Journal of the Operations Research
Society of Japan
Vol. 38, No. 4, December 1995

© 1995 The Operations Research Society of Japan

PSEUDO-CONSERVATION LAW FOR DISCRETE-TIME MULTI-QUEUE SYSTEMS WITH PRIORITY DISCIPLINES

Yoshitaka Takahashi B. Krishna Kumar
NTT Telecommunication Networks Laboratories

(Received January 10, 1994; Revised October 19, 1994)

Abstract We consider a discrete-time cyclic-service system consisting of multiple stations visited by a single server. Customers from several priority classes arrive at an individual station according to independent batch Bernoulli processes. We assume a non-preemptive priority rule and non-zero switch-over times of the server between consecutive stations. We derive an exact expression for a weighted sum of the mean waiting times for the individual priority classes: a pseudo-conservation law. Taking the limit of our result as the length tends to zero yields previously obtained continuous-time results.

1. Introduction

Multi-queue models served in cyclic order by a single server have been used to evaluate the performance of polling and token ring systems. Recently, the necessity for and importance of priority functions in a multi-queue model have been increasing. For example, a token ring system handling packetized and data traffic, where a voice packet has a higher (non-preemptive) priority than a data packet, reduces to a multi-queue priority model [10]. Motivated by this situation, we will treat a discrete-time multi-queue priority system and present an exact expression for a weighted sum of the mean waiting times for the individual priority classes, so-called *pseudo-conservation law*, that generalizes the previously obtained results [2, 3, 8, 17].

The pseudo-conservation law is shown to be important from both the practical and theoretical points of view, since it can be readily used to obtain (or test) the exact solutions and approximations for the mean waiting times in the individual queues in a multi-queue system [12, 17]. For continuous-time multi-queue priority systems, Fournier & Rosberg [8] and Shimogawa & Takahashi [17] have independently presented the pseudo-conservation law.

To the best of our knowledge, however, there are few literature on the pseudo-conservation laws for discrete-time multi-queue priority systems. In discrete-time systems, all events (e. g., arrivals and departures of customers, server-switches) are allowed to occur only at regularly spaced points in time, as seen in recently developed communication systems [20, 21]. Only Boxma & Groenendijk [3] have treated a discrete-time single-class (non-priority) multi-queue system.

The main goal of this paper is to obtain discrete-time analogs of the results for continuous-time priority systems [8, 17]. Section 2 describes our discrete-time priority systems in details. By using an ergodic argument we present some preliminary results for a general input system. In Section 3, assuming a batch Bernoulli process input, we derive a pseudo-conservation law. The approach taken here essentially relies on the simplified argument of Shimogawa & Takahashi [17]. We show how the argument in [17] can be generalized and applied to the

batch Bernoulli process input system. In Section 4, by letting the slot length tend to zero, we obtain the continuous-time results for Poisson input systems [2, 8, 17] as special cases.

2. Model description and preliminary results

Time is divided into slots which are equal to time unity (one) in length. This unit-time slot is assumed for Sections 2 and 3. We consider a multi-queue, single-server system with N stations, each with infinite queueing capacity. The stations are visited by the server in cyclic order. We assume four types of service strategy: exhaustive service, gated service, 1-limited service, and 1-decrementing service. See Takagi [18,19] for the definition of these types of service strategy for a single-class (non-priority) system.

We assume P priority classes of customers arriving at each station. We further assume a non-preemptive or head-of-the-line (HL) local priority, as in [7, 8, 10, 12, 17]. By ***local*** we mean that the customer class to be selected for service at a station depends only on the customers present at that station (and independent of the customers at the other stations). A class-i customer has precedence over a class-j customer if $i < j$ $(1 \leq i, j \leq P)$.

For every (exhaustive, gated, 1-limited, or 1-decrementing) service strategy, when the server visits a station i $(1 \leq i \leq N)$ and finds no customers at that station, immediately the server moves on to the next station $i+1$ (after the server visits station N, it moves on to station 1). The next station index will be denoted as $i(\text{mod } N)+1$.

