Journal of the Operations Research
Society of Japan
Vol. 42, No. 3, September 1999

ANALYSIS OF MULTICLASS M/G/1 QUEUES WITH A MIXTURE OF 1-LIMITED DISCIPLINES AND GATED DISCIPLINES

Tetsuji Hirayama
University of Tsukuba

(Received April 30, 1997; Final March 30, 1999)

Abstract In this paper, we consider single server queues with several groups and several classes of customers. We consider priority scheduling algorithms for the multiclass queues in which the server admits customers in each group into the service facility by 1-limited disciplines or by gated disciplines. Our objective is to show a method for deriving mean waiting times for these multiclass M/G/1 queues. From the analysis of the busy periods, we investigate some linear structure inherent in the mean waiting times conditioned on the system state at each customer's arrival epoch. The steady state mean waiting times can be derived from the linear structure by using the Little's formula and the PASTA property.

1. Introduction

The gated disciplines operate as follows. The server opens the gate of the service facility at some time epoch in order to serve some of customers in the system. After admitting these customers into the service facility, the server closes its gate to prohibit the other customers from entering the service facility. After completing services of the customers in the service facility, the server again opens its gate to serve other waiting customers. Based on timings for opening the gate and rules for which customers may be admitted into the service facility, the following two types of gated disciplines have been investigated:

1. (Pure) gated disciplines.
2. Gated batch disciplines.

For the pure gated disciplines, the server selects one of the stations and opens its gate, and then he admits all customers at the station at that time into the service facility and closes its gate. When all customers in the service facility complete their services, the server again selects one of the stations and opens its gate. These disciplines have been investigated in connection with polling systems [5]. We may also consider the gated disciplines for priority queues. For the gated batch disciplines, the server admits a customer into the service facility who arrives at the empty system and closes its gate. At his service completion epoch, the server again opens its gate to admit a 'batch' consists of all customers at the system at that time into the service facility. After completing their services, the server again opens its gate to admit a 'batch' consists of all customers at the system at that time into the service facility, and so forth [21]. The server serves customers in the service facility in a priority order in the gated batch priority discipline. A variation of the discipline where customers in the service facility are served in a processor sharing fashion has been analyzed [1, 18]. The other type of the gated discipline called the 'binomial gated discipline' is investigated in connection with polling systems [15]. An interesting variant called the 'selfish scheduling algorithm' has been investigated by Kleinrock [13].

We consider priority scheduling algorithms with a mixture of 1-limited disciplines and gated disciplines. Customers in the system are classified into J groups. Group i has a higher priority than group j if $i < j$. Further group i consists of L_i classes of customers $(i = 1, \cdots, J)$. According to the service disciplines adopted by the groups, they are classified into one of the following two types: 1-limited groups and gated groups. When a customer arrives at the empty system, the server immediately admits him into the service facility and then closes the gate. At his service completion epoch, the server selects a group with the highest priority and opens its gate in order to admit some of its customers into the service facility. If the selected group is the 1-limited group, one of the customers belonging to the group enters the service facility and then the gate is closed. If the selected group is the gated group, all customers belonging to the group at that time enter the service facility and then the gate is closed. The customers in the service facility are served according to a predetermined service order (FCFS or priority within the group). When their services are completed, the server again selects a group with the highest priority and opens its gate, and so on. If all groups are gated groups and $L_i = 1$ for all groups $(i = 1, \cdots, J)$, the scheduling algorithm is a pure gated discipline (for priority queues). If the system consists of a gated group ($J = 1$ and $L_1 \geq 1$), the scheduling algorithm is the gated batch (priority) discipline. If all groups are 1-limited groups and $L_i = 1$ for all groups $(i = 1, \cdots, J)$, the scheduling algorithm is an ordinary fixed priority discipline.

Analysis of priority queues has been accomplished by various methods. Cobham [4] solved a set of linear equations in order to obtain the mean waiting times. The method of supplementary variables is used to obtain the time dependent quantities [11]. The Laplace-Stieltjes transforms (LSTs) of the waiting time distributions are obtained by the method of embedded Markov chains [16]. Recently two methods for investigating distributions of waiting times for priority queues with server vacations are appeared. One is the method presented in [14] in which the delay cycle analysis effectively gives the conditional LSTs of the waiting time distributions. The method can be used to analyze various types of priority queues [21, 22]. The other is the method presented in [20] in which the LSTs of the virtual waiting time distributions are obtained by the 'level crossing analysis'. This method is also used to analyze M/G/1 queues with preemption-distance priorities [17]. Although the Cobham's method for the mean waiting times is simple, these methods for the waiting time distributions are so complicated that we should study many stochastic arguments. Our objective is to show a method for *mean* waiting times of the multiclass M/G/1 queues with mixtures of the 1-limited disciplines and the gated disciplines. We first define a stochastic process that represents the system states of the queues. We further define the mean waiting times conditioned on the system states at each customer's arrival epoch. Then we investigate some linear structure inherent in the conditional mean waiting times from the analysis of the busy periods. Similar linear structure can be found in the other multiclass M/G/1 queues with feedbacks [10]. The steady state mean waiting times can be derived in a straightforward manner from the linear structure by using the Little's formula and the PASTA property.

These composite priority queues have been used to analyze programmable terminal control units [6]. Recently, workgroup networking demands higher bandwidth as users increasingly share and access data across the network. More powerful workstations promote multiple classes of high-bandwidth networked applications using imaging, graphics, and multimedia. Switches are tools for increasing bandwidth, controlling traffic, and dispelling congestion. Using switches enables us to preserve the investment in an existing LAN infrastructure while maintaining the ability to easily grow the LAN in the future [8]. A switch fabric is a method used to actually route a packet from one (input) port to another (output)

port. There are two typical designs for the switch fabric [8, 23]. 'Cross-point matrix (or point-to-point matrix)' is one of these designs in which all ports are internally connected to all other ports. Another design is a 'shared bus (or bus-based)' in which an internal high-speed backplane is used to interconnect switch ports. These switches usually support packet priorities. Since packets at input ports can be transmitted one at a time through the shared bus, single server queues with multiple customer classes and priorities can be used to analyze effectiveness of scheduling algorithms at the shared bus when these scheduling algorithms do not quite affect behaviors of packets waiting for transmissions at output ports. Each class of application at each port corresponds to a class of customers. Behaviors of packets at the output ports of shared bus switches have been investigated in [2], and behaviors of cross-point matrix switches have been investigated in [3, 12]. When we design efficient packet transmissions under the prospective diversification of network service requirements, we should investigate the effects of their scheduling algorithms including various priorities and service orders of several types of packets on their system performances. This paper contributes to the quantitative evaluations of these scheduling algorithms.

2. Model Description

We consider multiclass M/G/1 queues. Customers in the system are classified into J groups. Customers belonging to group i stay at station i $(i = 1, \cdots, J)$. Further group i consists of L_i classes of customers $(i = 1, \cdots, J)$. Let $\mathcal{S} \equiv \{(i, \alpha) : i = 1, \cdots, J \text{ and } \alpha = 1, \cdots, L_i\}$ be a set of station-class pairs of the system and let $J_c \equiv \sum_{i=1}^{J} L_i$ be the total number of classes. Class α customers belonging to group i arrive at station i from outside the system according to a Poisson process with rate $\lambda_{i\alpha} > 0$ $((i, \alpha) \in \mathcal{S})$. Since these arrival processes are assumed to be independent, the overall arrival process at station i is also a Poisson process with rate $\lambda_i \equiv \sum_{\alpha=1}^{L_i} \lambda_{i\alpha}$, and the overall arrival process at the system is also a Poisson process with rate $\lambda \equiv \sum_{i=1}^{J} \lambda_i$. A class α customer at station i is called an (i, α)-*customer*. A single server serves customers at these stations. Service times $S_{i\alpha}$ of (i, α)-customers are independently, identically and arbitrarily distributed with mean $E[S_{i\alpha}] > 0$ and second moment $\overline{s^2}_{i\alpha}$ $((i, \alpha) \in \mathcal{S})$. Customers are served according to a predetermined *scheduling algorithm* defined below. After receiving services, they depart from the system. The arrival processes and the service times are assumed to be independent of each other. We define intensities ρ_j and ρ_j^+ in the following manner:

$$\rho_j \equiv \sum_{\alpha=1}^{L_j} \lambda_{j\alpha} E[S_{j\alpha}], \quad j = 1, \cdots, J,$$

$$\rho_j^+ \equiv \begin{cases} 0, & j = 0, \\ \sum_{i=1}^{j} \sum_{\alpha=1}^{L_i} \lambda_{i\alpha} E[S_{i\alpha}], & j = 1, \cdots, J. \end{cases}$$

Then we put the usual assumption that $\rho_J^+ < 1$.

The system is separated into two parts which are called the 'service facility' and the 'waiting rooms' of the stations. The server selects one of the stations at a time, and then opens its gate, which separates the service facility from its waiting room, in order to admit some customers in the station to the service facility. Then the server closes its gate and serves customers in the service facility until he empties it, and then selects another station and opens its gate. Since the gates of the stations that are not selected by the server are closed, all customers in such stations must wait for service at their waiting rooms. If there is at least a customer in the system, the server selects one of the stations to serve its customers.

Once a customer begins a service, his service is not interrupted by other customers' (non-preemptive). Hence, at any time epoch, customers in each station are classified into the following three types:

- a cutomer being served,
- customers in the service facility (who are not being served), and
- customers in the waiting rooms.

