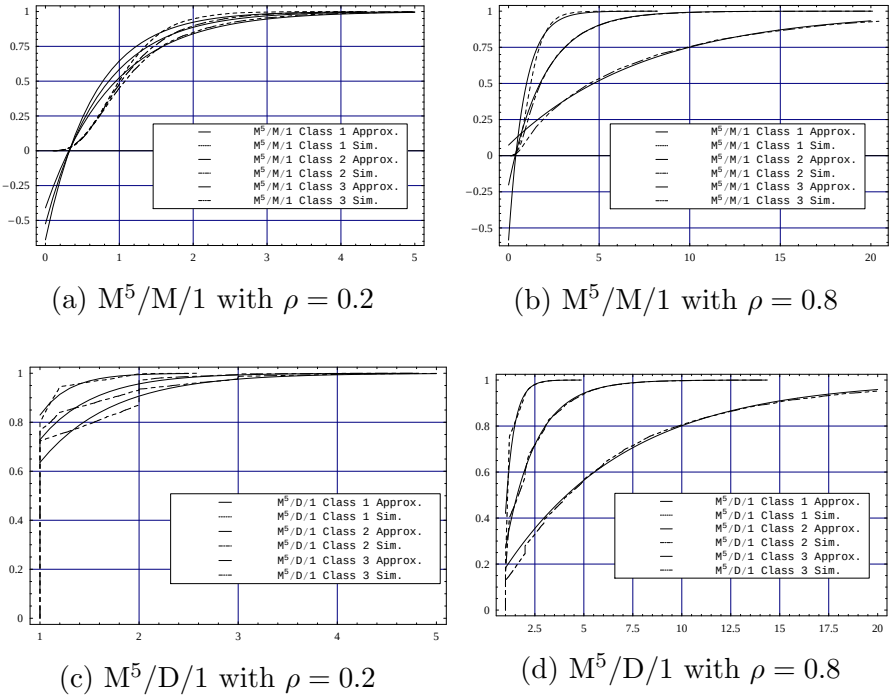
(a) $\rho = 0.2$ (b) $\rho = 0.8$

Figure 4: Delay distribution $D(t)$ in $M^5/M/1$ queues



(a) $c^2 = 3$ (b) $c^2 = 5$

Figure 5: Delay distribution $D(t)$ in $M/H_2/1$ queues



(a) $M^5/M/1$ with $\rho = 0.2$ (b) $M^5/M/1$ with $\rho = 0.8$

(c) $M^5/D/1$ with $\rho = 0.2$ (d) $M^5/D/1$ with $\rho = 0.8$

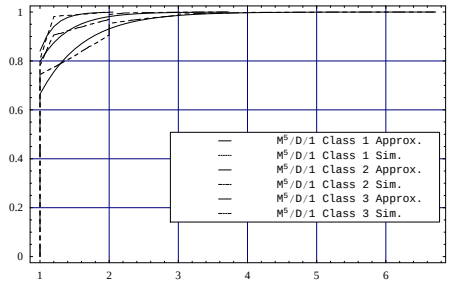
Figure 6: Delay distribution $D_n(t)$ in nonpreemptive priority queues

3.3. Numerical Test for Batch Arrival M/G/1 Priority Queues

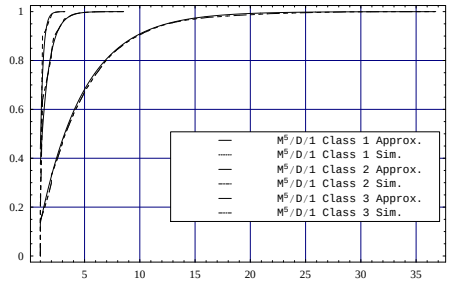
For priority queues with several classes, we do not have the exact explicit expression for the delay distribution $D_n(t) := P\{D_n \le t\}$. Therefore, we compare the approximate $D_n(t)$ given in (12) for each class with the result of simulation. In so doing, let us assume for the sake of simplicity that the batch size and the service time distribution are identical for all classes.

Our simulation was conducted by *OPNET*, Version 11.0A PL3 [14].