If the server finds any customer at station i $(1 \leq i \leq N)$ upon the server's arrival, the server remains at that station, according to one of the followings.

e) For *exhaustive* service, all customers at station i are served according to the HL priority rule until no more customers are left at that station. In other words, when the server leaves station i to move on to the next station $i(\text{mod } N)+1$, no customers are left in station i.

g) For *gated* service, only customers found at station i upon the server's arrival are served according to the HL priority rule. The station can be considered to have a gate. The gate is opened upon the server's arrival at the station, and it is shut just after the server's arrival. The server accepts and serves only those customers that have passed through the gate.

1*l*) For *1-limited* service, the highest priority class found at station i upon the server's arrival is selected, and only one customer of the selected class is served. The 1-limited service is sometimes called a pure limited service.

1d) For *1-decrementing* service, the highest priority class found at station i upon the server's arrival is selected, and the selected-class customers are served until the number of customers in that class becomes one less than there were at the server's arrival. The 1-decrementing service is sometimes called a pure decrementing service or a semi-exhaustive service.

The notation "1-*" means that only one priority class of customers is selected upon the server's arrival at a 1-* service station and only the selected priority class is served during this visit. See Karvelas and Leon-Garcia [10] for a practical example of this strategy.

Arrivals occur at the beginning of a slot, as in a discrete-time environment [3, 20, 21]. Customers from an individual priority class arrive at a station according to a batch Bernoulli process [20]. Let X_{ip} be the i. i. d. batch size [the number of class-p arriving customers at station i during a slot be independent and identically distributed (i. i. d.)] with first two

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

moments; $\lambda_{ip} := E[X_{ip}], \lambda_{ip}^{(2)} := E[X_{ip}^2]$. Similarly, we introduce for lumped arrivals:

$$X_i := \sum_{p=1}^{P} X_{ip} \text{ (for station } i) \text{ and } X := \sum_{i=1}^{N} \sum_{p=1}^{P} X_{ip} \text{ (for the total system),}$$

with first two moments $\lambda_i, \lambda_i^{(2)}$; and $\lambda, \lambda^{(2)}$, respectively.

For simplicity, we will refer to a class-p customer arriving at station i as a *class-*(i,p) customer. Let H_{ip} be the i. i. d. service time of a class-(i,p) customer with first two moments;

$$h_{ip} := E[H_{ip}], \text{ and } h_{ip}^{(2)} := E[H_{ip}^2].$$

The offered loads are then given as

$$\rho_{ip} := \lambda_{ip} h_{ip} \text{ (for class } (i,p)), \rho_i := \sum_{p=1}^{P} \rho_{ip} \text{ (for station } i), \text{ and}$$

$$\rho := \sum_{i=1}^{N} \sum_{p=1}^{P} \rho_{ip} \text{ (for the total system).}$$

Let S_i be the i. i. d. switch-over time of the server between stations i and $i(\text{mod } N)+1$, with first two moments $s_i, s_i^{(2)}$. The total switch-over time of the server during a cycle is given as

$$S := \sum_{i=1}^{N} S_i,$$

with first two moments s and $s^{(2)}$.

As in the literature [11, 12, 15, 20], the model described above will be referred to as a discrete-time $Geom^X/GI/1$ *type* multi-queue priority system, since positive batches for an individual class-(i,p) form a Bernoulli process, i. e., the batch inter-arrival times are geometrically distributed. Here, the positive batch means a batch with positive size, and the batch inter-arrival time means the time between two successive arrivals of positive batches, as in Takahashi & Hashida [20]. The model where the batch inter-arrival times are generally distributed but other assumptions are the same as the one described above, will be referred to as a discrete-time $G^X/GI/1$ *type* multi-queue priority system.

Let C be the cycle time, i. e., the time between two successive arrivals of the server at a station. By using the ergodic theorem and Little's law, we obtain the mean cycle time, denoted by c, as

$$c := E[C] = \frac{s}{1-\rho} \tag{2.1}$$

for the discrete-time $G^X/GI/1$ type multi-queue priority system. See Appendix for the proof of (2.1). It should be noted that (2.1) was previously proven for memory-less (batch Poisson and batch Bernoulli) input systems in the literature. In the appendix we have shown that (2.1) is still valid for a general input system.

