Journal of the Operations Research Society of Japan
Vol. 33, No. 3, September 1990

STRUCTURED PRIORITY QUEUE WITH BATCH ARRIVALS

Yoshitaka Takahashi
NTT Laboratories

Hideaki Takagi
IBM Research

(Received June 15, 1989; Revised March 5, 1990)

Abstract This paper investigates a single-server priority queueing system with batch arrivals where an arriving batch is composed of multi-class customers. The main motivation for studying this type of priority queue is its potential applicability to communication switching systems. Preemptive resume and nonpreemptive (head-of-the-line) rules are considered. The queue length and waiting time distributions are obtained via the supplementary variable technique and delay cycle analysis. The conservation law approach is applied to find the mean delay formulas.

1. Introduction

We consider a single-server priority queue with P priority classes of customers. Customers arrive in batches so that the time between two successively arriving batches is exponentially distributed with mean $1/\lambda$ and is independent of any other events in the system. We assume that a customer with a smaller index has precedence over a customer with a greater index (class 1 is the highest, and class P the lowest). Let $\underline{K} \equiv (K_1, K_2, \cdots, K_P)$ be a random vector where component K_i signifies the number of class i customers in an arriving batch. The cumulative size of all class customers in an arriving batch is then given by $K_1 + K_2 + \cdots + K_P$. The vector $\underline{K}$ is assumed to have the identical joint distribution for individual arriving batches. The joint probability distribution of the batch size vector $\underline{K}$, $a(k_1, k_2, \cdots, k_p) \equiv P[K_1 = k_1, \cdots, K_P = k_P]$ $(k_j \geq 0, 1 \leq j \leq P)$ is arbitrary. It is convenient to assume that $a(0, 0, \cdots, 0) = 0$ and to denote the vector with an under-bar, say $\underline{k} \equiv (k_1, k_2, \cdots, k_P)$. We will refer to this priority as "structured priority," as in Sidi, et al. [15].

The special case of independent batch arrivals, as in Hawkes [4] and Meister [12], is represented by a distribution in which the only non-zero probabilities are of the form $a(0, \cdots, k_j, \cdots, 0)$ for some $j (1 \leq j \leq P)$. The case of independent single arrivals, as in Jaiswal [5, 6], is represented by a distribution in which the only non-zero probabilities are of the form $a(0, \cdots, 1, \cdots, 0)$ for some j-th component. Structured priority queues then imply the ordinary independent arrival priority queues.

The generating function of batch size vector $\underline{K}$ is now introduced by

$$G(z_1, \cdots, z_P) \equiv G(\underline{z}) \equiv \sum_{k_j \geq 0 (1 \leq j \leq P)} a(k_1, k_2, \cdots, k_P) z_1^{k_1} z_2^{k_2} \cdots z_P^{k_P}.$$

The mean and the i-th factorial moments of the class p batch size $(1 \leq p \leq P)$ are then given by

$$g_p \equiv [\frac{\partial G(\underline{z})}{\partial z_p}]_{\underline{z}=\underline{1}},$$

and

$$g_p^{(i)} \equiv [\frac{\partial^i G(\underline{z})}{\partial^i z_p}]_{\underline{z}=\underline{1}}, (i = 2, 3, \cdots)$$

where $\underline{1} \equiv (1 \cdots, 1)$. We let $g_{ij} \equiv [\frac{\partial^2 G(\underline{z})}{\partial z_i \partial z_j}]_{\underline{z}=\underline{1}}$ for $j \neq i (1 \leq i, j \leq P)$. Note that the arrival rate of the class p customers is then given by λg_p.

Ordinary priority queues with independent single or batch arrivals (where $g_{ij} = 0$ for $j \neq i (1 \leq i, j \leq P)$; see Remark 4.1) have been extensively studied in the literature [2, 7]. However, there are few studies on the structured priority queue. A discrete-time structured nonpreemptive priority queue has been analyzed by Sidi, et al. [14–16] under the assumption that the service time is constant (invariant) for individual priority classes. A continuous-time structured priority with a service process different from that in this paper has been treated in [17]. We will treat a continuous-time structured preemptive-resume and nonpreemptive priority queues, which generalize the ordinary independent arrival priority queues analyzed in [1, 4, 5, 6, 12, 18, 19, 20].

Structured priority queues have a potential applicability to practical systems. For instance, a telephone call which is assumed to arrive at a switching system according to a Poisson process with rate λ requires a few of priority tasks. The arrival processes of individual class tasks at the system are not necessarily independent of each other in this example. The telephone call corresponds to a batch (group of customers), and the task corresponds to a customer in queueing terminology. It is an important but heavy job to identify or estimate the joint generating function $G(\underline{z})$ in the telephone switching system. The second example is a data communication system [17] where a message is composed of packets. Each packet (in an arriving message at a switching node) requires two processing operations: validity checking (protocol processing) and routing processing for subsequent transmission. The former operation has a higher priority than the latter operation. We assume a message arrives at a node according to a Poisson process with rate λ, and further assume that packets in a message arrive simultaneously as in [10, 11]. The message corresponds to a batch, and the packet corresponds to a customer in the queueing terminology. The packet switching system having such a batch input can be modeled by a special structured priority queue with $G(z_1, z_2) = H(z_1 \cdot z_2)$ where $H(z)$ is the generating function of packet length (the number of packets) in a message. In transportation engineering, which is the third example, containers of a variety of commodities arrive by tracks at a marshalling yard, and may be handled in the order of priority classes associated with commodities. Here, owing to the limited capacity, the number of containers (customers) for each commodity (priority) are correlated. There are many similar situations where customers arrive in a batch from a vehicle of limited capacity. For other examples, see Stuck and Arthurs [17, Chapter 10] and Sidi, et al. [14–16].

The following notation characterizing the service process is necessary. Customers are served one at a time with independent service times. The Laplace-Stieltjes transform (LST) of the distribution function (DF), the mean, and the i-th moment for each class p customer are denoted by $B_p^*(s), b_p$, and $b_p^{(i)} (i = 2, 3 \cdots)$, respectively. The probability density function of the service times can be expressed as

$$\frac{dB_p(x)}{dx} = \eta_p(x) \cdot \exp\{-\int_0^x \eta_p(y) dy\}$$

by using $\eta_p(x)$, age-specific failure rate in renewal theory. Service is first-come-first-served within each class, but a higher class has a preemptive resume or nonpreemptive (head-of-the-line) priority over a lower class.

Traffic intensity for class p is denoted by $\rho_p \equiv \lambda g_p b_p$, and traffic intensity for lumped class $\{1, 2, \cdots, p\}$ by ρ_p^+, i.e.,

$$\rho_p^+ \equiv \sum_{j=1}^{p} \rho_j \ (1 \leq p \leq P).$$

We also use the notation

$$\rho_0^+ \equiv 0,$$

and

$$\rho \equiv \rho_P^+$$

which is assumed to be less than unity for stability.

2. Queue Length Distribution

It seems complicated to formulate our structured priority queue with general P classes via the supplementary variable technique [7, 8], so we restrict ourselves to the case of $P = 2$, the same as in Hawkes [4], Jaiswal [5, 6] and Sidi [14, 16]. We sill see that the approach taken here requires a solution to a two-dimensional functional equation in this case ($P = 2$) and that the joint generating function for the queue lengths distribution is obtained. The queue length, here, is defined as the number of customers in the system (including the server). For the sake of convenience, we introduce the vector notation

$$\underline{0} \equiv (0, 0);$$
$$\underline{k}_0^0 \equiv (k_1, k_2), \underline{k}_0^1 \equiv (0, k_2), \underline{k}_0^2 \equiv (k_1, 0);$$

and

$\underline{u}^1 \equiv (1, 0), \underline{u}^2 \equiv (0, 1).$

2.1 Preemptive resume priority rule

When a customer being served is interrupted by a higher priority class customer under the preemptive resume rule, the interrupted customer enters into "limbo," a term introduced by Wolff [20]. The supplementary variables are chosen as the elapsed service times of customers in service and limbo. To be exact, we define the following probabilities:

Q_1 $(\underline{k}, x, y, t)dxdy$ $\equiv$ the probability that at time t there are k_j class j customers in the system ($1 \leq j \leq 2$), the elapsed service time on the customer under service (of class 1) lies between x and $x + dx$, and the head of the class 2 customers was preempted earlier (a class 2 customer is in limbo) when its elapsed service time was lying between y and $y + dy$;

P_1 $(\underline{k}, x, t)dx$ $\equiv$ the probability that at time t there are k_j class j customers in the system ($1 \leq j \leq 2$), the elapsed service time on the customer under service (of class 1) lies between x and $x + dx$, and none of the class 2 customers was preempted earlier (no one is in limbo);