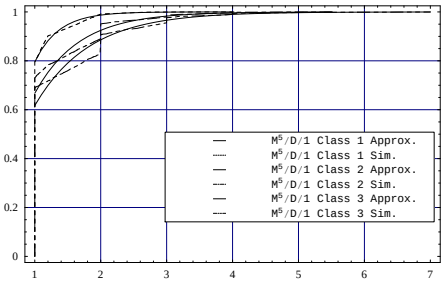
In Figure 6, we plot the $D_n(t)$ from simulation and the approximate $D_n(t)$ for $M^5/M/1$ and $M^5/D/1$ nonpreemptive priority queues with 3 classes, where $b = 0.2$ so that $\lambda = \rho/3$ for every class. For $M^5/M/1$ queues shown in Figures 6(a) and (b), the approximate $D_n(t)$ becomes negative for small t as in Figure 4. However, the accuracy is acceptable for large t when $D_n(t)$ is close to 1. For $M^5/D/1$ queues shown in Figures 6(c) and (d), the accuracy is not bad. In both queues, the accuracy seems better for larger values of ρ.



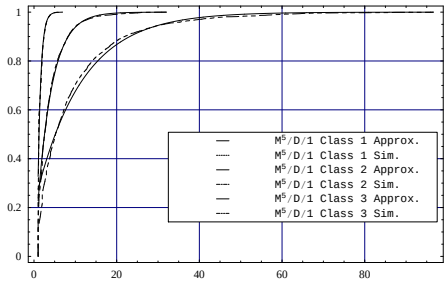
(a) $M^5/D/1$ with $\rho = 0.2$ (ratio 1, 2, 4)



(b) $M^5/D/1$ with $\rho = 0.8$ (ratio 1, 2, 4)



(c) $M^5/D/1$ with $\rho = 0.2$ (ratio 4, 2, 1)



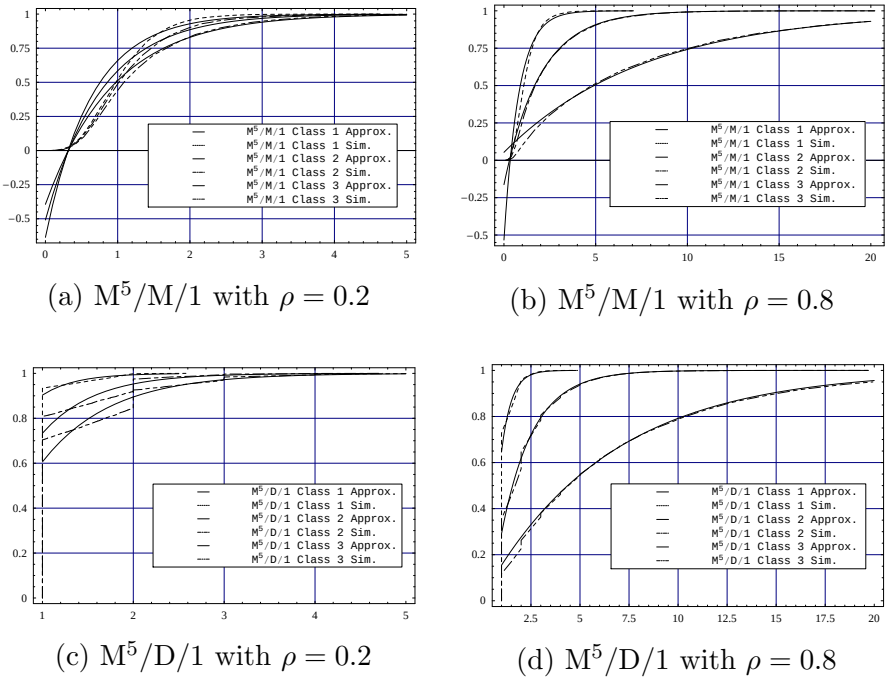
(d) $M^5/D/1$ with $\rho = 0.8$ (ratio 4, 2, 1)

Figure 7: Delay distribution $D_n(t)$ in nonpreemptive priority queues with different traffic intensities

In Figure 7, we show similar plot for $M^5/D/1$ nonpreemptive priority queues with different arrival rate for each class. Namely, we assume $\lambda_1 : \lambda_2 : \lambda_3 = 1 : 2 : 4$ in Figures 7(a) and (b), and $\lambda_1 : \lambda_2 : \lambda_3 = 4 : 2 : 1$ in Figures 7(c) and (d). The accuracy of approximation is good in these cases.