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

Let V_i be the visit period for station i, i. e., the length of time that the server stays at station i $(1 \leq i \leq N)$. The argument for deriving (2.1) is similarly applied to get

$$v_i := E[V_i] = \rho_i c = \rho_i \frac{s}{1-\rho} \tag{2.2}$$

for the $G^X/GI/1$ type multi-queue priority system with any(exhaustive, gated, 1-limited, or 1-decrementing) service strategy at station i.

Let (e, g, $1l$, and $1d$) be the partition of the station index set $\{1, 2, \cdots, N\}$ defined by
$e := \{j \mid \text{station } j \text{ is exhaustive}\}$,
$g := \{j \mid \text{station } j \text{ is gated}\}$,
$1l := \{j \mid \text{station } j \text{ is 1-limited}\}$, and
$1d := \{j \mid \text{station } j \text{ is 1-decrementing}\}$.

Let A_{ip} $(1 \leq i \leq N; 1 \leq p \leq P)$ denote the event where the server finds that the highest priority class is p upon its arrival at station i. The probability of event A_{ip}, denoted by $Pr[A_{ip}]$, is also the probability that class-(i,p) customers are being served during the server's visit period, if $i \in 1l \cup 1d$.

If station i adopts the 1-limited service strategy, it can be verified that

$$Pr[A_{ip}] = \lambda_{ip} \frac{s}{1-\rho} \ (1 \leq p \leq P, i \in 1l). \tag{2.3}$$

for the $G^X/GI/1$ type multi-queue priority system. See Appendix for the proof of (2.3).

If station i adopts the 1-decrementing service strategy, the mean visit period for class-(i,p) customers is

$$\frac{h_{ip}}{1-\rho_{ip}}$$

given A_{ip}, since this fraction corresponds to the mean busy period initiated by one class-(i,p) customer for a discrete-time $Geom^X/GI/1$ queue with batch size X_{ip} and service time H_{ip}. (See Takahashi & Hashida [20].) Noting that $\rho_{ip}c$ is the mean length of time the server serves class-(i,p) customers during each visit to station i, we have

$$\rho_{ip}c = Pr[A_{ip}] \frac{h_{ip}}{1-\rho_{ip}}, \tag{2.4}$$

which yields, from (2.1),

$$Pr[A_{ip}] = \frac{\lambda_{ip}(1-\rho_{ip})s}{1-\rho} \ (1 \leq p \leq P, i \in 1d). \tag{2.5}$$

We have assumed that the system is stable in our derivation of equations (2.1) through (2.5). Although it is beyond the scope of this paper to obtain a necessary and sufficient condition for stability, the following condition is necessary:

$$\text{“}\rho < 1 \ (i \in e \cup g),\text{”}$$
$$\text{“}\rho < 1 \text{ and } \frac{\lambda_i s}{1-\rho} < 1 \ (i \in 1l),\text{”}$$

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

or

$$\text{“}\rho < 1 \text{ and } \frac{\lambda_i(1-\rho_i)s}{1-\rho} < 1 \ (i \in 1d),\text{”}$$

which we will assume from now on.

3. The pseudo-conservation law

3.1 General form of the pseudo-conservation law

To derive our pseudo-conservation law, we start with the stochastic decomposition property for the work load in a single-server vacation model. We define

V_c: the amount of work required by all customers in the cyclic service system described in Section 2 at an arbitrary epoch,

V_o: the amount of work in the lumped discrete-time FCFS $Geom^X/GI/1$ system (without switch-over time) where the same arrivals and service times are assumed as in our cyclic service system at an arbitrary epoch, and

Y: the amount of work in the cyclic service system at an arbitrary epoch in a switching interval. Here, an arbitrary epoch is supposed to be the instant just after the beginning of a slot as customary in discrete-time queueing literature; See [3, 20, 21].