Let $\mathcal{R}, \mathcal{R}_+, \mathcal{I}_+$ be respectively a set of real numbers, a set of nonnegative real numbers, and a set of nonnegative integers. Then let (κ, a) denote the station-class pair of a customer being served and let r denote his remaining service time. Number of (i,α)-customers in the service facility (who are not being served) is denoted by $g_{i\alpha}$ $((i,\alpha) \in \mathcal{S})$. The vectors are denoted by $\boldsymbol{g}_i \equiv (g_{i\alpha} : \alpha = 1, \cdots, L_i) \in \mathcal{I}_+^{L_i}$ $(i = 1, \cdots, J)$, and $\boldsymbol{g} \equiv (\boldsymbol{g}_1, \cdots, \boldsymbol{g}_J) \in \mathcal{I}_+^{L_1} \times \cdots \times \mathcal{I}_+^{L_J}$. Number of (i,α)-customers in the waiting room is denoted by $n_{i\alpha}$ $((i,\alpha) \in \mathcal{S})$. The vectors are denoted by $\boldsymbol{n}_i \equiv (n_{i\alpha} : \alpha = 1, \cdots, L_i) \in \mathcal{I}_+^{L_i}$ $(i = 1, \cdots, J)$, and $\boldsymbol{n} \equiv (\boldsymbol{n}_1, \cdots, \boldsymbol{n}_J) \in \mathcal{I}_+^{L_1} \times \cdots \times \mathcal{I}_+^{L_J}$. Further let $(\kappa(t), a(t))$ denote the station-class pair of a customer being served at time t and let $r(t)$ denote his remaining service time at time t. The number of (i,α)-customers in the service facility at time t (who are not being served) is denoted by $g_{i\alpha}(t)$ and the number of (i,α)-customers in the waiting room at time t is denoted by $n_{i\alpha}(t)$. Let $\boldsymbol{g}_i(t) \equiv (g_{i\alpha}(t) : \alpha = 1, \cdots, L_i) \in \mathcal{I}_+^{L_i}$ $(i = 1, \cdots, J)$, and let $\boldsymbol{g}(t) \equiv (\boldsymbol{g}_1(t), \cdots, \boldsymbol{g}_J(t)) \in \mathcal{I}_+^{L_1} \times \cdots \times \mathcal{I}_+^{L_J}$. Further, let $\boldsymbol{n}_i(t) \equiv (n_{i\alpha}(t) : \alpha = 1, \cdots, L_i) \in \mathcal{I}_+^{L_i}$ $(i = 1, \cdots, J)$, and let $\boldsymbol{n}(t) \equiv (\boldsymbol{n}_1(t), \cdots, \boldsymbol{n}_J(t)) \in \mathcal{I}_+^{L_1} \times \cdots \times \mathcal{I}_+^{L_J}$.

Each time interval from when the server opens the gate to admit some customers into the service facility until the first time when the server completes all of their services and again opens the gate is called a *service period.* Specifically, if we would like to specify the station selected by the server, we call the period 'the service period of the station'. Further if we would like to specify that a specific customer is scheduled to serve during the period, we call it 'his service period'. If the system becomes empty, the server becomes idle $((\kappa, a) = (0,0))$. We denote a set of these periods of the stations by $\mathit{\Pi} \equiv \{0, 1, \cdots, J\}$.

Customers in the system are served according to a predetermined *scheduling algorithm.* A scheduling algorithm is a set of decision rules determining which customer will next be serviced and for how long [13]. We will prescribe the scheduling algorithms according to the following specifications:

- Selection orders of the stations.
- Customer selection rules when the server admits customers into the service facility.
- Service orders of customers in the service facility.

We consider *priority* selection orders of the stations for which group i has priority over group j if $i < j$. Then the server always selects a station (a group) with the highest priority among all non-empty stations. If a customer arrives at any empty system, the server selects to serve him immediately. The server should not be idle if there is at least a customer in the system (non-idling).

The customer selection rule for each station is either

- 1-limited, or
- gated.

When the server selects one of the groups with the 1-*limited* discipline, one of customers belonging to the group enters the service facility at a time. A group with the 1-limited discipline is called a 1-*limited group.* If the server selects a 1-limited group with the FCFS discipline, he admits a customer into the service facility who has arrived at the station earliest of all customers within the group. $\mathcal{H}_{1F}$ denotes the set of the 1-limited groups with

the FCFS discipline. If the server selects a 1-limited group with the fixed priority discipline, he admits a customer into the service facility who has the highest priority among all customers at the station. A 1-limited group with the fixed priority discipline is decomposed into multiple 1-limited groups with the FCFS discipline. Hence, for the 1-limited groups, we only consider the FCFS discipline. When the server selects one of the groups with the *gated* discipline, all customers belonging to the group at this moment enter the service facility in batches at a time. A group with the gated discipline is called a *gated group.*

The service order of customers in the service facility is either

- FCFS order, or
- fixed priority order.

The 1-limited groups are not concerned with these orders. If the server selects one of the gated groups with the FCFS order, he serves all customers in the service facility in a first come first served order. $\mathcal{H}_{gF}$ denotes the set of the gated groups with the FCFS order. If the server selects one of the gated group with the fixed priority order, he serves all customers in the service facility according to a fixed priority order where class α customers in the group has priority over class β customers in the group if $\alpha < \beta$. Customers in each class are served in a first come first served order. $\mathcal{H}_{gP}$ denotes the set of the groups with the fixed priority order.

Let us consider the system operated under some fixed scheduling algorithm. The e^{th} customer arrives from outside the system at epoch σ^e $(e = 1, 2, \cdots)$. We denote him by $\boldsymbol{c}^e$. We specify information of the system at time t: $L(t) \equiv \{(j_m(t), \beta_m(t), s_m(t)) : m = 1, 2, \cdots\}$ where $(j_m(t), \beta_m(t))$ is a station-class pair and $s_m(t)$ is a status of a customer who has arrived the m^{th} earliest of all customers in the system at time t $(m = 1, 2, \cdots)$ ($s_m(t)$ = expended service time at t if the customer is being served, = 's' if he is in the service facility, or = 'w' if he is in the waiting room). Let us consider transition epochs of these processes consist of customer arrival epochs and service completion epochs. Then let $(X(t), \Gamma(t))$ denote a station-class pair of an arriving customer at the last transition epoch before or on t $(t \geq 0)$; if it is not a customer arrival epoch, then $(X(t), \Gamma(t)) = (0, 0)$. $X(t)$ and $\Gamma(t)$ is right continuous with left-hand limits. Then we define the stochastic process $\mathcal{Q} = \{\boldsymbol{Y}(t) = (X(t), \Gamma(t), \kappa(t), a(t), r(t), \boldsymbol{g}(t), \boldsymbol{n}(t), L(t)) : t \geq 0\}$ that represents an evolution of the system. For any scheduling algorithm defined above, $\mathcal{Q}$ may embed a Markov process whose transition epochs consist of customer arrival epochs and service completion epochs. Possible values of $\boldsymbol{Y}(t)$ $(t \geq 0)$ are called *states.* The state space of $\mathcal{Q}$ is denoted by $\mathcal{E}$.

We would like to derive two types of cost functions defined below. First type of the cost functions is related to the waiting times of customers in the waiting room. Let e $(= 1, 2, \cdots)$ be a customer number. We define

$$C^e_{Wi\alpha}(t) \equiv \begin{cases} 1, & \text{if } \boldsymbol{c}^e \text{ stays in the waiting room at station } i \\ & \text{as a class } \alpha \text{ customer at time } t, \\ 0, & \text{otherwise,} \end{cases} \tag{2.1}$$

for $t \geq 0$ and $(i, \alpha) \in \mathcal{S}$. Further we define

$$W^e_{i\alpha} \equiv \int_0^\infty C^e_{Wi\alpha}(t)dt, \quad (i, \alpha) \in \mathcal{S}. \tag{2.2}$$

Then their expected values conditioned on the state of the system at his arrival epoch are defined by

$$W_{i\alpha}(\boldsymbol{Y}, e) \equiv E\Big[\int_{\sigma^e}^\infty C^e_{Wi\alpha}(t)dt | \boldsymbol{Y}(\sigma^e) = \boldsymbol{Y}\Big], \quad (i, \alpha) \in \mathcal{S}, \tag{2.3}$$

for $\boldsymbol{Y} \in \mathcal{E}$. $W_{i\alpha}(\boldsymbol{Y}, e) \equiv 0$ for $\boldsymbol{Y} = (j, \beta, \kappa, a, r, \boldsymbol{g}, \boldsymbol{n}, L) \in \mathcal{E}$ such that $(i, \alpha) \neq (j, \beta)$.

Second type of the cost functions is related to the waiting times of customers in the service facility. Let e $(= 1, 2, \cdots)$ be a customer number. We define

$$C^e_{Gi\alpha}(t) \equiv \begin{cases} 1, & \text{if } \boldsymbol{c}^e \text{ waits for service in the service facility} \\ & \text{as an } (i,\alpha)\text{-customer at time } t, \\ 0, & \text{otherwise,} \end{cases} \tag{2.4}$$

for $t \geq 0$ and $(i, \alpha) \in \mathcal{S}$. Further we define

$$G^e_{i\alpha} \equiv \int_0^\infty C^e_{Gi\alpha}(t)dt, \quad (i, \alpha) \in \mathcal{S}. \tag{2.5}$$

Then their expected values conditioned on the state of the system at his arrival epoch are defined by

$$G_{i\alpha}(\boldsymbol{Y}, e) \equiv E\Big[\int_{\sigma^e}^\infty C^e_{Gi\alpha}(t)dt | \boldsymbol{Y}(\sigma^e) = \boldsymbol{Y}\Big], \quad (i, \alpha) \in \mathcal{S}, \tag{2.6}$$

for $\boldsymbol{Y} \in \mathcal{E}$. $G_{i\alpha}(\boldsymbol{Y}, e) \equiv 0$ for $\boldsymbol{Y} = (j, \beta, \kappa, a, r, \boldsymbol{g}, \boldsymbol{n}, L) \in \mathcal{E}$ such that $(i, \alpha) \neq (j, \beta)$. Since every customer in each of the 1-limited groups receives a service immediately on entering the service facility, the cost function $G_{i\alpha}(\boldsymbol{Y}, e)$ is always equal to zero for 1-limited group i.