In Figure 8, we plot the $D_n(t)$ from simulation and the approximate $D_n(t)$ for $M^5/M/1$ and $M^5/D/1$ preemptive resume priority queues with 3 classes. We can make the same observation as in Figure 6. Again, the accuracy is better for larger values of ρ in both queues.

In Figure 9, we show similar plot for $M^5/D/1$ preemptive resume priority queues with different arrival rate for each class. We assume $\lambda_1 : \lambda_2 : \lambda_3 = 1 : 2 : 4$ in Figures 9(a) and (b), and $\lambda_1 : \lambda_2 : \lambda_3 = 4 : 2 : 1$ in Figures 9(c) and (d). The accuracy is good when the traffic intensity is high.



(a) $M^5/M/1$ with $\rho = 0.2$ (b) $M^5/M/1$ with $\rho = 0.8$

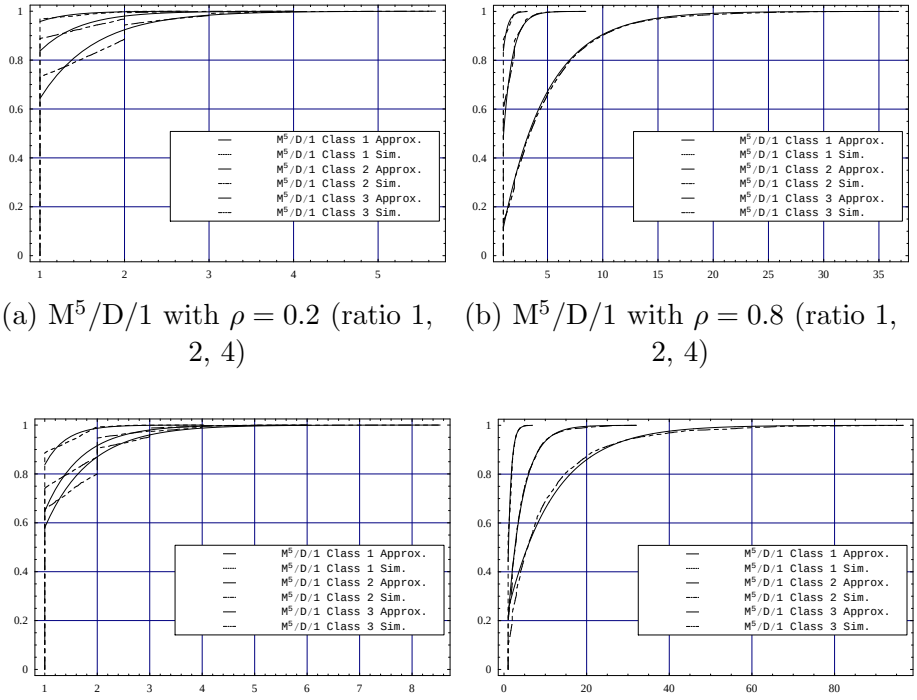
(c) $M^5/D/1$ with $\rho = 0.2$ (d) $M^5/D/1$ with $\rho = 0.8$

Figure 8: Delay distribution $D_n(t)$ in preemptive resume priority queues

From the numerical results shown in Figures 3 through 9, we generally observe that our approximate delay distribution given in (12) may not be accurate at small values of the delay especially when the traffic intensity is low. However these are cases for which the accurate evaluation of delay percentage is not critical in the application we have in mind. On the other hand, the accuracy our approximation is good at large values of the delay (e.g., 90 percentile) when the traffic intensity is high.

4. Conclusion

In this paper, we have proposed an approximation method for obtaining the distribution function of the delay for each class in batch arrival M/G/1 nonpreemptive and preemptive resume priority queues. We have shown validation of our approximation by comparison with exact results as well as simulation results. We have observed that the accuracy is good enough to calculate the delay



(a) $M^5/D/1$ with $\rho = 0.2$ (ratio 1, 2, 4)

(b) $M^5/D/1$ with $\rho = 0.8$ (ratio 1, 2, 4)

(c) $M^5/D/1$ with $\rho = 0.2$ (ratio 4, 2, 1)

(d) $M^5/D/1$ with $\rho = 0.8$ (ratio 4, 2, 1)

Figure 9: Delay distribution $D_n(t)$ in preemptive resume priority queues with different traffic intensities

percentile at high values. While the approximation method in [4] is developed for the single arrival M/G/1 nonpreemptive priority queue with discrete service times, our method can be applied to batch arrival M/G/1 nonpreemptive and preemptive resume priority queues with generally distributed service times.