For our discrete-time multi-queue priority model, it is straightforward to verify the following stochastic decomposition property, as in Boxma & Groenendijk [3]:

$$\mathbf{V_c} \stackrel{d}{=} \mathbf{V_o} + \mathbf{Y}, \tag{3.1}$$

where $\stackrel{d}{=}$ stands for equality in distribution. Especially, it follows that

$$v_c := E[\mathbf{V_c}] = E[\mathbf{V_o}] + E[\mathbf{Y}]. \tag{3.2}$$

It is well known (see in [20]) for the standard discrete-time FCFS $Geom^X/GI/1$ system that

$$E[\mathbf{V_o}] = \frac{\rho}{2(1-\rho)} \sum_{i=1}^{N}\sum_{p=1}^{P} \lambda_{ip} h_{ip}^{(2)} + \sum_{i=1}^{N}\sum_{p=1}^{P} \frac{\lambda_{ip}^{(2)} - \lambda_{ip}^2 - \lambda_{ip}}{2(1-\rho)} h_{ip}^2 + \sum_{i=1}^{N}\sum_{p=1}^{P} \rho_{ip}\left[\frac{h_{ip}^{(2)}}{2h_{ip}} + \frac{1}{2}\right]. \tag{3.3}$$

Let m_i denote the mean amount of work that is left at station i after an arbitrary departure of the server from that station, and w_{ip} be the mean waiting time of a class-(i,p) customer in station i $(1 \le i \le N, 1 \le p \le P)$. Following the argument in Boxma & Groenendijk [2, 3], which is also seen to be valid for our HL priority model, we have

$$E[\mathbf{Y}] = \rho\left[\frac{s^{(2)}}{2s} - \frac{1}{2}\right] + \frac{s}{2(1-\rho)}\left(\rho^2 - \sum_{i=1}^{N} \rho_i^2\right) + \sum_{i=1}^{N} m_i. \tag{3.4}$$

Since our priority rule is non-preemptive, we have (as seen in [20])

$$v_c = \sum_{i=1}^{N}\sum_{p=1}^{P} \rho_{ip} w_{ip} + \sum_{i=1}^{N}\sum_{p=1}^{P} \rho_{ip}\left[\frac{h_{ip}^{(2)}}{2h_{ip}} + \frac{1}{2}\right]. \tag{3.5}$$

Substituting (3.3) through (3.5) into (3.2) yields the following form for the pseudo-conservation law for our discrete-time priority system.

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

Lemma 3.1 For a discrete-time $Geom^X/GI/1$ type multi-queue priority system with mixed exhaustive, gated, 1-limited, and 1-decrementing service stations, we have

$$\sum_{i=1}^{N}\sum_{p=1}^{P}\rho_{ip}w_{ip} = \frac{\rho}{2(1-\rho)}\sum_{i=1}^{N}\sum_{p=1}^{P}\lambda_{ip}h_{ip}^{(2)} + \sum_{i=1}^{N}\sum_{p=1}^{P}\frac{\lambda_{ip}^{(2)}-\lambda_{ip}^2-\lambda_{ip}}{2(1-\rho)}h_{ip}^2 + \rho[\frac{s^{(2)}}{2s}-\frac{1}{2}]$$
$$+\frac{s}{2(1-\rho)}[\rho^2-\sum_{i=1}^{N}\rho_i^2]+\sum_{i=1}^{N}m_i, \tag{3.6}$$

where w_{ip} denotes the mean waiting time of a class p customer in station i, and m_i the mean amount of work that is left at station i after an arbitrary departure of the server from that station.

Remark 3.1 Bisdikian [1] pointed out an error is involved in Boxma and Groenendijk's [3] result even for a zero switch-over time system (but did not mention which equation is incorrect). We have corrected the error through replacing Eq. (2.3) of Boxma and Groenendijk [3] by our equation (3.3), according to Takahashi and Hashida [20]. □

3.2 Evaluation of m_i $(i = 1, \cdots, N)$

We are now in a position to evaluate m_i for each individual (exhaustive, gated, 1-limited and 1-decrementing) service strategy. The first lemma treats exhaustive and gated service stations, which will be verified by using a fairly straightforward argument.