For any e $(= 1, 2, \cdots)$, the total waiting time that $\boldsymbol{c}^e$ spends from when he arrives from outside the system until his service begins is given by

$$W^e_{X,\Gamma} + G^e_{X,\Gamma}, \tag{2.7}$$

where $(X, \Gamma) = (X(\sigma^e), \Gamma(\sigma^e))$. Its expected value given that he arrives at the system in state $\boldsymbol{Y} = (j, \beta, \kappa, a, r, \boldsymbol{g}, \boldsymbol{n}, L) \in \mathcal{E}$ is given by

$$W_{j\beta}(\boldsymbol{Y}, e) + G_{j\beta}(\boldsymbol{Y}, e). \tag{2.8}$$

(Although there may be some redundancies in the definitions of these cost functions and the other quantities, they are prepared for further extensions of the system. See, for example, [9].)

3. Busy Periods of the System

Now we consider the system with any scheduling algorithm defined in Section 2 is in state $\boldsymbol{Y} = (j, \beta, \kappa, a, r, \boldsymbol{g}, \boldsymbol{n}, L) \in \mathcal{E}$ at any arrival epoch τ. Let SA_P denote the scheduling algorithm. For any specified group k $(k \leq j)$, we consider busy period processes related to group k. Since it is convenient to consider customers in a mass who belong among groups $1, \cdots, k-1$ and who arrive after time τ, these customers are called $\boldsymbol{c}(k-1)$-customers. (For convenience, $\boldsymbol{c}(0)$-customers indicate an empty set of customers.)

First let us consider the case: $k \in \mathcal{H}_{1F}$. Let $B^{(k,1F)}_P$ be a length of time lasting from time τ until the first epoch after τ when all of the following customers are cleared from the system: 1) customers who are in the service facility (including a customer in service) at τ, 2) customers belonging among groups $1, \cdots, k$ who are in the waiting room at τ (except for a customer arriving at τ), and 3) $\boldsymbol{c}(k-1)$-customers. Its expected value conditioned on the system state $\boldsymbol{Y}$ at τ is denoted by $\overline{B^{(k,1F)}_P}(\boldsymbol{Y})$. Let $\mathcal{C}^1$ denote a set of customers who are in the service facility (including a customer in service) at τ and customers belonging among groups $1, \cdots, k$ who are in the waiting room at τ (except for a customer arriving at τ). We can characterize algorithm SA_P (after epoch τ) as follows. After completing a

service of a customer in service, it begins services of customers in the service facility at τ. Then it serves customers belonging among groups $1, \cdots, k$ in the waiting room at τ and $\boldsymbol{c}(k-1)$-customers (if any) in the priority order prescribed by SA_P. By the definition of the scheduling algorithm, customers other than customers in $\mathcal{C}^1$ and $\boldsymbol{c}(k-1)$-customers should not begin their services before epoch $\tau + B_P^{(k,1F)}$. Customers in $\mathcal{C}^1$ can be numbered in the (relative) order of their services such that $\boldsymbol{c}_0$ is the first customer to be served, $\boldsymbol{c}_1$ is the second customer, $\cdots$, and $\boldsymbol{c}_{N^1}$ is the final customer, where N^1+1 is the number of customers in $\mathcal{C}^1$.

Further we define another scheduling algorithm $SA_L^{(k,1F)}$ as follows. It takes the same algorithm as SA_P before epoch τ, and after the epoch it begins services of customers in $\mathcal{C}^1$ and $\boldsymbol{c}(k-1)$-customers in a non-preemptive LCFS order. (The relative service order of customers in $\mathcal{C}^1$ is as follows: $\boldsymbol{c}_0$ is the (relatively) first customer to be served, $\boldsymbol{c}_{N^1}$ is the (relatively) second customer, $\cdots$, and $\boldsymbol{c}_1$ is the (relatively) final customer.) Other customers are served in the same algorithm as SA_P. Let $B_L^{(k,1F)}$ be a length of time under $SA_L^{(k,1F)}$ lasting from time τ until the first epoch after τ when customers in $\mathcal{C}^1$ and $\boldsymbol{c}(k-1)$-customers are cleared from the system. Its expected value conditioned on the system state $\boldsymbol{Y}$ at τ is denoted by $\overline{B_L^{(k,1F)}}(\boldsymbol{Y})$.

Since scheduling algorithms SA_P and $SA_L^{(k,1F)}$ are work conserving, the total amount of remaining service times of $\boldsymbol{c}(k-1)$-customers plus remaining service times of customers in $\mathcal{C}^1$ under SA_P is equal to that under $SA_L^{(k,1F)}$. Since $B_P^{(k,1F)}$ and $B_L^{(k,1F)}$ are lengths of time between epoch τ and the first epoch when the total amount of remaining service times of $\boldsymbol{c}(k-1)$-customers and customers in $\mathcal{C}^1$ becomes zero, these are equivalent. Hence we would like to obtain the expected value $\overline{B_L^{(k,1F)}}(\boldsymbol{Y})$. The period $B_L^{(k,1F)}$ can be divided into N^1+1 sub-periods $B_{L,l}^{(k,1F)}$ $(l=0,1,\cdots,N^1)$ where $B_{L,l}^{(k,1F)}$ is initiated by the service of $\boldsymbol{c}_l$ and is terminated at the first epoch after $\boldsymbol{c}_l$'s service completion when the system is cleared of $\boldsymbol{c}(k-1)$-customers. Since $B_{L,l}^{(k,1F)}$ $(l=0,1,\cdots,N^1)$ is an *exceptional first service busy period* (Sec. 8–5 in [26]), its expected value is given by $\overline{s_l}/(1-\rho_{k-1}^+)$ where $\overline{s_l}$ is the expected value of $\boldsymbol{c}_l$'s service time. Since $B_L^{(k,1F)}$ is a sum of $B_{L,l}^{(k,1F)}$ $(l=0,1,\cdots,N^1)$, we have

$$\begin{aligned}\overline{B_P^{(k,1F)}}(\boldsymbol{Y}) &= \overline{B_L^{(k,1F)}}(\boldsymbol{Y}) = \sum_{l=0}^{N^1} \frac{\overline{s_l}}{1-\rho_{k-1}^+} \\ &= \frac{r + \sum_{i=1}^{J}\sum_{\alpha=1}^{L_i} g_{i\alpha} E[S_{i\alpha}] + \sum_{i=1}^{k}\sum_{\alpha=1}^{L_i} n_{i\alpha} E[S_{i\alpha}]}{1-\rho_{k-1}^+}. \end{aligned} \tag{3.1}$$

Next let us consider the case: $k \in \mathcal{H}_{gF} \cup \mathcal{H}_{gP}$. Let $B_P^{(k,g)}$ be a length of time lasting from time τ until the first epoch after τ when all of the following customers are cleared from the system: 1) customers who are in the service facility (including a customer in service) at τ, 2) customers belonging among groups $1, \cdots, k-1$ who are in the waiting room at τ, and 3) $\boldsymbol{c}(k-1)$-customers. Its expected value conditioned on the system state $\boldsymbol{Y}$ at τ is denoted by $\overline{B_P^{(k,g)}}(\boldsymbol{Y})$. Let $\mathcal{C}^g$ denote a set of customers who are in the service facility (including a customer in service) at τ and customers belonging among groups $1, \cdots, k-1$ who are in the waiting room at τ. By the definition of SA_P, customers other than customers in $\mathcal{C}^g$ and $\boldsymbol{c}(k-1)$-customers should not begin their services before epoch $\tau + B_P^{(k,g)}$. Customers in $\mathcal{C}^g$ can be numbered in the (relative) order of their services such that $\boldsymbol{c}'_0$ is the first customer to be served, $\boldsymbol{c}'_1$ is the second customer, $\cdots$, and $\boldsymbol{c}'_{N^g}$ is the final customer, where N^g+1 is the number of customers in $\mathcal{C}^g$.

Further we define another scheduling algorithm $SA_L^{(k,g)}$ as follows. It takes the same algorithm as SA_P before epoch τ, and after the epoch it begins services of customers in $\mathcal{C}^g$ and $\boldsymbol{c}(k-1)$-customers in a non-preemptive LCFS order. (The relative service order of customers in $\mathcal{C}^g$ is as follows: $\boldsymbol{c}'_0$ is the (relatively) first customer to be served, $\boldsymbol{c}'_{N^g}$ is the (relatively) second customer, $\cdots$, and $\boldsymbol{c}'_1$ is the (relatively) final customer.) Other customers are served in the same algorithm as SA_P. Let $B_L^{(k,g)}$ be a length of time under $SA_L^{(k,g)}$ lasting from time τ until the first epoch after τ when customers in $\mathcal{C}^g$ and $\boldsymbol{c}(k-1)$-customers are cleared from the system. Its expected value conditioned on the system state $\boldsymbol{Y}$ at τ is denoted by $\overline{B_L^{(k,g)}(\boldsymbol{Y})}$.