P_2 $(\underline{k}_0^1, x, t)dx$ $\equiv$ the probability that at time t there are no class 1 customers and k_2 class 2 customers in the system, the elapsed service time on the customer under service (of class 2) lies between x and $x + dx$;

P_0 (t) $\equiv$ the probability that the system is empty at time t;

and

Q_1 $(\underline{0}, x, y, t) \equiv Q_1(\underline{k}_0^1, x, y, t) \equiv Q_1(\underline{k}_0^2, c, y, t) \equiv 0, P_1(\underline{k}_0^1, x, t) \equiv P_2(\underline{k}_0^2, x, t) \equiv 0.$

To obtain the difference–differential equations for the state probabilities, we follow Keilson and Kooharian [8] relating the probabilities at time $t+dt$ to those at time t. These arguments lead to

$$\begin{aligned}
&Q_1(\underline{k}, x+dt, y, t+dt) = Q_1(\underline{k}, x, y, t)\{1-(\lambda+\eta_1(x))dt\}\\
&\qquad + \lambda \sum_{0\le \underline{n}\le \underline{k}} Q_1(\underline{n}, x, y, t)a(\underline{k}-\underline{n})dt + o(dt),\\
&P_1(\underline{k}, x+dt, t+dt) = P_1(\underline{k}, x, t)\{1-(\lambda+\eta_1(x))dt\}\\
&\qquad + \lambda \sum_{0\le \underline{n}\le \underline{k}} P_1(\underline{n}, x, t)a(\underline{k}-\underline{n})dt + o(dt),\\
&P_2(\underline{k}_0^1, x+dt, t+dt) = P_2(\underline{k}_0^1, x, t)\{1-(\lambda+\eta_2(x))dt\}\\
&\qquad + \lambda \sum_{0\le \underline{n}_0^1\le \underline{k}_0^1} P_2(\underline{n}_0^1, x, t)a(\underline{k}_0^1-\underline{n}_0^1)dt + \int_0^\infty Q_1(\underline{k}_0^1+\underline{u}^1, x_0, x, t)\eta_1(x_0)dx_0dt + o(dt).
\end{aligned}$$

Taking the limit of the above equations as $dt \to 0$ yields the basic equations

$$\frac{\partial Q_1(\underline{k}, x, y, t)}{\partial t} + \frac{\partial Q_1(\underline{k}, x, y, t)}{\partial x} + (\lambda+\eta_1(x))Q_1(\underline{k}, x, y, t) = \lambda \sum_{0\le \underline{n}\le \underline{k}} Q_1(\underline{n}, x, y, t)a(\underline{k}-\underline{n}), \tag{2.1}$$

$$\frac{\partial P_i(\underline{k}_0^{i-1}, x, t)}{\partial t} + \frac{\partial P_i(\underline{k}_0^{i-1}, x, t)}{\partial x} + (\lambda+\eta_i(x))P_i(\underline{k}_0^{i-1}, x, t)$$
$$= \lambda \sum_{0\le \underline{n}_0^{i-1}\le \underline{k}_0^{i-1}} P_i(\underline{n}_0^{i-1}, x, t)a(\underline{k}_0^{i-1}-\underline{n}_0^{i-1}) + \int_0^\infty Q_{i-1}(\underline{k}_0^{i-1}+\underline{u}^{i-1}, x_0, x, t)\eta_{i-1}(x_0)dx_0, \tag{2.2}$$
$$(1\le i\le 2)$$

where $Q_0(\underline{k}, x, y, t)$ is assumed to be zero. Similarly, for the empty state, we have

$$\frac{d}{dt}P_0(t) + \lambda P_0(t) = \sum_{i=1}^{2}\int_0^\infty P_i(\underline{u}^i, x, t)\eta_i(x)dx. \tag{2.3}$$

These equations are to be solved under the following boundary conditions:

$$Q_1(\underline{k}, 0, y, t) = \int_0^\infty Q_1(\underline{k}+\underline{u}^1, x, y, t)\eta_1(x)dx + \lambda \sum_{0\le \underline{n}_0^1\le \underline{k}_0^1} P_2(\underline{n}_0^1, y, t)a(\underline{k}-\underline{n}_0^1), \tag{2.4}$$

$$P_i(\underline{k}_0^{i-1}, 0, t) = \sum_{j=1}^{i}\int_0^\infty P_j(\underline{k}_0^{i-1}+\underline{u}^j, x, t)\eta_j(x)dx + \lambda P_0(t)a(\underline{k}_0^{i-1})\ (1\le i\le 2). \tag{2.5}$$

The initial condition is assumed to be $P_0(0)=1$, i.e., we start with an empty system.

We define the generating functions as

$$\begin{aligned}
f_m(x, y, z_2, t) &\equiv \sum_{n\ge 1} z_2^n Q_1((m,n), x, y, t),\\
F(x, y, \underline{z}, t) &\equiv \sum_{m\ge 1}\sum_{n\ge 1} z_1^m z_2^n Q_1((m,n)x, y, t) = \sum_{m\ge 1} z_1^m f_m(x, y, z_2, t),\\
h_m(x, z_2, t) &\equiv \sum_{n\ge 1} z_2^n P_1((m,n), x, t),\\
H_1(x, \underline{z}, t) &\equiv \sum_{m\ge 1}\sum_{n\ge 1} z_1^m z_2^n P_1((m,n), x, t) = \sum_{m\ge 1} z_1^m h_m(x, z_2, t),
\end{aligned}$$

and

$$H_2(x, z_2, t) \equiv \sum_{n\geq 1} z_2^n P_2((0,n), x, t).$$

Multiplying equations (2.1) and (2.2) with appropriate powers of z_1 and z_2, we have

$$\frac{\partial F}{\partial t} + \frac{\partial F}{\partial x} + \{\lambda(1 - G(\underline{z})) + \eta_1(x)\}F = 0, \tag{2.6}$$

$$\frac{\partial H_1}{\partial t} + \frac{\partial H_1}{\partial x} + \{\lambda(1 - G(\underline{z})) + \eta_1(x)\}H_1 = 0, \tag{2.7}$$

$$\frac{\partial H_2}{\partial t} + \frac{\partial H_2}{\partial x} + \{\lambda(1 - G(0, z_2)) + \eta_2(x)\}H_2 = \int_0^\infty f_1(x_0, x, z_2, t)\eta_1(x_0)dx_0. \tag{2.8}$$

Similarly the boundary conditions (2.4) and (2.5) become

$$\begin{aligned} F(0, y, \underline{z}, t) &= \frac{1}{z_1}\int_0^\infty F(x, y, \underline{z}, t)\eta_1(x)dx - \int_0^\infty f_1(x, y, z_2, t)\eta_1(x)dx \\ &+ \lambda[G(\underline{z}) - G(0, z_2)]H_2(y, z_2, t), \\ H_1(0, \underline{z}, t) &= \frac{1}{z_1}\int_0^\infty H_1(x, \underline{z}, t)\eta_1(x)dx - \int_0^\infty h_1(x, z_2, t)\eta_1(x)dx \\ &+ \lambda[G(\underline{z}) - G(0, z_2)]P_0(t), \end{aligned}$$

and with the help of (2.3),

$$H_2(0, z_2, t) = \frac{1}{z_2}\int_0^\infty H_2(x, z_2, t)\eta_2(x)dx + \int_0^\infty h_1(x, z_2, t)\eta_1(x)dx$$

$$-\lambda[1 - G(0, z_2)]P_0(t) - \frac{d}{dt}P_0(t).$$

We now take Laplace transforms with respect to t and denote them by stars (*), so that, for example,

$$F^*(x, y, \underline{z}, s) \equiv \int_0^\infty e^{-st}F(x, y, \underline{z}, t)dt \ (Re(s) > 0).$$

Equations (2.6) through (2.8) then become

$$\frac{\partial F^*}{\partial x} + \{\lambda(1 - G(\underline{z})) + s + \eta_1(x)\}F^* = 0, \tag{2.9}$$

$$\frac{\partial H_1^*}{\partial x} + \{\lambda(1 - G(\underline{z})) + s + \eta_1(x)\}H_1^* = 0, \tag{2.10}$$

$$\frac{\partial H_2^*}{\partial x} + \{\lambda(1 - G(0, x_2)) + s + \eta_2(x)\}H_2^* = \int_0^\infty f_1^*(x_0, x, z_2, s)\eta_1(x_0)dx_0. \tag{2.11}$$