Lemma 3.2 The mean amount of work left behind at station i after the departure of the server from that station, m_i, is given by

$$m_i = 0 \ (i \in e), \tag{3.7}$$

and

$$m_i = c\rho_i^2 \ (i \in g). \tag{3.8}$$

Proof. Equation (3.7) comes immediately from the definition of the exhaustive service station. If station i adopts the gated service strategy, m_i corresponds to the mean amount of work required by customers that arrive at station i during the station i visit period of the server. Since the visit period is given by $\rho_i c$ from (2.2), we have

$$m_i = \sum_{p=1}^{P}\lambda_{ip}(\rho_i c)h_{ip} = c\rho_i^2,$$

which yields (3.8), completing the proof.□

The next lemma treats the 1-limited and 1-decrementing service stations, which will require somewhat more work than Lemma 3.1. We will show how the simple argument by Shimogawa & Takahashi [17] can be generalized and applied to our discrete-time system. (See also Remarks 3.2 and 4.1.)

Lemma 3.3 Let M_i be the amount of work left behind at station i after the departure of the server from that station. The expectation m_i of M_i is expressed as

$$m_i = c[\rho_i^2 + \sum_{p=1}^{P}\frac{\lambda_{ip}^{(2)}-\lambda_{ip}}{2}h_{ip} + \sum_{p=1}^{P}\{\lambda_{ip}\sum_{u=p+1}^{P}\rho_{iu} + \rho_{ip}\sum_{u=1}^{p}\lambda_{iu}\}w_{ip}] \ (i \in 1l), \tag{3.9}$$

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

and

$$m_i = c[\rho_i^2 - \sum_{p=1}^{P} \rho_{ip}^2 + \sum_{p=1}^{P}\{\lambda_{ip}(1-\rho_{ip}) \sum_{u=p+1}^{P} \rho_{iu} + \rho_{ip} \sum_{u=1}^{p} \lambda_{iu}(1-\rho_{iu})\} w_{ip}$$
$$- \sum_{p=1}^{P}[\frac{\lambda_{ip}^2 h_{ip}^{(2)}}{2(1-\rho_{ip})}\{\sum_{u=p+1}^{P} \rho_{iu}(3-\rho_{ip}) + (1-\rho_{ip})\rho_{ip}\}$$
$$+ \lambda_{ip}(1-\rho_{ip})\{\sum_{u=p+1}^{P} \frac{\lambda_{iu}\rho_{iu}h_{iu}^{(2)}}{2(1-\rho_{iu})^2}(2-\rho_{iu})\} + R_{ip}(\text{disc})]] \ (i \in 1d), \qquad (3.10)$$

where R_{ip} (disc) is defined by

$$R_{ip}(\text{disc}) := \frac{1}{2}[\frac{\rho_{ip}h_{ip}}{1-\rho_{ip}}(\lambda_{ip}^{(2)} - \lambda_{ip}^2 - \lambda_{ip})(1-\rho_{ip}+\rho_{ip}\sum_{u=p+1}^{P}\rho_{iu}) - \frac{\lambda_{ip}^{(2)} - \lambda_{ip}}{\lambda_{ip}}(1-\rho_{ip})\rho_{ip}$$
$$+ \lambda_{ip}(1-\rho_{ip})\{\sum_{u=p+1}^{P}(\lambda_{iu}^{(2)} - \lambda_{iu}^2 - \lambda_{iu})(1-\rho_{iu}+\rho_{iu}^2)\frac{h_{ip}}{\lambda_{iu}(1-\rho_{iu})^2}\}$$
$$+ \lambda_{ip}\sum_{u=p+1}^{P}\rho_{iu}\{\rho_{ip}^2 - (1-\rho_{ip})(1+\rho_{iu})\}].$$

Proof. To derive $m_i(:= E[M_i])$, we use the following decomposition:

$$m_i = \sum_{p=1}^{P} m_{ip} Pr[A_{ip}], \qquad (3.11)$$

where m_{ip} denotes the conditional expectation of M_i given A_{ip}, i. e.,

$$m_{ip} := E[M_i \mid A_{ip}] \ (1 \le p \le P).$$

Since we already know $Pr[A_{ip}]$ from (2.3) and (2.5) for the 1-limited and 1-decrementing service stations, all we have to do is to evaluate the quantity m_{ip}.