Since scheduling algorithm $SA_L^{(k,g)}$ is also work conserving, the total amount of remaining service times of $\boldsymbol{c}(k-1)$-customers plus remaining service times of customers in $\mathcal{C}^g$ under SA_P is equal to that under $SA_L^{(k,g)}$. Let $\overline{s_l}'$ be the expected value of $\boldsymbol{c}'_l$'s service time $(l=0,1,\cdots,N^g)$. Then we have

$$\begin{aligned}\overline{B_P^{(k,g)}(\boldsymbol{Y})} &= \overline{B_L^{(k,g)}(\boldsymbol{Y})} = \sum_{l=0}^{N^g} \frac{\overline{s_l}'}{1-\rho_{k-1}^+} \\ &= \frac{r + \sum_{i=1}^{J}\sum_{\alpha=1}^{L_i} g_{i\alpha}E[S_{i\alpha}] + \sum_{i=1}^{k-1}\sum_{\alpha=1}^{L_i} n_{i\alpha}E[S_{i\alpha}]}{1-\rho_{k-1}^+}. \end{aligned} \tag{3.2}$$

4. Values of the Cost Functions at Arbitrary States

In this section, we obtain values of the cost functions defined in Section 2. Let us consider the system with any scheduling algorithm defined in the section. Let e $(=1,2,\cdots)$ be a customer number, and we consider the situation that $\boldsymbol{c}^e$ arrives at the system in state $\boldsymbol{Y}=(j,\beta,\kappa,a,r,\boldsymbol{g},\boldsymbol{n},L)\in\mathcal{E}$. Since $W_{i\alpha}(\boldsymbol{Y},e)=0$ and $G_{i\alpha}(\boldsymbol{Y},e)=0$ for $(i,\alpha)\neq(j,\beta)$, we consider the case that $(i,\alpha)=(j,\beta)\in\mathcal{S}$.

1-limited groups.

We consider the 1-limited group j that adopts the FCFS discipline $(j\in\mathcal{H}_{1F})$. From the analysis of the last section, we have

$$W_{j\beta}(\boldsymbol{Y},e) = \overline{B_P^{(j,1F)}(\boldsymbol{Y})} = \frac{r + \sum_{i=1}^{J}\sum_{\alpha=1}^{L_i} g_{i\alpha}E[S_{i\alpha}] + \sum_{i=1}^{j}\sum_{\alpha=1}^{L_i} n_{i\alpha}E[S_{i\alpha}]}{1-\rho_{j-1}^+}. \tag{4.1}$$

For the discipline, the customer immediately receives his service when his service period begins. Hence we have

$$G_{j\beta}(\boldsymbol{Y},e)=0. \tag{4.2}$$

Gated groups.

We consider the gated group j $(j\in\mathcal{H}_{gF}\cup\mathcal{H}_{gP})$. $W_{j\beta}(\cdot)$ has the same expression for any service order adopted by the group. From the analysis of the last section, we have

$$W_{j\beta}(\boldsymbol{Y},e) = \overline{B_P^{(j,g)}(\boldsymbol{Y})} = \frac{r + \sum_{i=1}^{J}\sum_{\alpha=1}^{L_i} g_{i\alpha}E[S_{i\alpha}] + \sum_{i=1}^{j-1}\sum_{\alpha=1}^{L_i} n_{i\alpha}E[S_{i\alpha}]}{1-\rho_{j-1}^+}. \tag{4.3}$$

The expected waiting time $G_{j\beta}(\cdot)$ of $\boldsymbol{c}^e$ is different between service orders adopted by the group. First we consider the group that adopts the FCFS order. His waiting time $G_{j\beta}^e$ is a time to complete services of customers who are already in the group at his arrival epoch. Then

$$G_{j\beta}(\boldsymbol{Y},e) = \sum_{\alpha=1}^{L_j} n_{j\alpha}E[S_{j\alpha}], \tag{4.4}$$

for $j \in \mathcal{H}_{gF}$.

Next we consider the group that adopts the fixed priority order. Since (j,β)-customers are served in first come first served order within their class, the waiting time $G^e_{j\beta}$ of $\boldsymbol{c}^e$ is a time to complete services of customers who are already in classes $1,\cdots,\beta-1$ at station j at the beginning epoch of his service period and services of (j,β)-customers who are already in station j at his arrival epoch. Then

$$G_{j\beta}(\boldsymbol{Y},e) = \sum_{\alpha=1}^{\beta-1}(n_{j\alpha}+\lambda_{j\alpha}W_{j\beta}(\boldsymbol{Y},e))E[S_{j\alpha}] + n_{j\beta}E[S_{j\beta}] \tag{4.5}$$

$$= \sum_{\alpha=1}^{\beta} n_{j\alpha}E[S_{j\alpha}] + \left(\sum_{\alpha=1}^{\beta-1}\lambda_{j\alpha}E[S_{j\alpha}]\right)\frac{r+\sum_{i=1}^{J}\sum_{\alpha=1}^{L_i} g_{i\alpha}E[S_{i\alpha}] + \sum_{i=1}^{j-1}\sum_{\alpha=1}^{L_i} n_{i\alpha}E[S_{i\alpha}]}{1-\rho^{+}_{j-1}},$$

for $j \in \mathcal{H}_{gP}$.

Common structures of the expressions.

The cost functions derived above are shown to be linear combinations of $(r,\boldsymbol{g},\boldsymbol{n})$ of state $\boldsymbol{Y}=(j,\beta,\kappa,a,r,\boldsymbol{g},\boldsymbol{n},L)\in\mathcal{E}$ at the $\boldsymbol{c}^e$'s arrival epoch. Then by appropriately choosing the coefficients, we can express as

$$W_{j\beta}(\boldsymbol{Y},e) = \varphi_{j\beta}r + \sum_{i=1}^{J}\sum_{\alpha=1}^{L_i}\psi_{i\alpha,j\beta}g_{i\alpha} + \sum_{i=1}^{J}\sum_{\alpha=1}^{L_i}\phi_{i\alpha,j\beta}n_{i\alpha}, \quad (j,\beta)\in\mathcal{S}, \tag{4.6}$$

$$G_{j\beta}(\boldsymbol{Y},e) = \eta_{j\beta}r + \sum_{i=1}^{J}\sum_{\alpha=1}^{L_i}\theta_{i\alpha,j\beta}g_{i\alpha} + \sum_{i=1}^{J}\sum_{\alpha=1}^{L_i}\zeta_{i\alpha,j\beta}n_{i\alpha}, \quad (j,\beta)\in\mathcal{S}. \tag{4.7}$$

Of course, every scheduling algorithm has its own coefficients. For simplicity, we define the following vectors:

$$\begin{aligned}
\boldsymbol{\phi}_{j\beta} &\equiv (\phi_{11,j\beta},\phi_{12,j\beta},\cdots,\phi_{JL_J,j\beta})' \in \mathcal{R}^{J_c\times 1},\\
\boldsymbol{\psi}_{j\beta} &\equiv (\psi_{11,j\beta},\psi_{12,j\beta},\cdots,\psi_{JL_J,j\beta})' \in \mathcal{R}^{J_c\times 1},\\
\boldsymbol{\zeta}_{j\beta} &\equiv (\zeta_{11,j\beta},\zeta_{12,j\beta},\cdots,\zeta_{JL_J,j\beta})' \in \mathcal{R}^{J_c\times 1},\\
\boldsymbol{\theta}_{j\beta} &\equiv (\theta_{11,j\beta},\theta_{12,j\beta},\cdots,\theta_{JL_J,j\beta})' \in \mathcal{R}^{J_c\times 1},
\end{aligned}$$

for $(j,\beta)\in\mathcal{S}$ where $'$ denotes a transposition of a vector.

Then the cost functions defined by (2.3) and (2.6) of the system operated under the given scheduling algorithm are given by

$$W_{i\alpha}(\boldsymbol{Y},e) = \begin{cases} r\varphi_{j\beta}+\boldsymbol{g}\boldsymbol{\psi}_{j\beta}+\boldsymbol{n}\boldsymbol{\phi}_{j\beta}, & (i,\alpha)=(j,\beta),\\ 0, & (i,\alpha)\neq(j,\beta),\end{cases} \tag{4.8}$$

$$G_{i\alpha}(\boldsymbol{Y},e) = \begin{cases} r\eta_{j\beta}+\boldsymbol{g}\boldsymbol{\theta}_{j\beta}+\boldsymbol{n}\boldsymbol{\zeta}_{j\beta}, & (i,\alpha)=(j,\beta),\\ 0, & (i,\alpha)\neq(j,\beta).\end{cases} \tag{4.9}$$

The important thing to consider about these expressions is that the component $(j,\beta,r,\boldsymbol{g},\boldsymbol{n})$ of state $\boldsymbol{Y}$ is sufficient to derive values of the cost functions. Further each function is linear with respect to $(r,\boldsymbol{g},\boldsymbol{n})$.

5. Steady State Values of the Cost Functions

We consider the system defined in Section 2. Although we have considered the system in arbitrary states, the system operated sufficiently long time may enter some steady state.

Now we evaluate steady state values of the waiting times. Finally, some numerical examples are provided.

Definitions and assumptions.

Now we define

$$\bar{w}_{i\alpha} \equiv \lim_{N\to\infty} \frac{\sum_{e=1}^{N}(W_{i\alpha}^{e}+G_{i\alpha}^{e})}{\sum_{e=1}^{N}\mathbf{1}\{(X(\sigma^{e}),\Gamma(\sigma^{e}))=(i,\alpha)\}}, \quad (i,\alpha)\in\mathcal{S}. \tag{5.1}$$

$\bar{w}_{i\alpha}$ in (5.1) denotes the *average waiting time* that (i,α)-customers spend in the system until their services begin, since $W_{i\alpha}^{e}=0$ and $G_{i\alpha}^{e}=0$ for $(X(\sigma^{e}),\Gamma(\sigma^{e}))\neq(i,\alpha)$. Further we define

$$\bar{W}_{i\alpha} \equiv \lim_{N\to\infty}\frac{1}{N}\sum_{e=1}^{N}W_{i\alpha}^{e}, \quad (i,\alpha)\in\mathcal{S}, \tag{5.2}$$

$$\bar{G}_{i\alpha} \equiv \lim_{N\to\infty}\frac{1}{N}\sum_{e=1}^{N}G_{i\alpha}^{e}, \quad (i,\alpha)\in\mathcal{S}, \tag{5.3}$$

if these limits may exist. $\bar{W}_{i\alpha}$ and $\bar{G}_{i\alpha}$ in (5.2) and (5.3) are introduced for convenience, and they may not be the average waiting times of (i,α)-customers.