The boundary conditions become

$$\begin{aligned} F^*(0, y, \underline{z}, s) &= \frac{1}{z_1}\int_0^\infty F^*(x, \underline{z}, s)\eta_1(x)dx - \int_0^\infty f_1^*(x_0, x, z_2, s)\eta_1(x_0)dx_0 \\ &+ \lambda[G(\underline{z}) - G(0, z_2)]H_2^*(y, z_2, s), \end{aligned} \tag{2.12}$$

$$\begin{aligned} H_1^*(0, \underline{z}, s) &= \frac{1}{z_1}\int_0^\infty H_1^*(x, \underline{z}, s)\eta_1(x)dx - \int_0^\infty h_1^*(x, z_2, s)\eta_1(x)dx \\ &+ \lambda[G(\underline{z}) - G(0, z_2)]P_0^*(s), \end{aligned} \tag{2.13}$$

and making use of the initial condition $P_0(0) = 1$,

$$H_2^*(0, z_2, s) = \frac{1}{z_2}\int_0^\infty H_2^*(x, z_2, s)\eta_2(x)dx + \int_0^\infty h_1^*(x, z_2, s)\eta_1(x)dx$$
$$-[\lambda(1 - G(0, z_2)) + s]P_0^*(s) + 1. \tag{2.14}$$

The solutions of (2.9) and (2.10) are given by

$$F^*(x, y, \underline{z}, s) = F^*(0, y, \underline{z}, s)\exp\{-\int_0^x \eta_1(y)dy - [\lambda(1 - G(\underline{z})) + s]x\}, \tag{2.15}$$

$$H_1^*(x, \underline{z}, s) = H_1^*(0, \underline{z}, s)\exp\{-\int_0^z \eta_1(y)dy - [\lambda(1 - G(\underline{z})) + s]x\}. \tag{2.16}$$

It remains to determine $F^*(0, y, \underline{z}, s), H_1^*(0, \underline{z}, s)$, and $H_2^*(x, z_2, s)$. Substituting (2.15) and (2.16) into (2.12) and (2.13) gives that

$$F^*(0, y, \underline{z}, s)\{1 - B_1^*[\lambda(1 - G(\underline{z})) + s]/z_1\}$$
$$= \lambda[G(\underline{z}) - G(0, z_2)]H_2^*(y, z_2, s) - \int_0^\infty f_1^*(x, y, z_2, s)\eta_1(x)dx, \tag{2.17}$$
$$H_1^*(0, \underline{z}, s)\{1 - B_1^*[\lambda(1 - G(\underline{z})) + s]/z_1\}$$
$$= \lambda[G(\underline{z}) - G(0, z_2)]P_0^*(s) - \int_0^\infty h_1^*(x, z_2, s)\eta_1(x)dx. \tag{2.18}$$

By applying Rouché's theorem (see Appendix), it is easy to show that there exists the unique root $\tilde{z}_1 \equiv \tilde{z}_1(z_2, s)(|\tilde{z}_1| < 1)$ of the equation:

$$\tilde{z}_1 - B_1^*[\lambda(1 - G(\tilde{z}_1, z_2)) + s] = 0$$

for $Re(s) > 0$ and $|z_2| \leq 1$. It follows now from (2.17) and (2.18) that

$$\int_0^\infty f_1^*(x, y, z_2, s)\eta_1(x)dx = \lambda[G(\tilde{z}_1, z_2) - G(0, z_2)]H_2^*(y, z_2, s), \tag{2.19}$$

$$\int_0^\infty h_1^*(x, z_2, s)\eta_1(x)dx = \lambda[G(\tilde{z}_1, z_2) - G(0, z_2)]P_0^*(s). \tag{2.20}$$

Substituting (2.19) into (2.11) gives that

$$\frac{\partial H_2^*}{\partial x} + \{\lambda(1 - G(\tilde{z}_1, z_2)) + s + \eta_2(x)\}H_2^* = 0,$$

from which it follows that

$$H_2^*(x, z_2, s) = H_2^*(0, z_2, s)\exp\{-\int_0^x \eta_2(y)dy - [\lambda(1 - G(\tilde{z}_1, z_2)) + s]x\}. \tag{2.21}$$

Substituting (2.21) into (2.14), we get from (2.20) that

$$H_2^*(0, z_2, s) = \frac{1 - [\lambda\{1 - G(\tilde{z}_1, z_2)\} + s]P_0^*(s)}{1 - \frac{1}{z_2}B_2^*[\lambda(1 - G(\tilde{z}_1, z_2)) + s]}. \tag{2.22}$$

It follows from (2.17) through (2.20) that

$$F^*(0, y, \underline{z}, s) = \frac{\lambda\{G(\underline{z}) - G(\tilde{z}_1, z_2)\}}{1 - \frac{1}{z_1}B_1^*[\lambda(1 - G(\underline{z})) + s]}H_2^*(y, z_2, s), \tag{2.23}$$

$$H_1^*(0,\underline{z},s) = \frac{\lambda\{G(\underline{z}) - G(\tilde{z}_1, z_2)\}}{1 - \frac{1}{z_1}B_1^*[\lambda(1-G(\underline{z})) + s]} P_0^*(s). \tag{2.24}$$

Thus $F^*(x,y,\underline{z},s), H_1^*(x,\underline{z},s)$, and $H_2^*(x,z_2,s)$ are determined by (2.15), (2.16), and (2.21) through (2.24) except for $P_0^*(s)$. To determine $P_0^*(s)$, we observe that $H_2^*(0,z_2,s)$ is regular for $|z_2| < 1$ and therefore

$$P_0^*(s) = \frac{1}{\lambda\{1 - G(\tilde{z}_1(\tilde{z}_2,s),\tilde{z}_2)\} + s} \tag{2.25}$$

where $\tilde{z}_2 \equiv \tilde{z}_2(s)$ is the unique root of the equation:

$$\tilde{z}_2 - B_2^*[\lambda(1 - G(\tilde{z}_1(\tilde{z}_2,s),\tilde{z}_2)) + s] = 0.$$

The existence and uniqueness of $\tilde{z}_2$ follow from the similar argument to that employed for $\tilde{z}_1$.

Now if $\Pi(\underline{z},t)$ is the joint probability generating function for the queue lengths at time t, its Laplace transform is given by

$$\Pi^*(\underline{z},s) = \int_0^\infty\int_0^\infty F^*(x,y,\underline{z},s)dxdy + \int_0^\infty H_1^*(x,\underline{z},s)dx + \int_0^\infty H_2^*(s,z_2,s)dx + P_0^*(s).$$

The equilibrium distribution of queue lengths is

$$\Pi(\underline{z}) \equiv \lim_{t\to\infty}\Pi(\underline{z},t) = \lim_{s\to 0} s\Pi^*(\underline{z},s),$$

provided that the first limit exists. Similarly, the limit of the Laplace transform multiplied by s as $s \to 0$ will be denoted by omitting the star (*), for example,

$$P_0 \equiv \lim_{s\to 0} sP_0^*(s).$$

Assuming that the equilibrium distribution does exist, we have

$$\Pi(\underline{z}) = F(\underline{z}) + H_1(\underline{z}) + H_2(z_2) + P_0, \tag{2.26}$$

where

$$\begin{aligned} F(\underline{z}) &\equiv \lim_{s\to 0} s\int_0^\infty\int_0^\infty F^*(x,y,\underline{z},s)dxdy \\ &= \frac{G(\underline{z}) - G(\tilde{z}_1,z_2)}{1-G(\tilde{z})}\cdot\frac{1-B_1^*[\lambda(1-G(\underline{z}))]}{1-\frac{1}{z_1}B_1^*[\lambda(1-G(\underline{z}))]}\cdot H_2(z_2), \\ H_1(\underline{z}) &\equiv \lim_{s-0} s\int_0^\infty H_1^*(x,\underline{z},s)dx \\ &= \frac{G(\underline{z}) - G(\tilde{z}_1,z_2)}{1-G(\underline{z})}\cdot\frac{1-B_1^*[\lambda(1-G(\underline{z}))]}{1-\frac{1}{z_1}B_1^*[\lambda(1-G(\underline{z}))]}\cdot P_0, \end{aligned}$$

and

$$\begin{aligned} H_2(z_2) &\equiv \lim_{s\to 0} s\int_0^\infty H_2^*(x,z_2,s)dx \\ &= (-1)\cdot\frac{1-B_2^*[\lambda(1-G(\tilde{z}_1,z_2))]}{1-\frac{1}{z_2}B_2^*[\lambda(1-G(\tilde{z}_1,z_2))]}\cdot P_0. \end{aligned}$$

Here, $\tilde{z}_1 \equiv \tilde{z}_1(z_2)$ is the root of the equation:

$$\tilde{z}_1 - B_1^*[\lambda(1-G(\tilde{z}_1,z_2))] = 0. \tag{2.27}$$

The normalizing condition $\Pi(1,1) = 1$ implies that

$$P_0 = 1 - \rho_1 - \rho_2.$$

The moments of the queue lengths can be determined from (2.26). The mean number of class 1 customers in the system, $E[L_1]$, is

$$E[L_1] = \frac{\partial}{\partial z_1}\Pi(\underline{z}) \mid_{\underline{z}=\underline{1}} = \rho_1 + \frac{\lambda^2 g_1^2 b_1^{(2)} + \rho_1 g_1^{(2)}/g_1}{2(1-\rho_1)}. \tag{2.28}$$

The mean number of class 2 customers in the system, $E[L_2]$, is given by

$$\begin{aligned} E[L_2] &= \frac{\partial}{\partial z_2}\Pi(\underline{z}) \mid_{\underline{z}=\underline{1}} \\ &= \frac{\rho_2}{1-\rho_1} + \frac{\lambda g_2\{\lambda g_2 b_2^{(2)} + \lambda g_1^{(2)} b_1^2 + \lambda g_1 b_1^{(2)}\}}{2(1-\rho_1)(1-\rho_1-\rho_2)} \\ &\quad + \frac{\rho_2 g_2^{(2)}}{2(1-\rho_1-\rho_2)g_2} + \frac{\lambda g_{12} b_1}{1-\rho_1-\rho_2}. \end{aligned} \tag{2.29}$$

Remark 2.1.