In the present work, we have limited our validation to the cases in which the batch size and the service time distribution are the same for all classes of customers. In order to demonstrate the robustness of our method in a variety of situations, further validation is necessary for the cases in which the traffic parameters differ significantly among the classes. Through such extensive validation we will be able to recommend our approximation method for the delay percentile calculation in communication systems, as in [11], which was the original motivation of the present study.

References

[1] 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Physical layer aspects of UTRA high speed downlink packet access, 3GPP TR25.848 V 4.0, 2001-03.

[2] V.E. Benes, On queues with Poisson arrivals, *Annals of Mathematical Statistics*, **28**, No. 3 (1957), 670-677.

[3] E. Brockmeyer, H.L. Halstrøm, A. Jensen, The life and works of A.K. Erlang, *Transactions of the Danish Academy of Technical Sciences*, No. 2 (1948), 133.

[4] T. Irnich, B. Walke, Spectrum estimation methodology for next generation wireless systems, In: *The 15-th Annual IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC 2004)*, Barcelona (2004), 5-8.

[5] ITU-R, Methodology for calculation of spectrum requirements for the future development of the terrestrial component of IMT-2000 and systems beyond IMT-2000, Recommendation ITU-R M.1768 (2006).

[6] L. Kleinrock, *Queueing Systems, Volume 2: Computer Applications*, John Wiley and Sons, New York (1976).

[7] R. Syski, *Introduction to Congestion Theory in Telephone Systems*, Second Edition, Elsevier, Amsterdam (1986).

[8] H. Takagi, *Queueing Analysis: Foundation of Performance Evaluation, Volume 1: Vacation and Priority Systems*, Elsevier, Amsterdam (1991).

[9] H. Takagi, Y. Takahashi, Priority queues with batch Poisson arrivals, *Operations Research Letters*, **10**, No. 4 (1991), 225-232.

[10] H. Takagi, H. Yoshino, N. Matoba, M. Azuma, Methodology for calculation of spectrum requirements for the next generation mobile communication systems, *IEICE Transactions on Communications*, **J-89-B**, No. 2 (2006), 135-142, In Japanese.

[11] H. Takagi, H. Yoshino, N. Matoba, M. Azuma, M. Shirakabe, System capacity calculation for packet-switched traffic in the next generation wireless systems, part II: batch arrival M/G/1 nonpreemptive priority queueing model for transmission over a radio channel, In: *Performance Challenges*

for Efficient Next Generation Networks: Proceedings of the 19-th International Tetetraffic Congress - ITC19 (Ed-s: X. Liang, Z. Xin, V.B. Iversen, G.S. Kuo), Beijing (2005), 21-30.

[12] H.C. Tijms, *Stochastic Models: An Algorithmic Approach*, John Wiley and Sons, Chichester (1994).

[13] S.S. Wilks, *Mathematical Statistics*, John Wiley and Sons, New York (1962).

[14] http://www.opnet.com

Appendix A: Distribution of $\mathcal{T}_n$ in Batch Arrival M/G/1 Priority Queues

In a batch arrival M/G/1 priority queue described in Section 2, recall that $\mathcal{T}_n$ denotes the time interval from the start of service for the first customer to the end of service for the last customer in a batch of class n.

We first consider a nonpreemptive priority queue. If the batch contains k customers, $\mathcal{T}_n$ consists of (i) $k-1$ *delay cycles* [6, Section 3.3, p. 111], each of which consists of the *initial delay* for the service of a customer of class n and the subsequent *delay busy periods* generated by the customers of class 1 through $n-1$, and (ii) the uninterrupted service time for the last customer. This consideration leads to the following LST of the DF for $\mathcal{T}_n$:

$$T_n(s) := \int_0^\infty e^{-st} dP\{\mathcal{T}_n \le t\} = \frac{G_n\{B_n[\sigma_{\le n-1}(s)]\}}{B_n[\sigma_{\le n-1}(s)]} B_n(s), \tag{20}$$

where $B_n(s)$ is the LST of the DF for $\mathcal{B}_n$, the service time of a customer of class n, $G_n(z)$ is the PGF of $\mathcal{G}_n$, the number of customers included in a batch of class n, and

$$\sigma_{\le n-1}(s) := s + \lambda_{\le n-1} - \lambda_{\le n-1}\Theta_{\le n-1}(s)$$