The customer average values $\bar{\boldsymbol{g}}_i \equiv (\bar{g}_{i\alpha} : \alpha=1,\cdots\cdots,L_i)$, $\bar{\boldsymbol{n}}_i \equiv (\bar{n}_{i\alpha} : \alpha=1,\cdots\cdots,L_i)$, $\bar{\boldsymbol{g}} \equiv (\bar{\boldsymbol{g}}_1,\cdots,\bar{\boldsymbol{g}}_J)$, $\bar{\boldsymbol{n}} \equiv (\bar{\boldsymbol{n}}_1,\cdots,\bar{\boldsymbol{n}}_J)$, and $\bar{\boldsymbol{Y}} \equiv (\bar{X},\bar{\Gamma},\bar{\kappa},\bar{a},\bar{r},\bar{\boldsymbol{g}},\bar{\boldsymbol{n}},\bar{L})$ of the state at arrival epochs of customers are defined by

$$\bar{\boldsymbol{Y}} \equiv \lim_{N\to\infty}\frac{1}{N}\sum_{e=1}^{N}\boldsymbol{Y}(\sigma^{e}). \tag{5.4}$$

The time average values $\tilde{\boldsymbol{g}}_i \equiv (\tilde{g}_{i\alpha} : \alpha=1,\cdots\cdots,L_i)$, $\tilde{\boldsymbol{n}}_i \equiv (\tilde{n}_{i\alpha} : \alpha=1,\cdots,L_i)$, $\tilde{\boldsymbol{g}} \equiv (\tilde{\boldsymbol{g}}_1,\cdots,\tilde{\boldsymbol{g}}_J)$, $\tilde{\boldsymbol{n}} \equiv (\tilde{\boldsymbol{n}}_1,\cdots,\tilde{\boldsymbol{n}}_J)$, and $\tilde{\boldsymbol{Y}} \equiv (\tilde{X},\tilde{\Gamma},\tilde{\kappa},\tilde{a},\tilde{r},\tilde{\boldsymbol{g}},\tilde{\boldsymbol{n}},\tilde{L})$ of the state of the system are defined by

$$\tilde{\boldsymbol{Y}} \equiv \lim_{t\to\infty}\frac{1}{t}\int_{0}^{t}\boldsymbol{Y}(s)ds. \tag{5.5}$$

We state the following three assumptions:

[A-1] The process $\mathcal{Q}$ is regenerative [19], and the system is initially empty.

[A-2] $E[N_B]<\infty$ where N_B is the number of customers served during a regenerative cycle.

[A-3] The customer average value (5.4) exists, and

$$E[\sum_{e=1}^{N_B}r(\sigma^{e})]<\infty, \quad E[\sum_{e=1}^{N_B}\boldsymbol{g}(\sigma^{e})]<\infty, \quad \text{and} \quad E[\sum_{e=1}^{N_B}\boldsymbol{n}(\sigma^{e})]<\infty.$$

Derivation of a matrix equation on the average numbers of customers.

Note that the quantities $\bar{W}_{i\alpha}(\bar{G}_{i\alpha})$ include $W_{i\alpha}^{e}(G_{i\alpha}^{e})$ for customers $\boldsymbol{c}^{e}$ who are not (i,α)-customers. The proportion of arriving customers who really become (i,α)-customers is equal to $\lambda_{i\alpha}/\lambda$. Then we can get the following representations concerning with the steady state values of the cost functions. From (4.8), (5.2) and (5.4), we can show that

$$\bar{W}_{i\alpha} = (\lambda_{i\alpha}/\lambda)\{\bar{r}\varphi_{i\alpha}+\bar{\boldsymbol{g}}\boldsymbol{\psi}_{i\alpha}+\bar{\boldsymbol{n}}\boldsymbol{\phi}_{i\alpha}\}, \quad (i,\alpha)\in\mathcal{S}. \tag{5.6}$$

Further, from (4.9), (5.3) and (5.4), we can show that

$$\bar{G}_{i\alpha} = (\lambda_{i\alpha}/\lambda)\{\bar{r}\eta_{i\alpha}+\bar{\boldsymbol{g}}\boldsymbol{\theta}_{i\alpha}+\bar{\boldsymbol{n}}\boldsymbol{\zeta}_{i\alpha}\}, \quad (i,\alpha)\in\mathcal{S}. \tag{5.7}$$

The steady state average value $\tilde{r}$ of remaining service times of customers being served is given by

$$\tilde{r} = \sum_{i=1}^{J} \sum_{\alpha=1}^{L_i} \frac{\lambda_{i\alpha} \overline{s^2}_{i\alpha}}{2}. \tag{5.8}$$

We use the generalized Little's formula ($H = \lambda G$) [7, 24, 25] that equates the time average values of the costs with the customer average values of the costs to obtain

$$\tilde{n}_{i\alpha} = \lambda \bar{W}_{i\alpha}, \quad (i,\alpha) \in \mathcal{S}, \tag{5.9}$$
$$\tilde{g}_{i\alpha} = \lambda \bar{G}_{i\alpha}, \quad (i,\alpha) \in \mathcal{S}. \tag{5.10}$$

From (5.6) and (5.9), we have

$$\tilde{n}_{i\alpha} = \lambda_{i\alpha}\{\bar{r}\varphi_{i\alpha} + \bar{\boldsymbol{g}}\boldsymbol{\psi}_{i\alpha} + \bar{\boldsymbol{n}}\boldsymbol{\phi}_{i\alpha}\}, \quad (i,\alpha) \in \mathcal{S}. \tag{5.11}$$

From (5.7) and (5.10), we have

$$\tilde{g}_{i\alpha} = \lambda_{i\alpha}\{\bar{r}\eta_{i\alpha} + \bar{\boldsymbol{g}}\boldsymbol{\theta}_{i\alpha} + \bar{\boldsymbol{n}}\boldsymbol{\zeta}_{i\alpha}\}, \quad (i,\alpha) \in \mathcal{S}. \tag{5.12}$$

From the PASTA property [26, 27], we have

$$\tilde{n}_{i\alpha} = \lambda_{i\alpha}\{\tilde{r}\varphi_{i\alpha} + \tilde{\boldsymbol{g}}\boldsymbol{\psi}_{i\alpha} + \tilde{\boldsymbol{n}}\boldsymbol{\phi}_{i\alpha}\}, \quad (i,\alpha) \in \mathcal{S}, \tag{5.13}$$
$$\tilde{g}_{i\alpha} = \lambda_{i\alpha}\{\tilde{r}\eta_{i\alpha} + \tilde{\boldsymbol{g}}\boldsymbol{\theta}_{i\alpha} + \tilde{\boldsymbol{n}}\boldsymbol{\zeta}_{i\alpha}\}, \quad (i,\alpha) \in \mathcal{S}. \tag{5.14}$$

For the notational simplicity, we define the following vectors and matrices.

$$\begin{aligned}
&\boldsymbol{s} \equiv (\varphi_{i\alpha} : (i,\alpha) \in \mathcal{S}) \in \mathcal{R}^{1\times J_c}, \quad \boldsymbol{s}_g \equiv (\eta_{i\alpha} : (i,\alpha) \in \mathcal{S}) \in \mathcal{R}^{1\times J_c},\\
&\boldsymbol{S}_1 \equiv (\boldsymbol{\phi}_{11}, \boldsymbol{\phi}_{12}, \cdots, \boldsymbol{\phi}_{JL_J}) \in \mathcal{R}^{J_c\times J_c}, \quad \boldsymbol{S}_2 \equiv (\boldsymbol{\zeta}_{11}, \boldsymbol{\zeta}_{12}, \cdots, \boldsymbol{\zeta}_{JL_J}) \in \mathcal{R}^{J_c\times J_c},\\
&\boldsymbol{R}_1 \equiv (\boldsymbol{\psi}_{11}, \boldsymbol{\psi}_{12}, \cdots, \boldsymbol{\psi}_{JL_J}) \in \mathcal{R}^{J_c\times J_c}, \quad \boldsymbol{R}_2 \equiv (\boldsymbol{\theta}_{11}, \boldsymbol{\theta}_{12}, \cdots, \boldsymbol{\theta}_{JL_J}) \in \mathcal{R}^{J_c\times J_c},\\
&\boldsymbol{\Lambda} \equiv \mathrm{diag}(\lambda_{i\alpha} : (i,\alpha) \in \mathcal{S}) \in \mathcal{R}^{J_c\times J_c},
\end{aligned} \tag{5.15}$$

where $\varphi_{i\alpha}, \eta_{i\alpha}, \boldsymbol{\phi}_{i\alpha}, \boldsymbol{\zeta}_{i\alpha}, \boldsymbol{\psi}_{i\alpha}$ and $\boldsymbol{\theta}_{i\alpha}$ ($(i,\alpha) \in \mathcal{S}$) are defined in the last section (eqs. (4.8) and (4.9)). Then we arrive at the equations that determine steady state average values of the components of the system state

$$\tilde{\boldsymbol{n}} = \{\tilde{r}\boldsymbol{s} + \tilde{\boldsymbol{g}}\boldsymbol{R}_1 + \tilde{\boldsymbol{n}}\boldsymbol{S}_1\}\boldsymbol{\Lambda}, \tag{5.16}$$
$$\tilde{\boldsymbol{g}} = \{\tilde{r}\boldsymbol{s}_g + \tilde{\boldsymbol{g}}\boldsymbol{R}_2 + \tilde{\boldsymbol{n}}\boldsymbol{S}_2\}\boldsymbol{\Lambda}. \tag{5.17}$$

Or equivalently, we have

$$(\tilde{\boldsymbol{n}}, \tilde{\boldsymbol{g}}) = \left\{\tilde{r}(\boldsymbol{s}, \boldsymbol{s}_g) + (\tilde{\boldsymbol{n}}, \tilde{\boldsymbol{g}}) \begin{pmatrix} \boldsymbol{S}_1 & \boldsymbol{S}_2 \\ \boldsymbol{R}_1 & \boldsymbol{R}_2 \end{pmatrix}\right\} \begin{pmatrix} \boldsymbol{\Lambda} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Lambda} \end{pmatrix}, \tag{5.18}$$

where $\boldsymbol{O} \in \mathcal{R}^{J_c\times J_c}$ is a zero matrix.