We consider a preemptive resume priority queue with an arriving batch being composed of a single class, and a single customer only. This model has been treated by Jaiswal [5]. We denote by λ_p the arrival rate for class p ($p = 1,2$). We have for a single independent arrival model $G(\underline{z}) = (\lambda_1 z_1 + \lambda_2 z_2)/\lambda$, from which it follows that $g_{12} = [\frac{\partial^2 G(\underline{z})}{\partial z_1 \partial z_2}]_{\underline{z}=\underline{1}} = 0$. It is seen that equations (2.28) and (2.29) are consistent with Jaiswal's formula [5, (34), (35)]. Meister [12] analyzed an independent batch arrival preemptive resume model via the imbedded Markov approach, but evaluated only the waiting time of the first customer in a batch (a supercustomer). It is seen that Meister's formula follows from the above equations (2.28) and (2.29) by using a result in Section 4. □

2.2 Nonpreemptive (Head–of–the–line) priority rule

As remarked in Jaiswal [5], the nonpreemptive (head–of–the–line) priority is comparatively easier to solve because of the absence of the preempted time variable required to make the process Markovian. Under the nonpreemptive priority rule, we do not have to consider a customer in limbo [20]. It is enough to define the following probabilities:

P_i $(\underline{k}, x, t)dx \equiv$ the probability that at time t there are k_j class j customers in the system ($1 \le j \le 2$), the elapsed service time on the customer under service (of class i) lies between x and $x + dx$;

P_0 $(t) \equiv$ the probability that the system is empty at time t;

and

P_i $(\underline{0}, x, t) \equiv P_1(\underline{k}_0^1, x, t) \equiv P_2(\underline{k}_0^2, x, t) \equiv 0$.

An argument to derive the basic equations (2.1) through (2.3) under the preemptive resume rule is now applied similarly under the nonpreemptive priority rule. We have

$$\begin{aligned} &\frac{\partial P_i(\underline{k}, x, t)}{\partial t} + \frac{\partial P_i(\underline{k}, x, t)}{\partial x} + (\lambda + \eta_i(x))P_i(\underline{k}, x, t) \\ &= \lambda \sum_{0 \le \underline{n} \le \underline{k}} P_i(\underline{n}, x, t)a(\underline{k} - \underline{n})(1 \le i \le 2), \end{aligned} \tag{2.30}$$

$$\frac{d}{dt}P_0(t) + \lambda P_0(t) = \sum_{i=1}^{2}\int_0^\infty P_i(\underline{u}^i, x, t)\eta_i(x)dx.$$

These equations are to be solved under the following boundary conditions:

$$P_i(\underline{k}_0^{i-1}, 0, t) = \sum_{j=1}^{2}\int_0^\infty P_j(\underline{k}_0^{i-1} + \underline{u}^j, x, t)\eta_j(x)dx + \lambda P_0(t)a(\underline{k}_0^{i-1}) \ (1 \le i \le 2). \tag{2.31}$$

The initial condition is assumed to be $P_0(0) = 1$. Again, we start with an empty system.

We define the generating functions as

$$f_i^m(x, z_2, t) \equiv \sum_{n\ge 1} z_2^n P_i((m,n), x, t),$$
$$F_i(x, \underline{z}, t) \equiv \sum_{m\ge 1}\sum_{n\ge 1} z_1^m z_2^n P_i((m,n), x, t) = \sum_{m\ge 1} z_1^m f_i^m(x, z_2, t).$$

Multiplying equation (2.30) with appropriate powers of z_1 and z_2, we have

$$\frac{\partial F_i}{\partial t} + \frac{\partial F_i}{\partial x} + \{\lambda(1 - G(\underline{z})) + \eta_i(x)\}F_i = 0 \ (1 \le i \le 2),$$

Similarly the boundary conditions (2.31) become

$$\begin{aligned} F_1(0, \underline{z}, t) &= \frac{1}{z_1}\int_0^\infty [F_1(x, \underline{z}, t) - f_1^1(x, z_2, t)]\eta_1(x)dx \\ &+ \frac{1}{z_2}\int_0^\infty [F_2(x, \underline{z}, t) - F_2(x, \underline{z}_0^1, t)]\eta_2(x)dx + \lambda P_0(t)[G(\underline{z}) - G(\underline{z}_0^1)], \\ F_2(0, \underline{z}_0^1, t)&[\frac{1}{z_2}\int_0^\infty F_2(x, \underline{z}_0^1, t)\eta_2(x)dx + \int_0^\infty f_1^1(x, z_2, t)\eta_1(x)dx \\ &- \sum_{i=1}^{2}\int_0^\infty P_i(\underline{u}^i, x, t)\eta_i(x)dx + \lambda P_0(t)G(\underline{z}_0^1). \end{aligned}$$

Taking Laplace transforms, we have

$$\begin{aligned} &\frac{\partial F_i^*}{\partial x} + \{(1 - G(\underline{z})) + \eta_i(x) + s\}F_i^* = 0(1 \le i \le 2), \\ &(s+\lambda)P_0^*(s) = \sum_{i=1}^{2}\int_0^\infty P_i(\underline{u}^i, x, s)\eta_i(x)dx + 1, \\ &F_1^*(0, \underline{z}, s) = \frac{1}{z_1}\int_0^\infty [F_1^*(x, \underline{z}, s) - z_1 f_1^{1*}(x, z_2, s)]\eta_1(x)dx \\ &\quad + \frac{1}{z_2}\int_0^\infty [F_2^*(x, \underline{z}, s) - F_2^*(x, \underline{z}_0^1, s)]\eta_2(x)dx + \lambda P_0^*(s)[G(\underline{z}) - G(\underline{z}_0^1)], \end{aligned} \tag{2.32}$$

and

$$\begin{aligned} F_2^*(0, \underline{z}_0^1, s) = \frac{1}{z_2}\int_0^\infty F_2^*(x, \underline{z}_0^1, s)\eta_2(x)dx + \int_0^\infty f_1^{1*}(x, z_2, s)\eta_1(x)dx \\ +1 - \{s + \lambda - \lambda G(\underline{z}_0^1)\}P_0^*(s). \end{aligned}$$

The solution of (2.32) is given by

$$F_i^*(x, \underline{z}, s) = F_i^*(0, \underline{z}, s)\exp\{-\int_0^x \eta_i(y)dy - [\lambda(1 - G(\underline{z})) + s)]x\} \ (1 \le i \le 2). \tag{2.33}$$

Recall that $P_2(\underline{k}, 0, t) = 0$ for all $k_1 > 0$ from the definition of the nonpreemptive priority rule. It then follows that $F_i^*(0, \underline{z}, s) = F_i^*(0, \underline{z}_0^1, s)$. Almost the same derivation as under the preemptive resume rule leads to

$$F_2^*(0, \underline{z}_0^1, s) = \frac{1 - [\lambda\{1 - G(\tilde{z}_1, z_2)\} + s]P_0^*(s)}{1 - \frac{1}{z_2}B_2^*[\lambda(1 - G(\tilde{z}_1, z_2)) + s]}, \tag{2.34}$$

$$\begin{aligned} F_1^*(0, \underline{z}, s) &= \frac{1 - [\lambda\{1 - G(\underline{z})\} + s]P_0^*(s)}{1 - \frac{1}{z_1}B_1^*[\lambda(1 - G(\underline{z})) + s]} \\ &\quad - F_2^*(0, \underline{z}_0^1, s)\frac{1 - \frac{1}{z_2}B_2^*[\lambda(1 - G(\underline{z})) + s]}{1 - \frac{1}{z_1}B_1^*[\lambda(1 - G(\underline{z})) + s]}, \end{aligned} \tag{2.35}$$

where $\tilde{z}_1 \equiv \tilde{z}_1(z_2, s)$ is the unique root of the equation:

$$\tilde{z}_1 - B_1^*[\lambda(1 - G(\tilde{z}_1, z_2)) + s] = 0.$$

(See also Appendix.) Note that $P_0^*(s)$ is given by equation (2.25), since $P_0^*(s)$ is invariant under the work-conserving rule [20], and since the preemptive resume and nonpreemptive priority rules are work-conserving.