with

$$\lambda_{\le n-1} := \sum_{i=1}^{n-1} \lambda_i \ .$$

Furthermore, $\Theta_{\le n-1}(s)$ is the LST of the DF for the duration $\Theta_{\le n-1}$ of a busy period generated by the customers of classes 1 through $n-1$. It is given as the solution to the equation

$$\Theta_{\le n-1}(s) = B_{\le n-1}[s + \lambda_{\le n-1} - \lambda_{\le n-1}\Theta_{\le n-1}(s)],$$

where

$$B_{\le n-1}(s) := \frac{1}{\lambda_{\le n-1}} \sum_{i=1}^{n-1} \lambda_i G_i[B_i(s)].$$

The mean T_n and the second moment $T_n^{(2)}$ given in (6) and (7) can be derived from (20).

For a preemptive resume priority queue, the LST of the DF for $\mathcal{T}_n$ is simply given by

$$T_n(s) = G_n\{B_n[\sigma_{\le n-1}(s)]\}. \tag{21}$$

Then we get the mean T_n and the second moment $T_n^{(2)}$ given in (10) and (11), respectively.

Appendix B: Derivation of $D(t)$ in an M/D/1 Queue

We derive (17) in such a way that utilizes the convolution of a uniform distribution, which is different from the derivation by Erlang shown in [7, p. 220].

The LST of the DF for the delay in the M/D/1 queue with arrival rate λ and service time b is given by

$$D^*(s) = \frac{(1-\rho)se^{-bs}}{s-\lambda+\lambda e^{-bs}}$$
$$= \frac{(1-\rho)e^{-bs}}{1-\dfrac{\lambda}{s}\left(1-e^{-bs}\right)} = (1-\rho)e^{-bs}\sum_{k=0}^{\infty}\rho^k\left(\frac{1-e^{-bs}}{bs}\right)^k, \tag{22}$$

where $\rho = \lambda b$. This expansion was first presented by Benes [2] for an M/G/1 queue. Note that $(1-e^{-bs})/(bs)$ is the LST of the DF for the residual life of the fixed service time b which has a uniform distribution over $[0, b]$. Therefore, $[(1-e^{-bs})/(bs)]^k$ is the LST of the k-fold self convolution of the uniform distribution for which the DF is given in [13, p. 204] (the result is due to Pierre Simon Laplace in 1814). Moreover, the factor e^{-bs} on the right-hand side of (22) shifts the origin of t by b. Thus we have the transform

$$e^{-bs}\left(\frac{1-e^{-bs}}{bs}\right)^k \quad \Longleftrightarrow$$

$$\begin{cases} 0\,, & t < b\,, \\ \dfrac{1}{k!b^k}\displaystyle\sum_{i=1}^{\lfloor t/b\rfloor}(-1)^{i-1}\binom{k}{i-1}(t-ib)^k\,, & b \le t < (k+1)b\,, \\ 1\,, & t \ge (k+1)b\,. \end{cases}$$

It follows that, for j such that $b \le jb \le t < (j+1)b$, we get

$$\begin{aligned} D(t) = \;& (1-\rho)\left[\sum_{k=0}^{j-1}\rho^k + \sum_{k=j}^{\infty}\frac{\rho^k}{k!b^k}\sum_{i=1}^{j}(-1)^{i-1}\binom{k}{i-1}(t-ib)^k\right] \\ &= 1-\rho^j + (1-\rho)\sum_{i=1}^{j}\frac{(-1)^{i-1}}{(i-1)!}\sum_{k=j}^{\infty}\frac{(\lambda t - i\rho)^k}{(k-i+1)!} \\ &= 1-\rho^j + (1-\rho)\sum_{i=1}^{j}\frac{(i\rho-\lambda t)^{i-1}}{(i-1)!}\left[e^{\lambda t - i\rho} - \sum_{k=0}^{j-i}\frac{(\lambda t - i\rho)^k}{k!}\right]. \end{aligned}$$

However it can be proved, say, by mathematical induction on j that

$$\sum_{i=1}^{j}\frac{(i\rho-\lambda t)^{i-1}}{(i-1)!}\sum_{k=0}^{j-i}\frac{(\lambda t - i\rho)^k}{k!} = \frac{1-\rho^j}{1-\rho}\,, \quad j \ge 1.$$

Hence we obtain (17).