The average numbers of customers and the average waiting times.

From the above analysis, we have an expression for the vector $(\tilde{\boldsymbol{n}}, \tilde{\boldsymbol{g}})$ of the average numbers of customers as follows:

$$(\tilde{\boldsymbol{n}}, \tilde{\boldsymbol{g}}) = \tilde{r}(\boldsymbol{s}, \boldsymbol{s}_g)\left\{\begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Lambda}^{-1} \end{pmatrix} - \begin{pmatrix} \boldsymbol{S}_1 & \boldsymbol{S}_2 \\ \boldsymbol{R}_1 & \boldsymbol{R}_2 \end{pmatrix}\right\}^{-1}, \tag{5.19}$$

where $\tilde{r}$ is defined in (5.8) and where the other constants (vectors and matrices) in the right-hand side are defined by the expressions in (5.15). The existence of the above inverse matrix is shown in Appendix. Finally, we can get the vector $\bar{\boldsymbol{w}} \equiv (\bar{w}_{i\alpha} : (i,\alpha) \in \mathcal{S})$ of the average waiting times

$$\bar{\boldsymbol{w}} = (\tilde{\boldsymbol{n}} + \tilde{\boldsymbol{g}})\boldsymbol{\Lambda}^{-1}. \tag{5.20}$$

The average waiting times of all customers belonging to group i are given by

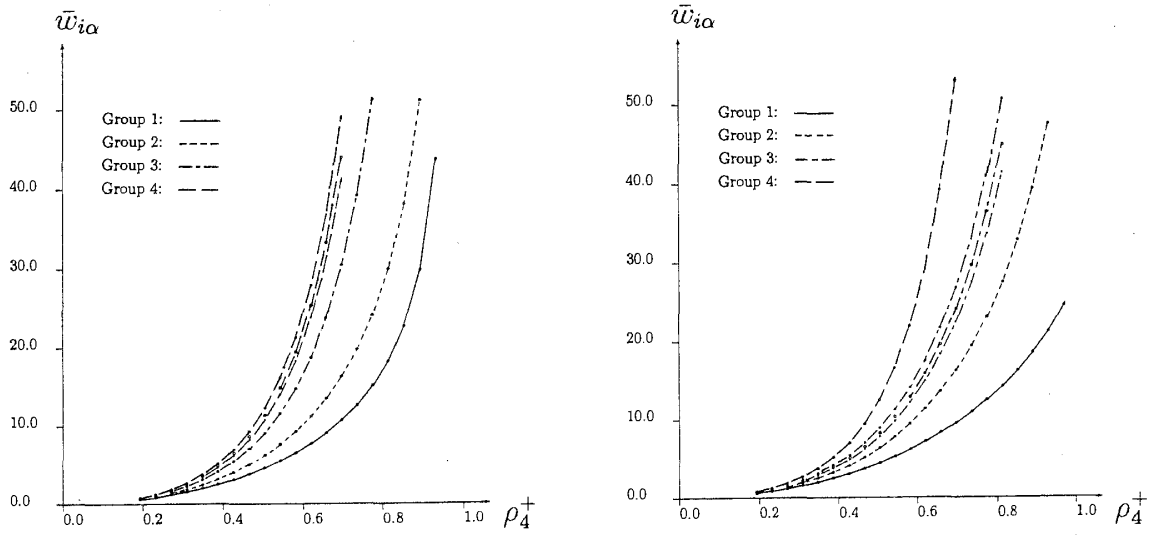
$$\bar{w}_i = \sum_{\alpha=1}^{L_i} \frac{\lambda_{i\alpha}}{\lambda_i} \bar{w}_{i\alpha}, \quad i = 1, \cdots, J. \tag{5.21}$$

The average waiting time of all customers is given by

$$\bar{w} = \sum_{i=1}^{J} \sum_{\alpha=1}^{L_i} \frac{\lambda_{i\alpha}}{\lambda} \bar{w}_{i\alpha}. \tag{5.22}$$

Remark. From (5.19), we can see that the average numbers of customers in the waiting rooms and customers in the service facility are 'linear' in the remaining service time $\tilde{r}$. Hence from (5.20) if $\tilde{r}$ is increased twice while keeping the arrival rates and the mean service times constants, all average waiting times are also increased twice. The property is often observed for M/G/1 system. For example, see the P-K mean value formula and the Cobham's formula for a nonpreemptive HOL [4, 13].

Numerical examples and graphs.



(a) Model 1. (b) Model 2.

Figure 1. Mean waiting times.

Now we calculate the steady state performance measures for two models described below.

- Model 1.
  1. Group 1 (with 3 classes): 1-limited group with the FCFS discipline.
  2. Group 2 (with 3 classes): Gated group with the FCFS order.
  3. Group 3 (with 3 classes): 1-limited group with the FCFS discipline.
  4. Group 4 (with 3 classes): Gated group with the priority order.
- Model 2.

1. Group 1 (with 3 classes): Gated group with the FCFS order.
2. Group 2 (with 3 classes): 1-limited group with the FCFS discipline.
3. Group 3 (with 3 classes): Gated group with the priority order.
4. Group 4 (with 3 classes): 1-limited group with the FCFS discipline.

In order to compare a difference between the two scheduling algorithms, we use the same set of parameters (Table 1) for calculating the performance measures of these two models. All service times have 5 stage Erlang distributions. The mean service times are varied for plotting the average waiting times with different values of ρ_4^+.

Table 1. Set of parameters.

	Arrival rates			Mean service times		
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Group 1	1/100.0	1/150.0	1/200.0	1.5 ~	3.0 ~	4.5 ~
Group 2	1/100.0	1/150.0	1/200.0	1.5 ~	3.0 ~	4.5 ~
Group 3	1/180.0	1/360.0	1/540.0	1.0 ~	5.5 ~	10.0 ~
Group 4	1/180.0	1/360.0	1/540.0	1.0 ~	5.5 ~	10.0 ~

For example, we consider a packet switching local area network where 3 users (classes) who transmit packets with 4 priority groups are connected through a switch. For a gated group, all packets of all users are admitted into the service facility at a time. Further if packets in the service facility belong to a gated group with the priority order, user 1's packets have the highest priority and user 3's packets have the lowest priority. Since the average waiting times of packets in each 1-limited group with the FCFS discipline or in each gated group with the FCFS order are identical, their graphs in Figure 1 are overlapped. If we would like to minimize the waiting times of packets with priority, we will adopt Model 2. On the other hand, if we would like to give a prioritized service maintaining certain degree of fairness, we will adopt Model 1.

6. Conclusions

We have investigated the multiclass M/G/1 queues with a mixture of the 1-limited disciplines and the gated disciplines. The average waiting times of customers and the average numbers of customers are obtained.

Our approach takes four steps to obtain steady state values of these systems performance measures. First, we define the system states and the stochastic process associated with them, and then define the system performance measures as cost functions (conditional expectations) of the system states. The cost functions $W_{i\alpha}(\cdot, e)$ and $G_{i\alpha}(\cdot, e)$ $((i,\alpha) \in \mathcal{S})$ denote the conditional expected waiting time of the e^{th} customer in the waiting room and his conditional expected waiting time in the service facility, respectively. Second, we analyze busy periods. Third, from their analysis we derive the expressions of the cost functions for every station-class pair and for every service discipline. The important things to consider about these expressions are that the component $(j, \beta, r, \boldsymbol{g}, \boldsymbol{n})$ of system state $\boldsymbol{Y} = (j, \beta, \kappa, a, r, \boldsymbol{g}, \boldsymbol{n}, L) \in \mathcal{E}$ is sufficient to derive these values and that each function is linear with respect to $(r, \boldsymbol{g}, \boldsymbol{n})$. Finally, we evaluate their steady state values. Since these values are expressed in matrix forms, an algorithm for yielding their actual values can be easily constructed.

Acknowledgement The author would like to thank the referees for their valuable comments and suggestions.

Appendix

In this appendix, we prove the existence of the inverse matrix in (5.19). The matrix in (5.19) is converted as

$$\begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Lambda}^{-1} \end{pmatrix} - \begin{pmatrix} \boldsymbol{S}_1 & \boldsymbol{S}_2 \\ \boldsymbol{R}_1 & \boldsymbol{R}_2 \end{pmatrix} = \left\{ \begin{pmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} \end{pmatrix} - \begin{pmatrix} \boldsymbol{S}_1\boldsymbol{\Lambda} & \boldsymbol{S}_2\boldsymbol{\Lambda} \\ \boldsymbol{R}_1\boldsymbol{\Lambda} & \boldsymbol{R}_2\boldsymbol{\Lambda} \end{pmatrix} \right\} \begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Lambda}^{-1} \end{pmatrix}. \tag{A.1}$$

where $\boldsymbol{I} \in \mathcal{R}^{J_c \times J_c}$ is an identity matrix. Then we show the following lemma.

Lemma. The inverse matrix of the matrix

$$\begin{pmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} \end{pmatrix} - \begin{pmatrix} \boldsymbol{S}_1\boldsymbol{\Lambda} & \boldsymbol{S}_2\boldsymbol{\Lambda} \\ \boldsymbol{R}_1\boldsymbol{\Lambda} & \boldsymbol{R}_2\boldsymbol{\Lambda} \end{pmatrix}$$

in (A.1) exists.