The equilibrium distribution of queue lengths is given by

$$\Pi(\underline{z}) \equiv \lim_{t\to\infty} \Pi(\underline{z}, t) = \lim_{s\to 0} s\Pi^*(\underline{z}, s) = \lim_{s\to 0} s\{\sum_{i=1}^{2} \int_0^\infty F_i(x, \underline{z}, s)dx + P_0^*(s)\},$$

provided that the first limit exists. Assuming that the equilibrium distribution does exist, we have

$$\Pi(\underline{z}) = \sum_{i=1}^{2} F_i(\underline{z}_0^{i-1}) + P_0,$$

where

$$\begin{aligned} F_1(\underline{z}) &\equiv \lim_{s\to 0} s\int_0^\infty F_1^*(x, \underline{z}, s)dx \\ &= (-P_0)\cdot\frac{1 - B_1^*[\lambda(1 - G(\underline{z}))]}{1 - \frac{1}{z_1}B_1^*[\lambda(1 - G(\underline{z}))]} \\ &\quad + P_0\cdot\frac{1 - G(\tilde{z}_1, z_2)}{1 - G(\tilde{z})}\cdot\frac{1 - B_1^*[\lambda(1 - G(\underline{z}))]}{1 - \frac{1}{z_1}B_1^*[\lambda(1 - G(\underline{z}))]}\cdot\frac{1 - \frac{1}{z_2}B_2^*[\lambda(1 - G(\underline{z}))]}{1 - \frac{1}{z_2}B_2^*[\lambda(1 - G(\tilde{z}_1, z_2))]}, \\ F_2(\underline{z}_0^1) &\equiv \lim_{s\to 0} s\int_0^\infty F_2^*(x, \underline{z}_0^1, s)dx \\ &= (-P_0)\cdot\frac{1 - G(\tilde{z}_1, z_2)}{1 - G(\underline{z})}\cdot\frac{1 - B_2^*[\lambda(1 - G(\underline{z}))]}{1 - \frac{1}{z_2}B_2^*[\lambda(1 - G(\tilde{z}_1, z_2))]}. \end{aligned}$$

Here, $P_0 = 1 - \rho_1 - \rho_2$, and $\tilde{z}_1 \equiv \tilde{z}_1(z_2)$ is the root of the equation (2.27). The moments of the queue lengths can be determined from the above moment generating function $\Pi(\underline{z})$. The mean number of class 1 customers in the system, $E[L_1]$, is

$$E[L_1] = \frac{\partial}{\partial z_1}\Pi(\underline{z})\mid_{\underline{z}=\underline{1}} = \rho_1 + \frac{\lambda g_1\{\lambda g_1 b_1^{(2)} + \lambda g_2 b_2^{(2)}\} + \rho_1 g_1^{(2)}/g_1}{2(1 - \rho_1)}. \tag{2.36}$$

The mean number of class 2 customers in the system, $E[L_2]$, is given by

$$E[L_2] = \frac{\partial}{\partial z_2}\Pi(\underline{z}) \mid_{\underline{z}=\underline{1}} = \rho_2 + \frac{\lambda g_2\{\lambda g_2 b_2^{(2)} + \lambda g_1^{(2)} b_1^2 + \lambda g_1 b_1^{(2)}\}}{2(1-\rho_1)(1-\rho_1-\rho_2)} + \frac{\rho_2 g_2^{(2)}}{2(1-\rho_1-\rho_2)g_2} + \frac{\lambda g_{12} b_1}{1-\rho_1-\rho_2}. \quad (2.37)$$

Remark 2.2.

Consider a nonpreemptive priority queue with an arriving batch being composed of a single class only, which is treated by Hawkes [4]. As in [4], denote by λ_p and $K_p(z_p)$ the arrival rate of batches and the generating function of batch size for class p $(p = 1,2)$. We then have

$$G(\underline{z}) = \frac{\lambda_1}{\lambda}K_1(z_1) + \frac{\lambda_2}{\lambda}K_2(z_2),$$

from which it follows that $g_{12} = [\frac{\partial^2 G(\underline{z})}{\partial z_1 \partial z_2}]_{\underline{z}=\underline{1}} = 0$. Equations (2.36) and (2.37) are seen to be consistent with Hawkes' formula [4, (25), (26)]. □

3. Waiting Time Distribution

3.1 Preliminaries

We consider an arbitrary class p customer who will be called as a class p "tagged" customer $(1 \leq p \leq P)$. The batch to which the class p tagged customer belongs will be called as the (class p) "associated" batch. It is verified that the class p associated batch sees time averages $(1 \leq p \leq P)$. This is because if Poisson arrivals with rate λ are independently selected with probability $1 - G(\underline{1}_{0p})$, the selected arrivals also form a Poisson process with rate $\lambda(1 - G(\underline{1}_{0p}))$. Here, $\underline{1}_{0p}$ signifies the vector where the i-th component $(i \neq p)$ is equal to unity but the p-th component is equal to zero, i.e., $\underline{1}_{0p} \equiv (1, \cdots, 1, 0, 1, \cdots, 1)$.

Let W_p be the waiting time of the class p tagged customer. All the customers already in the system at the arrival of the tagged customer who are to be served before his service are refereed to as "senior" customers. The senior customers consist of two types:
under PR;

1) the customer already in service whose priority index is less than or equal to p when the tagged customer arrives at the system,
2) the customers already in the queue and in limbo of classes $\{1, 2, \cdots, p\}$ when he arrives; and under NP;
1) the customer already in service when the tagged customer arrives at the system,
2) the customers already in the queue of classes $\{1, 2, \cdots, p\}$ when he arrives,

where the abbreviations PR and NP stand for preemptive resume and nonpreemptive priorities. Let τ_p be the time required to serve all the senior customers.

The sequence of customers served during W_p is generally complicated. For example, the higher class $\{1, 2, \cdots, p-1\}$ customers who arrive during W_p might get ahead of the senior customers. To simplify the argument, we adopt the following resequence as in Fujiki and Gambe [3, Chapter 12.2]. The senior customers are firstly served, all the higher class $\{1, 2, \cdots, p-1\}$ customers who subsequently arrive during the delay cycle generated by the higher class customers with an initial delay of τ_p are secondly served, all customers in the associated batch who have precedence over the tagged customer are thirdly served, and all the subsequently arriving higher class customers who precede the entrance into service of the tagged customer are finally served. Note that W_p is invariant after this resequence, since the resequence covers all the services to be done during W_p (see [3, p. 379]). In Sections 3 and 4, we will employ the resequence tacitly.

A key observation is the fact that the waiting time W_p is the summation of two independent random variables $W_{g,p}$ and T_p, where

$W_{g,p}$: the delay cycle generated by the higher class $\{1, 2, \cdots, p-1\}$ customers with an initial delay of τ_p,

T_p : the delay cycle generated by the higher class $\{1, 2, \cdots, p-1\}$ customers with an initial delay of total required service times to the customers in the associated batch having precedence over the tagged customer.

From the definition of our resequence, $W_{g,p}$ is the waiting time of the first customer (regardless of class) in the class p associated batch, and T_p is the time from start of service to the first customer in the associated batch to start of service to the tagged customer.

Henceforth, we will denote the LST of the distribution of a random variable by star (*), so that for example,

$$W_p^*(s) \equiv \int_0^\infty e^{-st} dP(W_p \le t) \quad (Re(s) > 0).$$

It follows from the definitions that

$$E[W_p] = E[W_{g,p}] + E[T_p] \ (1 \le p \le P) \quad \text{under } PR \text{ and } NP, \tag{3.1}$$

and

$$W_p^*(s) = W_{g,p}^*(s) T_p^*(s) \ (1 \le p \le P) \quad \text{under } PR \text{ and } NP. \tag{3.2}$$

For simplicity, we restrict ourselves to the case of $P = 2$ as in Section 2 to find the LSTs $W_{g,p}^*(s)$ and $T_p^*(s)(1 \le p \le 2)$.