Proof. Let

$$Q \equiv \begin{pmatrix} \boldsymbol{S}_1\boldsymbol{\Lambda} & \boldsymbol{S}_2\boldsymbol{\Lambda} \\ \boldsymbol{R}_1\boldsymbol{\Lambda} & \boldsymbol{R}_2\boldsymbol{\Lambda} \end{pmatrix}.$$

From the definition, components of Q are given by

$$\begin{aligned}
(\boldsymbol{S}_1\boldsymbol{\Lambda})_{i\alpha,j\beta} &= \phi_{i\alpha,j\beta}\lambda_{j\beta} = b_{i\alpha}e^{(0)}_{ij}c_{j\beta}, \\
(\boldsymbol{R}_1\boldsymbol{\Lambda})_{i\alpha,j\beta} &= \psi_{i\alpha,j\beta}\lambda_{j\beta} = b_{i\alpha}c_{j\beta}, \\
(\boldsymbol{S}_2\boldsymbol{\Lambda})_{i\alpha,j\beta} &= \zeta_{i\alpha,j\beta}\lambda_{j\beta} = b_{i\alpha}e^{(1)}_{ij}d_{j\beta} + b_{i\alpha}e^{(2)}_{i\alpha,j\beta}\lambda_{j\beta}, \\
(\boldsymbol{R}_2\boldsymbol{\Lambda})_{i\alpha,j\beta} &= \theta_{i\alpha,j\beta}\lambda_{j\beta} = b_{i\alpha}d_{j\beta},
\end{aligned}$$

$((i,\alpha),(j,\beta) \in \mathcal{S})$ where

$$\begin{aligned}
b_{i\alpha} &\equiv E[S_{i\alpha}], \\
c_{j\beta} &\equiv \lambda_{j\beta}/(1-\rho^{+}_{j-1}), \\
d_{j\beta} &\equiv \begin{cases} (\sum_{\alpha=1}^{\beta-1}\lambda_{j\alpha}E[S_{j\alpha}])\lambda_{j\beta}/(1-\rho^{+}_{j-1}), & j \in \mathcal{H}_{gP} \\ 0, & j \in \mathcal{H}_{1F} \cup \mathcal{H}_{gF}, \end{cases} \\
e^{(0)}_{ij} &\equiv \begin{cases} 1, & i \le j-1, \\ 0, & i > j-1, \end{cases} \quad j \in \mathcal{H}_{gF} \cup \mathcal{H}_{gP}, \\
&\equiv \begin{cases} 1, & i \le j, \\ 0, & i > j, \end{cases} \quad j \in \mathcal{H}_{1F}, \\
e^{(1)}_{ij} &\equiv \begin{cases} 1, & i \le j-1, \\ 0, & i > j-1, \end{cases} \\
e^{(2)}_{i\alpha,j\beta} &\equiv \begin{cases} 1, & i = j \text{ and } \alpha \le \beta, \\ 0, & \text{otherwise}, \end{cases} \quad j \in \mathcal{H}_{gP}, \\
&\equiv \begin{cases} 1, & i = j, \\ 0, & i \ne j, \end{cases} \quad j \in \mathcal{H}_{gF}, \\
&\equiv 0, \quad j \in \mathcal{H}_{1F}.
\end{aligned}$$

Now we consider any vector $\boldsymbol{x} = ((\boldsymbol{x}^{(1)})', (\boldsymbol{x}^{(2)})')' \in \mathcal{R}^{2J_c \times 1}$ that satisfies the equation:

$$(I - Q)\boldsymbol{x} = \mathbf{o},$$

where $I \in \mathcal{R}^{2J_c \times 2J_c}$ is an identity matrix. Each component of the equation is given by

$$x_{i\alpha}^{(1)} = \sum_k \sum_\gamma b_{i\alpha} e_{ik}^{(0)} c_{k\gamma} x_{k\gamma}^{(1)} + \sum_k \sum_\gamma (b_{i\alpha} e_{ik}^{(1)} d_{k\gamma} + b_{i\alpha} e_{i\alpha,k\gamma}^{(2)} \lambda_{k\gamma}) x_{k\gamma}^{(2)}, \quad (i,\alpha) \in \mathcal{S}, \tag{A.2}$$

$$x_{i\alpha}^{(2)} = \sum_k \sum_\gamma b_{i\alpha} c_{k\gamma} x_{k\gamma}^{(1)} + \sum_k \sum_\gamma b_{i\alpha} d_{k\gamma} x_{k\gamma}^{(2)}, \quad (i,\alpha) \in \mathcal{S}. \tag{A.3}$$

Now we define

$$y_k^{(1)} \equiv \sum_{\gamma=1}^{L_k} c_{k\gamma} x_{k\gamma}^{(1)}, \quad y_k^{(2)} \equiv \sum_{\gamma=1}^{L_k} d_{k\gamma} x_{k\gamma}^{(2)}, \quad k = 1, \cdots, J,$$

$$C \equiv \sum_{k=1}^{J} (y_k^{(1)} + y_k^{(2)}).$$

Then from (A.3) and the definition of $y_k^{(2)}$ and C, we have

$$x_{i\alpha}^{(2)} = b_{i\alpha} C, \quad (i,\alpha) \in \mathcal{S}, \tag{A.4}$$

$$y_k^{(2)} = \xi_k C, \quad k = 1, \cdots, J, \tag{A.5}$$

where $\xi_k \equiv \sum_{\gamma=1}^{L_k} d_{k\gamma} b_{k\gamma}$ $(k = 1, \cdots, J)$. By substituting them into (A.2), we have

$$x_{i\alpha}^{(1)} = b_{i\alpha} \left\{ \sum_{k=1}^{J} e_{ik}^{(0)} y_k^{(1)} + \left(\sum_{k=1}^{J} e_{ik}^{(1)} \xi_k + \sum_{k=1}^{J} \sum_\gamma e_{i\alpha,k\gamma}^{(2)} \lambda_{k\gamma} b_{k\gamma} \right) C \right\}. \tag{A.6}$$

Let

$$\nu_i \equiv \sum_{\alpha=1}^{L_i} c_{i\alpha} b_{i\alpha}, \qquad i = 1, \cdots, J,$$

$$\chi_i \equiv \sum_{\alpha=1}^{L_i} c_{i\alpha} b_{i\alpha} \left(\sum_{k=1}^{J} e_{ik}^{(1)} \xi_k + \sum_{k=1}^{J} \sum_{\gamma=1}^{L_k} e_{i\alpha,k\gamma}^{(2)} \lambda_{k\gamma} b_{k\gamma} \right), \qquad i = 1, \cdots, J,$$

$$\xi^+ \equiv \sum_{k=1}^{J} \xi_k = \sum_{k=1}^{J} \sum_{\gamma=1}^{L_k} d_{k\gamma} b_{k\gamma}.$$

Then from the definition of $y_k^{(1)}$ and C, (A.5) and (A.6), we have a set of equations:

$$C = \sum_{k=1}^{J} y_k^{(1)} + \xi^+ C, \tag{A.7}$$

$$y_i^{(1)} = \sum_{k=1}^{J} \nu_i e_{ik}^{(0)} y_k^{(1)} + \chi_i C, \quad i = 1, \cdots, J. \tag{A.8}$$

Let

$$y \equiv \begin{pmatrix} y_1^{(1)} \\ y_2^{(1)} \\ \vdots \\ \vdots \\ y_{J-1}^{(1)} \\ y_J^{(1)} \\ C \end{pmatrix}, \quad Q_0 \equiv \begin{pmatrix} \nu_1 e_1 & \nu_1 & \nu_1 & \cdots & \nu_1 & \nu_1 & \chi_1 \\ 0 & \nu_2 e_2 & \nu_2 & \cdots & \nu_2 & \nu_2 & \chi_2 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & & & \ddots & \nu_{J-1} e_{J-1} & \nu_{J-1} & \chi_{J-1} \\ 0 & 0 & 0 & \cdots & 0 & \nu_J e_J & \chi_J \\ 1 & 1 & 1 & \cdots & 1 & 1 & \xi^+ \end{pmatrix},$$

where $e_j \equiv 1$ if $j \in \mathcal{H}_{1F}$, or $e_j \equiv 0$ if $j \in \mathcal{H}_{gF} \cup \mathcal{H}_{gP}$. Then we have

$$\boldsymbol{y} = Q_0 \boldsymbol{y}. \tag{A.9}$$

Let

$$U \equiv \begin{pmatrix} 1-\nu_1 e_1 & -\nu_1 & -\nu_1 & \cdots & -\nu_1 & -\nu_1 & -\chi_1 \\ 0 & 1-\nu_2 e_2 & -\nu_2 & \cdots & -\nu_2 & -\nu_2 & -\chi_2 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & & & \ddots & 1-\nu_{J-1}e_{J-1} & -\nu_{J-1} & -\chi_{J-1} \\ 0 & 0 & 0 & \cdots & 0 & 1-\nu_J e_J & -\chi_J \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1-\xi^+ \end{pmatrix},$$

$$L \equiv \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & & & \ddots & 1 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ \ell_1 & \ell_2 & \ell_3 & \cdots & \ell_{J-1} & \ell_J & \ell_{J+1} \end{pmatrix},$$

where $(\ell_1, \ell_2, \cdots, \ell_J, \ell_{J+1})$ satisfies the equation:

$$(\ell_1, \ell_2, \cdots, \ell_J, \ell_{J+1})U = (-1, -1, \cdots, -1, 1-\xi^+). \tag{A.10}$$

Then it can be easily shown that

$$LU = I - Q_0,$$

where $I \in \mathcal{R}^{(J+1)\times(J+1)}$ is an identity matrix. Further it can be shown that

$$\det(I - Q_0) = \det(L)\det(U) = \ell_{J+1}\left(\prod_{j=1}^{J}(1-\nu_j e_j)\right)(1-\xi^+).$$

By definition and $\rho_J^+ < 1, 1-\nu_j e_j > 0$ $(j = 1, \cdots, J)$ and $1-\xi^+ > 0$. Hence if we may show that $\ell_{J+1} > 0$, then $\det(I - Q_0) > 0$.