3.2 Waiting time for class 1 in the two-class priority queue

Let $r_p(i)$ be the probability that an arbitrary class p customer is served i-th within the same class in its batch $(1 \le p \le 2)$. It is verified that

$$r_1(i) = \sum_{k_1=i}^{\infty} \sum_{k_2=0}^{\infty} a(k_1, k_2)/g_1; r_2(i) = \sum_{k_1=0}^{\infty} \sum_{k_2=i}^{\infty} a(k_1, k_2)/g_2 = \sum_{k_1=0}^{\infty} r_2(k_1, i), \tag{3.3}$$

where $r_2(k_1, i) \equiv \sum_{k_2=i}^{\infty} a(k_1, k_2)/g_2$ denotes the probability that a class 2 customer is served i-th in its batch having k_1 class 1 customers. (See (4.1) in the subsequent section.) It then follows from (3.3) that

$$T_1^*(s) = \sum_{i=1}^{\infty} r_1(i)[B_1^*(s)]^{i-1} = \frac{1 - G(B_1^*(s), 1)}{[1 - B_1^*(s)]g_1} \quad \text{under } PR \text{ and } NP. \tag{3.4}$$

The remaining work for us is to find $W_{g,1}^*(s)$. For the moment, we assume a modified model in which a batch of class 1 is considered as a single customer (called a supercustomer) with arrival rate of $\lambda' \equiv (1 - G(0, 1))$ and the LST of service time distribution given by $B_1^{'*}(s) \equiv [G(B_1^*(s), 1) - G(0, 1)]/[1 - G(0, 1)]$. For this modified model we use dashes (′) on each of the corresponding functions. Note that $W_1^{'*}(s) = W_{g,1}^*(s)$ under PR and NP. The derivation of $W_{g,1}^*(s)$ under PR is straightforward. In fact, it follows from the well-known Pollaczek-Khintchine formula that

$$\begin{aligned} W_{g,1}^*(s) &= W_1^{'*}(s) \\ &= \frac{(1 - \lambda' E[B_1'])s}{s - \lambda'[1 - B_1^{'*}(s)]} \\ &= \frac{(1 - \rho_1)s}{s - \lambda[1 - G(B_1^*(s), 1)]} \quad \text{under } PR. \end{aligned} \tag{3.5}$$

It becomes complicated to derive $W^*_{g,1}(s)$ under NP, since we cannot neglect class 2 customers in service. We use $V-cycle$ [9], which is defined by the delay cycle generated by class 1 supercustomers with initial delay V. The LST of the waiting time distribution of a supercustomer of class 1 that arrives during V-cycle is given by

$$\begin{aligned} W^*_{g,1}(s \mid V-\text{cycle}) &= \frac{1-V^*(s)}{sE[V]}\cdot\frac{(1-\lambda' E[B'_1])s}{s-\lambda'[1-B'^*_1(s)]} \\ &= \frac{1-V^*(s)}{sE[V]}\cdot\frac{s(1-\rho_1)}{s-\lambda+\lambda G(B^*_1(s),1)}. \end{aligned} \tag{3.6}$$

(see Kella and Yechiali [9]). Let B'_1-cycle and B_2-cycle be the V-cycle with $V^*(s)=B'^*_1(s)$, and the V-cycle with $V^*(s)=B^*_2(s)$, respectively. Denote by P_k the probability that a supercustomer of class 1 arrives when the system is in B'_1-cycle for $k=1$, or in B_2-cycle for $k=2$. Since a supercustomer of class 1 arrives during B'_1-cycle or during B_2-cycle unless he finds the system empty, we have

$$P_1+P_2=\rho. \tag{3.7}$$

Note that the number of times the system enters B_2-cycle per unit time is λg_2 on the average, and that each cycle lasts $b_2/(1-\rho)$ on the average. We then have

$$P_2=\lambda g_2\frac{b_2}{(1-\rho)}=\frac{\rho_2}{1-\rho}, \tag{3.8}$$

and from (3.7),

$$P_1=\frac{\rho_1(1-\rho)}{1-\rho_1}. \tag{3.9}$$

Thus we finally have from (3.6) through (3.9) that

$$\begin{aligned} W^*_{g,1}(s) &= 1-\rho+P_1W^*_{g,1}(s \mid B'_1-\text{cycle})+P_2W^*_{g,1}(s \mid B_2-\text{cycle}) \\ &= \frac{(1-\rho)s+\lambda g_2(1-B^*_2(s))}{s-\lambda+\lambda G(B^*_1(s),1)} \quad \text{under } NP. \end{aligned} \tag{3.10}$$

3.3 Waiting time for class 2 in the two-class priority queue

If the tagged customer is served i-th within class 2, and the associated batch has k_1 class 1 customers, the LST of the conditional delay time induced by the customers in the associated batch, $T^*_{2|k_1,i}(s)$, is obtained as

$$T^*_{2|k_1,i}(s)=[\Theta^*_1(s)]^{k_1}\cdot[C^*_2(s)]^{i-1}.$$

We have from (3.3) that

$$\begin{aligned} T^*_2(s) &= \sum_{k_1=0}^{\infty}\sum_{i=1}^{\infty} r_2(k_1,i)T^*_{2|k_1,i}(s) \\ &= \frac{G(\Theta^*_1(s),1)-G(\Theta^*_1(s),C^*_2(s))}{[1-C^*_2(s)]g_2} \quad \text{under } PR \text{ and } NP. \end{aligned} \tag{3.11}$$

To find $W^*_{g,2}(s)$, we use the completion time for class 2, C_2. The completion time C_2 is the time that elapses before next class 2 customer can receive service, see [3, 7]. Note that the conditional LST of the distribution of the time taken to clear the resulting class

1 customers is $[\Theta_1^*(s)]^j$ where $\Theta_1^*(s)$ is the LST of the busy period distribution of class 1 customers, given that an arriving batch includes j class 1 customers. We then have

$$\begin{aligned} C_2^*(s) &= \sum_{n=0}^{\infty} e^{-sx} e^{-\lambda x} \frac{(\lambda x)^n}{n!} [\sum_{k_1=0}^{\infty} \sum_{k_2=0}^{\infty} a(k_1 k_2) \Theta_1^*(s)^{k_1}]^n dB_2(x) \\ &= B_2^*[\lambda(1 - G(\Theta_1^*(s), 1)) + s] \end{aligned}$$

under PR and NP rules. $\Theta_1^*(s)$ itself follows by a similar argument replacing $B_2^*(s)$ by $B_1^*(s)$ in the above equation. Namely, $\Theta_1^*(s)$ satisfies the following equation

$$\Theta_1^*(s) = B_1^*[\lambda(1 - G(\Theta_1^*(s), 1)) + s].$$

We next consider the total unfinished work of classes 1 and 2, U_T. Note that U_T is identical to the waiting time in the $M/G/1$ queue where customers arrival rate is λ and the LST of the service time distribution is given by $G(B_1^*(s), B_2^*(s))$. Therefore, the LST of the total unfinished work $U_T^*(s)$ is given by

$$U_T^*(s) = \frac{(1 - \rho_1 - \rho_2)s}{s - \lambda[1 - G(B_1^*(s), B_2^*(s))]}, \tag{3.12}$$

from the Pollaczek-Khintchine formula. Following the derivation of $C_2^*(s)$, we have that

$$\begin{aligned} W_{g,2}^*(s) &= U_T^*[\lambda(1 - G(\Theta_1^*(s), 1)) + s] \\ &= \frac{(1 - \rho_1 - \rho_2)[s + \lambda(1 - G(\Theta_1^*(s), 1))]}{s - \lambda G(\Theta_1^*(s), 1) + \lambda G(B_1^*\{s + \lambda - \lambda G(\Theta_1^*(s), 1)\}, B_2^*\{s + \lambda - \lambda G(\Theta_1^*(s), 1)\})} \end{aligned} \tag{3.13}$$

under PR and NP.

We can find the mean waiting times for individual classes from the LSTs (3.2), (3.4), (3.5), (3.10), (3.11) and (3.13) as in Section 2. In fact it can be seen that for $1 \le p \le 2$,

$$E[W_p] = E[L_p]/(\lambda g_p) - b_p/(1 - \rho_{p-1}^+) \quad \text{under } PR, \tag{3.15}$$

and

$$E[W_p] = E[L_p]/(\lambda g_p) - b_p \quad \text{under } NP. \tag{3.16}$$

See also Remark 4.1. As far as the mean waiting time only is concerned, a different approach from that in this section will be developed in Section 4, which enables us to treat general P class priority queue straightforwardly.

4. Mean Delay Formula

4.1 Mean waiting time

To find the mean waiting time for a customer of an individual class, we will apply the approach used by Schrage [13] and Wolff [20] deriving a conservation law. We will see that this approach restricts us to dealing with only the mean performance measures of the system, but it is much simpler and more intuitive than the preceding approach.

Recall that Burke [1] treated a single class (nonpriority) batch arrival ($M^X/G/1$) queue and evaluated the probability that an arbitrary customer is served i-th in its batch. Applying the argument in [1] to the multi-class batch arrival case, the probability that an arbitrary class p customer is served i-th (within class p) in its batch, $r_p(i)$, is given by

$$r_p(i) = \sum_{k_j \geq 0 (j \neq p)} \sum_{l=i}^{\infty} a(k_1, \cdots, k_{p-1}, l, k_{p+1}, \cdots, k_P)/g_p \ (i = 1, 2, \cdots). \tag{4.1}$$

The customer of class p who is served i-th within its class in its batch will be refereed to as "the i-th customer" of class p. It should be noted that under both the PR and NP rules, the i-th customer of class p in its batch must be deferred for the delay cycle generated by customers of classes $\{1, 2, \cdots, p-1\}$, with a mean initial delay of

$$(i-1)b_p + \sum_{j=1}^{p-1} k_j b_j,$$

if the batch size of class j is $k_j (1 \leq j \leq p-1)$. This initial delay signifies the unfinished work to be served before the i-th customer's service, when the i-th customer of class p arrives at the system. From the definition of T_p (in Section 3.1), it follows that

$$E[T_p] = \sum_{i=1}^{\infty} \sum_{k_j \geq 0 (j \neq p)} \sum_{l=i}^{\infty} \frac{a(k_1, \cdots, k_{p-1}, l, k_{p+1}, \cdots, k_P)}{g_p} \cdot \frac{(i-1)b_p + \sum_{j=1}^{p-1} k_j b_j}{1 - \rho_{p-1}^{+}}.$$

Changing the order of the summation, we finally find that

$$E[T_p] = \frac{g_p^{(2)} b_p}{2g_p(1 - \rho_{p-1}^{+})} + \frac{\sum_{j=1}^{p-1} g_{jp} b_j}{g_p(1 - \rho_{p-1}^{+})} \quad \text{under both } PR \text{ and } NP. \tag{4.2}$$

We now consider the relationship between the unfinished work and waiting time as in [13, 19, 20]. For the time being, we will treat the NP rule. We introduce the following notation for class $p (1 \leq p \leq P)$:

Q_p : the number of class p customers in the queue (except for the server and limbo [20]),

U_p : unfinished work (load) for class p.

Note that the expected unfinished work of class p in the server is the mean forward recurrence time of the service time $b_p^{(2)}/(2b_p)$. Since the server is busy for class p with probability ρ_p, we have,

$$E[U_p] = E[Q_p]b_p + \rho_p \frac{b_p^{(2)}}{2b_p} . (1 \leq p \leq P)$$

The waiting time of the first customer in the associated batch, $W_{g,p}$, does not depend on the unfinished work of the lower class $\{p+1, \cdots, P\}$ customers except for that of a customer in service. The total unfinished work to be served when the tagged customer arrives, denoted by $U_p(B)$, is then given by

$$E[U_p(B)] = \sum_{j=1}^{p} E[Q_j]b_j + \sum_{j=1}^{P} \rho_j \frac{b_j^{(2)}}{2b_j} = \sum_{j=1}^{p} E[U_j] + \sum_{j=p+1}^{P} \rho_j \frac{b_j^{(2)}}{2b_j}.$$

The waiting time $W_{g,p}$ is the delay busy cycle generated by customers of the higher classes $\{1,2,\cdots,p-1\}$ with an initial delay of the unfinished work $U_p(B)$. Therefore, we have

$$\begin{aligned} E[W_{g,p}] &= \frac{E[U_p(B)]}{1-\rho_{p-1}^{+}} \\ &= \{\sum_{j=1}^{p} E[Q_j]b_j + \sum_{j=1}^{P} \rho_j \frac{b_j^{(2)}}{2b_j}\} \cdot \frac{1}{1-\rho_{p-1}^{+}}. \end{aligned} \tag{4.3}$$

It should be noted that (4.3) can be rigorously proven by using Brumelle's formula as in Takahashi [19]. From Little's formula $E[Q_p] = \lambda g_p E[W_p]$, (3.1), and (4.3), we have a recursive scheme for $E[W_p]$ as

$$E[W_p] = \frac{1}{1-\rho_p^{+}}\{\frac{1}{2}\sum_{j=1}^{P} \lambda g_j b_j^{(2)} + \frac{g_p^{(2)} b_p + s\sum_{j=1}^{p-1} g_{jp} b_j}{2g_p} + \sum_{j=1}^{p-1} \rho_j E[W_j]\}.$$

Thus we find the mean waiting time for individual class $p(1 \leq p \leq P)$

$$\begin{aligned} E[W_p] &= \frac{\sum_{j=1}^{P} \lambda g_j b_j^{(2)} + \sum_{j=1}^{p} \lambda g_j^{(2)} b_j^2}{2(1-\rho_{p-1}^{+})(1-\rho_p^{+})} + \frac{g_p^{(2)}}{2g_p} \cdot \frac{b_p}{1-\rho_{p-1}^{+}} \\ &+ \sum_{j=2}^{p-1} \rho_j \cdot \frac{\sum_{k=1}^{j-1} g_{kj} b_k}{g_j(1-\rho_{p-1}^{+}(1-\rho_p^{+})} + \sum_{k=1}^{p-1} \frac{g_{kp} b_p}{g_p(1-\rho_p^{+})} \quad \text{under } NP, \end{aligned} \tag{4.4}$$

where the empty sum ($\sum_{k=1}^{0}$, etc.) is assumed to be zero.

Under the PR rule, we get $E[W_p]$ by letting $g_{p+1} = \cdots = gp = 0$ in (4.4). We finally have

$$\begin{aligned} E[W_p] &= \frac{\sum_{j=1}^{p} \lambda[g_j b_j^{(2)} + g_j^{(2)} b_j^2]}{2(1-\rho_{p-1}^{+})(1-\rho_p^{+})} + \frac{g_p^{(2)}}{2g_p} \cdot \frac{b_p}{1-\rho_{p-1}^{+}} \\ &+ \sum_{j=2}^{p-1} \rho_j \cdot \frac{\sum_{k=1}^{j-1} g_{kj} b_k}{g_j(1-\rho_{p-1}^{+})(1-\rho_p^{+})} + \sum_{k=1}^{p-1} \frac{g_{kp} b_p}{g_p(1-\rho_p^{+})} \quad \text{under } PR. \end{aligned} \tag{4.5}$$

4.2 Other mean performance measures

From (4.4), (4.5) and Little's formula, the mean number of customers in the queue is obtained as,

$$E[Q_p] = \lambda g_p E[W_p] (1 \leq p \leq P) \tag{4.6}$$

under the PR and NP rules. Let L_p be the number of class p customers in the system (queue, server, and limbo). Since the mean completion time is $b_p/(1-\rho_{p-1}^{+})$ (see Wolff [20]), it follows from Little's formula that the mean number of class p customers in the server or in limbo is given by $\rho_p/(1-\rho_{p-1}^+)$ under PR. We then have for $1 \le p \le P$,

$$E[L_p] = \frac{\rho_p}{1-\rho_{p-1}^+} + E[Q_p] \quad \text{under } PR, \tag{4.7}$$

and

$$E[L_p] = \rho_p + E[Q_p] \quad \text{under } NP, \tag{4.8}$$

which has been directly shown for $P = 2$ in Section 2.

The mean sojourn time of a customer at the system, $E[S_p]$, is obtained again from Lettle's formula,

$$E[S_p] = E[L_p]/(\lambda g_p) \quad (1 \le p \le P) \tag{4.9}$$

under the PR and NP rules. The delay time of a class p customer is, here, defined as the sojourn time at the queue and at limbo. The mean delay time $E[D_p]$ is then obtained from

$$E[D_p] = E[S_p] - b_p \quad (1 \le p \le P) \tag{4.10}$$

under both the PR and NP rules.

Remark 4.1.

If $P = 2$, it is seen that the above mean formulas (4.7) and (4.8) are reduced to equations (2.28), (2.29) and (2.36), (2.37) obtained in Section 2.