Now we consider equation (A.10). These elements are given by

$$-\sum_{j=1}^{i-1}\ell_j\nu_j + \ell_i(1-\nu_i e_i) = -1, \quad i = 1, \cdots, J, \tag{A.11}$$

$$-\sum_{j=1}^{J}\ell_j\chi_j + \ell_{J+1}(1-\xi^+) = 1-\xi^+. \tag{A.12}$$

Hence we have

$$\ell_i = \frac{-1 + \sum_{j=1}^{i-1}\ell_j\nu_j}{1-\nu_i e_i}, \quad i = 1, \cdots, J,$$

$$\ell_{J+1} = \frac{1-\xi^+ + \sum_{j=1}^{J}\ell_j\chi_j}{1-\xi^+}.$$

From the definition of χ_j and (A.11), it can be easily shown that

$$\begin{aligned}\sum_{j=1}^{J}\ell_j\chi_j &= \sum_{j=1}^{J}\ell_j\left(\nu_j\sum_{k=j+1}^{J}\xi_k+\sum_{\alpha}c_{j\alpha}b_{j\alpha}\tilde{\rho}_{j\alpha}\right)\\ &= \sum_{k=1}^{J}\xi_k\sum_{j=1}^{k-1}\ell_j\nu_j+\sum_{j=1}^{J}\ell_j\sum_{\alpha}c_{j\alpha}b_{j\alpha}\tilde{\rho}_{j\alpha} \qquad \left(\sum_{j=1}^{0}\ell_j\nu_j=0\right)\\ &= \xi^+ + \sum_{k=1}^{J}\ell_k\left(\xi_k(1-\nu_k e_k)+\sum_{\alpha}c_{k\alpha}b_{k\alpha}\tilde{\rho}_{k\alpha}\right),\end{aligned}$$

where

$$\tilde{\rho}_{k\alpha}\equiv\sum_{j=1}^{J}\sum_{\gamma=1}^{L_j}e^{(2)}_{k\alpha,j\gamma}\lambda_{j\gamma}b_{j\gamma}=\begin{cases}\sum_{\gamma=\alpha}^{L_k}\lambda_{k\gamma}b_{k\gamma}, & k\in\mathcal{H}_{gP},\\ \rho_k, & k\in\mathcal{H}_{gF},\\ 0, & k\in\mathcal{H}_{1F},\end{cases}\quad \alpha=1,\cdots,L_k.$$

Further

$$\xi_k(1-\nu_k e_k)+\sum_{\alpha}c_{k\alpha}b_{k\alpha}\tilde{\rho}_{k\alpha}=\begin{cases}\nu_k\rho_k, & k\in\mathcal{H}_{gP}\cup\mathcal{H}_{gF},\\ 0, & k\in\mathcal{H}_{1F}.\end{cases}$$

Then

$$\sum_{j=1}^{J}\ell_j\chi_j=\xi^+ + \sum_{k\notin\mathcal{H}_{1F}}\ell_k\rho_k\nu_k=\xi^+ + \sum_{k=1}^{J}(1-e_k)\ell_k\rho_k\nu_k.$$

Hence we have

$$\ell_{J+1}=\frac{1}{1-\xi^+}\left(1+\sum_{k=1}^{J}\ell_k(1-e_k)\nu_k\rho_k\right).$$

Further from equation (A.11), we have

$$-\rho_k\sum_{j=1}^{k-1}\ell_j\nu_j+\rho_k\ell_k(1-\nu_k e_k)=-\rho_k,\quad k=1,\cdots,J. \tag{A.13}$$

Then we have

$$-\sum_{k=1}^{J}\rho_k\sum_{j=1}^{k}\ell_j\nu_j+\sum_{k=1}^{J}(1-e_k)\ell_k\nu_k\rho_k+\sum_{k=1}^{J}\ell_k\rho_k=-\rho_J^+.$$

After some calculations, we have

$$1+\sum_{k=1}^{J}(1-e_k)\ell_k\rho_k\nu_k=(1-\rho_J^+)\left\{1-\sum_{j=1}^{J}\ell_j\nu_j\right\}.$$

Because it can be easily shown that $\ell_i\le -1$ $(i=1,\cdots,J)$, we have

$$1-\sum_{j=1}^{J}\ell_j\nu_j=-\ell_J\nu_J-\ell_J(1-\nu_J e_J)>0.$$

Hence we have

$$\ell_{J+1}=\frac{(1-\rho_J^+)\{1-\sum_{j=1}^{J}\ell_j\nu_j\}}{1-\xi^+}>0.$$

Since we have shown that $\det(I-Q_0)>0, (I-Q_0)^{-1}$ exists and $\boldsymbol{y}=\mathbf{0}$ in (A.9). From (A.4) and (A.6), we have

$$\boldsymbol{x}^{(1)}=\mathbf{0},\qquad \boldsymbol{x}^{(2)}=\mathbf{0}.$$

Hence $(I-Q)^{-1}$ exists. □

References

[1] B. Avi-Itzhak and S. Halfin: Response times in gated M/G/1 queues: The processor-sharing case. *Queueing Systems*, **4** (1989) 263–279.

[2] J.S.-C. Chen, R. Guérin and T.E. Stern: Markov-modulated flow model for the output queues of a packet switch. *IEEE Transactions on Communications*, **40** (1992) 1098–1110.

[3] J.S.-C. Chen and R. Guérin: Performance study of an input queueing packet switch with two priority classes. *IEEE Transactions on Communications*, **39** (1991) 117–126.

[4] A. Cobham: Priority assignment in waiting line problems. *Operations Research*, **2** (1954) 70–76.

[5] M.J. Ferguson and Y.J. Aminetzah: Exact results for nonsymmetric token ring systems. *IEEE Transactions on Communications*, **33** (1985) 223–231.

[6] T.W. Gay and P.H. Seaman: Composite priority queue. *IBM Journal of Research and Development*, **19** (1975) 78–81.

[7] P.W. Glynn and W. Whitt: Extensions of the queueing relations $L = \lambda W$ and $H = \lambda G$. *Operations Research*, **37** (1989) 634–644.

[8] Hewlett Packard: *Network Design Guide: Designing HP Advance Stack Networks* (Hewlett Packard, 1998).

[9] T. Hirayama: Mean sojourn times of multiclass feedback queues with gated service disciplines. *Abstracts of the 1997 Fall National Conference of ORSJ* (Operations Research Society of Japan, 1997), 1-C-3.

[10] S.J. Hong, T. Hirayama and K. Yamada: The mean sojourn time of multiclass M/G/1 queues with feedback. *Asia-Pacific Journal of Operational Research*, **10** (1993) 233–249.

[11] N.K. Jaiswal: *Priority Queues* (Academic Press, New York, 1968).

[12] M.J. Karol, M.G. Hluchyj and S.P. Morgan: Input versus output queueing on a space-division packet switch. *IEEE Transactions on Communications*, **35** (1987) 1347–1356.

[13] L. Kleinrock: *Queueing Systems Vol. II: Computer Applications* (Wiley, New York, 1976).

[14] O. Kella and U. Yechiali: Priorities in M/G/1 queue with server vacations. *Naval Research Logistics*, **35** (1988) 23–34.

[15] H. Levy: Binomial-gated service: A method for effective operation and optimization of polling systems. *IEEE Transactions on Communications*, **39** (1991) 1341–1350.

[16] R.G. Miller, Jr.: Priority queues. *The Annals of Mathematical Statistics*, **31** (1960) 86–103.

[17] M. Paterok and M. Ettl: Sojourn time and waiting time distributions for M/GI/1 queues with preemption-distance priorities. *Operations Research*, **42** (1994) 1146–1161.

[18] K.M. Rege and B. Sengupta: A single server queue with gated processor-sharing discipline. *Queueing Systems*, **4** (1989) 249–261.

[19] S.M. Ross: *Applied Probability Models with Optimization Applications* (Holden-Day, San Francisco, 1970).

[20] J.G. Shanthikumar: Level crossing analysis of priority queues and a conservation identity for vacation models. *Naval Research Logistics*, **36** (1989) 797–806.

[21] H. Takagi: *Queueing Analysis: A Foundation of Performance Evaluation, Vol. 1: Vacation and Priority Systems, Part 1* (North-Holland, Amsterdam, 1991).

[22] Y. Takahashi and S. Shimogawa: Composite priority single-server queue with structured batch inputs. *Commun. Statist.-Stochastic Models*, **7** (1991) 481–497.

[23] 3 Com: Building multi-gigabit backbones: A new switching architecture for beyond 2000. *3 Com Technical Papers*, 500663-001 (1998).

[24] W. Whitt: A review of $L = \lambda W$ and extensions. *Queueing Systems*, **9** (1991) 235–268.

[25] W. Whitt: Correction note on $L = \lambda W$. *Queueing Systems*, **12** (1992) 431–432.

[26] R.W. Wolff: *Stochastic Modeling and the Theory of Queues* (Prentice-Hall, Englewood Cliffs, NJ, 1989).

[27] R.W. Wolff: Poisson arrivals see time averages. *Operations Research*, **30** (1982) 223–231.

Tetsuji Hirayama
Institute of Information Sciences and Electronics
University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-0006, Japan
E-mail: hirayama@is.tsukuba.ac.jp