We now consider a priority queue with an arriving batch being composed of a single class only, which is treated by Hawkes [4], Meister [12], Takagi, et al. [18], and Takahashi [19]. We denote by λ_p and $G_p(z_p)$ the arrival rate of batches and the generating function of batch size for class $p (1 \le p \le P)$. We then have

$$G(\underline{z}) = \sum_{p=1}^{P} \frac{\lambda_p}{\lambda} G_p(z_p),$$

from which it follows that $g_{ij} = [\frac{\partial^2 G(\underline{z})}{\partial z_i \partial z_j}]_{\underline{z}=\underline{1}} = 0$, if $i \neq j$. It can be seen that equations (4.4) and (4.5) are reduced to the previous results by [4, 12, 18, 19]. In other words, denoting by $E[W_p(IN)]$ the mean waiting time in an ordinary independent arrival queue considered in this remark, it follows from equations (4.4) and (4.5) that

$$E[W_p(ST) = E[W_p(IN)] + \sum_{j=2}^{p-1} \rho_j \cdot \frac{\sum_{k=1}^{j-1} g_{kj} b_k}{g_j(1-\rho_{p-1}^+)(1-\rho_p^+)} + \sum_{k=1}^{p-1} \frac{g_{kp} b_p}{g_p(1-\rho_p^+)} \tag{4.11}$$

under the PR and NP rules $(1 \le p \le P)$. Here, $E[W_p(ST)]$ is the mean waiting time in a structured priority queue. Note that the second and third terms in the right-hand side of (4.11) represent the influence of the dependency between classes in a batch upon the mean waiting time. Since it is easily verified that $g_{ij} \ge 0 \ (1 \le i, j \le P)$, we find under both the PR and NP rules

$$E[W_p(ST) \ge E[W_p(IN)] \ (1 \le p \le P).$$

This inequality is also valied between a couple of the above-derived mean performance measures in the ordinary independent arrival and structured priority queues. It should be noted that the mean performance measure of the highest priority class (class 1) customers is independent of $g_{ij}(1 \leq i, j \leq P)$ and so that we have for example, $E[D_1(ST)] = E[D_1(IN)]$. □

We now consider a structured priority queue with $G(\underline{z})$ having a special form as $G(z_1, z_2, \cdots, z_P) = H(z_1 \cdot z_2 \cdot \cdots \cdot z_P)$,
where $H(z)$ is the generaing function of a (discrete) random variable. This structured priority queue can be seen in a packet communication system (see Section 1). In such a system, $H(z)$ corresponds to the generating function of packet length (the number of packest in a message). Each packet in a message requires P operations. We assume that each packet makes $P-1$ copies of itself, and assume that each (original) packet belongs to class 1 and the copied packets belong to classes p $(2 \leq p \leq P)$. The i-th original packet in an arriving message which is assumed to be of class 1 generates the i-th copied packets of the lower classes (class 2, class 3, $\cdots$, and class P). One of the desirable performance measures in this example is the sojourn time of the i-th original and copied packets, i.e., the time that starts with the i-th original packet arriving at the system and ends when the i-th copied packet of class P departs the system. We denote by $S(i)$ the sojourn time of the i-th original and copied packets at the system. It follows that

$$E[S(i)] = E[W_{g,P}] + i \cdot b_P/(1 - \rho^+_{P-1}) \quad \text{under } PR, \tag{4.12}$$

and

$$E[S(i)] = E[W_{g,P}] + (i-1) \cdot b_P/(1 - \rho^+_{P-1}) + b_P \quad \text{under } NP. \tag{4.13}$$

Let $\{h_i\}$ be the packet length distribution in an arriving message with mean h and the n-th factorial moment $h^{(n)}$, so that

$$h \equiv \sum_{i=1}^{\infty} i \cdot h_i = \frac{d}{dz} H(z) \mid_{z=1} \text{ and } h^{(n)} = \frac{d^n}{dz^n} H(z) \mid_{z=1} (n = 2, 3, \cdots).$$

Applying Burke's result [1], the probability that an arbitrary customer is served i-th in its batch, $r(i)$, is given by

$$r(i) = \sum_{l=i}^{\infty} h_l / h. \tag{4.14}$$

The mean sojourn time of packets, $E[S']$, is given by

$$E[S'] = \sum_{i=1}^{\infty} r(i) E[S(i)].$$

Substituting (4.12) through (4.14) into the above equation, we find that

$$E[S'] = E[W_{g,P}] + b_P \cdot \frac{h^{(2)} + 2h}{2(1 - \rho^+_{P-1})} \quad \text{under } PR,$$

and

$$E[S'] = E[W_{g,P}] + b_P \cdot [1 + \frac{h^{(2)}}{2(1 - \rho^+_{P-1})}] \quad \text{under } NP.$$

5. Conclusion

We have derived the joint generating function of the queue lengths distribution and the LST of the waiting time distribution in a structured priority queue with two classes ($P = 2$).

We have also provided the mean waiting time formula for an individual class in the queue with general P classes. The results obtained here are shown to be reduced to the previously derived formulas [1, 4, 5, 6, 12, 18, 19] as special cases. A structured priority queue with a general arrival process will be a future research topic.

Appendix

The root $z = \tilde{z}_1$ of the equation: $z - B_1^*[\lambda(1 - G(z, z_2)) + s] = 0$.

We introduce the following functions of z for $Re(s) > 0, |z_2| \leq 1$.
$f(z) \equiv z$, and
$g(z) \equiv B_1^*[\lambda(1 - G(z, z_2)) + s]$.
It is obvious that $f(z)$ has one and only one zero inside the unit circle $C \equiv \{z : |z| = 1\}$. If we show $|f(z) > |g(z)|$ on $C, f(z)$ and $f(z) - g(z)$ have the same number of zeros inside C from Rouché's theorem, i.e., $f(z) - g(z)$ has one and only one zero inside C.
Now we have for all $z \in C$,

$$\begin{aligned} |g(z)| &= |\int_0^\infty \exp\{-\lambda(1 - G(z, z_2))t + st\} dB_1(t)| \\ &\leq \int_0^\infty \exp\{-Re(s)t\} dB_1(t) \\ &< 1 = |f(z)|, \end{aligned}$$

since $Re(G(z, z_2)) \leq |G(z, z_2) \leq 1$ and since $Re(s) > 0$. This validates the existence and uniqueness of the root $z = \tilde{z}_1$ (inside C) of the equation: $z - B_1^*[\lambda(1 - G(z, z_2)) + s] = 0$. □

Acknowledgement

The authors are grateful to Professor On Hashida, The University of Tsukuba, Tokyo (Former Research Fellow of NTT Laboratories) for his helpful suggestions and constructive comments.

References

[1] Burke, P.J.: Delays in single-server queues with batch input, *Operations Res.*, Vol. 23, (1975), 830–833.
[2] Chaudhry, M.L. and Templeton, J.G.C.: *A First Course in Bulk Queues*, John Wiley & Sons, Inc., New York, 1983.
[3] Fujiki, M. and Gambe, E.: *Teletraffic Theory*, Maruzen, Tokyo, 1980. (*in Japanese*)
[4] Hawkes, A.G.: Time-dependent solution of a priority queue with bulk arrival, *Operations Res.*, Vol. 13, (1965), 586–595.
[5] Jaiswal, N.K.: Preemptive resume priority queue, *Operations Res.*, Vol. 9, (1961), 732–742.
[6] Jaiswal, N.K.: Time-dependent solution of the head-of-the-line priority queue, *J. Roy. Stat. Soc.*, Series B, Vol. 24, (1962), 91–101.
[7] Jaiswal, N.K.: *Priority Queues*, Academic Press, New York, 1968.
[8] Keilson, J. and Kooharian, A.: On time-dependent queueing processes, *Ann. Math. Stat.*, Vol. 31 (1960), 104–112.
[9] Kella, O. and Yechiali, U.: Priorities in $M/G/1$ queue with server vacations, *Naval Res. Logistices*, Vol. 35 (1988), 23–34.

[10] Kuribayashi, S. and Takahashi, Y.: Diffusion approximation for a multi-server non-preemptive priority queue with batch Poisson inputs and general service time, *Trans. IEICE*, J71-B (1988), 121-128. (*in Japanese*).
[11] Manfield, D.R. and Tran-Gia, P.: Analysis of a finite storage system with batch input arrising out of message packetization, *IEEE Trans. Commun.*, COM-30 (1982), 456-463.
[12] Meister, B.: Waiting time in a preemptive resume system with compound Poisson input, *Computing*, Vol. 25 (1980), 17-28.
[13] Schrage, L.: An alternative proof of a conservation law for the queue $G/G/1$, *Operations Res.*, Vol. 18 (1970), 185-187.
[14] Sidi, M. and Segall, A.: Two interfering queues in packet-radio networks, *IEEE Trans. Commun.*, COM-31 (1983), 123-129.
[15] Sidi, M. and Segall, A.: Structured priority queueing systems with applications to packet-radio networks, *Performance Evaluation*, Vol. 3 (1983), 265-275.
[16] Sidi, M.: Two competing discrete-time queues with priority, *Queueing Systems*, Vol. 3 (1988), 347-362.
[17] Stuck, B.W. and Arthurs, E.: *A Computer and Communications Network Performance Analysis Primer*, Prentice-Hall, New Jersey, 1985.
[18] Takagi, H. and Takahashi, Y.: Priority queues with batch Poisson arrivals, (submitted to *Operations Res. Letters.*)
[19] Takahashi, Y.: On the relationship between work load and waiting time in single server queues with batch inputs, *Operations Res. Letters*, Vol. 7 (1988), 51-56.
[20] Wolff, R.W.: Work-conserving priorities, *J. Appl. Prob.*, Vol. 7 (1970), 327-337.

Yoshitaka Takahashi
Teletraffic Research Department
NTT Communication Switching Laboratories
3-9-11 Midori-cho, Musashino-shi,
Tokyo 180,
Japan

Hideaki Takagi
IBM Research,
Tokyo Research Laboratory
5-19 Sanban-cho, Chiyoda-ku,
Tokyo 102,
Japan