As in Shimogawa & Takahashi [17], we decompose the quantity m_{ip} in the following way. Let κ_{ipq} be the conditional mean number of class-(i,q) customers just after the departure of the server given $A_{ip}(1 \le p, q \le P)$. Since our priority is non-preemptive, we have

$$m_{ip} = \sum_{q=1}^{P} \kappa_{ipq} h_{iq} \ (1 \le p \le P). \qquad (3.12)$$

To evaluate κ_{ipq} appearing in (3.12), we need the following notation. Let T_{ip} be the sojourn time (or visit period) of the server at station i under the condition that the server finds class p the highest priority, i. e., A_{ip}. We can then have the mean sojourn time of the server, $t_{ip}(:= E[T_{ip}])$, as

$$t_{ip} = h_{ip}(i \in 1l), \text{ and } t_{ip} = \frac{h_{ip}}{1-\rho_{ip}}(i \in 1d). \qquad (3.13)$$

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

We consider an arbitrary class-(i,p) customer who has just finished its service during T_{ip} and is going to depart from the system. We will refer to this arbitrary class-(i,p) customer as a *tagged customer*.

We can easily evaluate κ_{ipq} for $p > q$. It follows from our priority rule that if service is given to the tagged customer, then no higher class customers are present at that station upon the server's arrival. Thus κ_{ipq} equals the mean number of class-(i,q) customers who arrive during T_{ip}, i. e., $\kappa_{ipq} = \lambda_{iq} t_{ip}$ or

$$\kappa_{ipq} = \lambda_{iq} h_{ip} \text{ (for } p > q, i \in 1l); \text{ and } \kappa_{ipq} = \lambda_{iq} \frac{h_{ip}}{1 - \rho_{ip}} \text{ (for } p > q, 1 \in 1d). \tag{3.14}$$

It now remains for us to evaluate κ_{ipq} for $p \leq q$. We will discuss the $i \in 1l$ and the $i \in 1d$ cases separately.

The $i \in 1l$ case. We first evaluate κ_{ipq} for $p = q$. Consider class-(i,p) (the same class) customers who arrive at the same slot as the tagged customer, but are served after the tagged customer. We will refer to these customers as *pseudo-subsequent* customers. As seen in Takahashi & Hashida [20], it follows that the number of pseudo-subsequent customers is given as the backward recurrence time of class-(i,p) batch size. The mean number of pseudo-subsequent customers is thus given by $[\lambda_{ip}^{(2)} - \lambda_{ip}]/(2\lambda_{ip})$. The number of customers in κ_{ipp}, except for these pseudo-subsequent customers, equals the number of customers who arrive during the sojourn time (with mean $w_{ip} + h_{ip}$) of the tagged customer. This observation leads to

$$\kappa_{ipp} = \lambda_{ip}(w_{ip} + h_{ip}) + \frac{\lambda_{ip}^{(2)} - \lambda_{ip}}{2\lambda_{ip}} \text{ (for } p = q, i \in 1l). \tag{3.15}$$

We then evaluate κ_{ipq} for $p < q$. To make the discussion clear, we decompose κ_{ipq} into two terms:

κ_{ipq} (senior): the mean number of class-(i,q) customers who were already present upon the arrival of the tagged customer and remain there until the end of the tagged customer's service, and

κ_{ipq} (subseqt): the mean number of class-(i,q) customers who arrive during the sojourn time of the tagged customer and remain there until the end of the tagged customer's service.

Note that the GASTA (Geometric Arrivals See Time Averages) property [9] implies that the tagged customer sees time averages, since the batch inter-arrival time of class (i,p) is geometrically distributed. We then have

$$\kappa_{ipq}(\text{senior}) = L_{iq} = \lambda_{iq} w_{iq},$$

where L_{iq} denotes the mean number of class-(i,q) waiting customers. With probability ρ_{iq}, the tagged customer found the server busy with a class-$(i.q)$ customer upon the tagged customer's arrival, but this class-(i,q) customer leaves from the system before the tagged customer's service (since we assume the non-preemptive priority rule). From our 1-l service strategy and $p < q$, it follows that no other class-(i,q) customers than this class-(i,q) customer will be served until the end of the tagged customer's service. Hence, this case (with probability ρ_{iq}) no longer contributes to κ_{ipq} (senior).

From the independence between the system state and the arrival processes, we have

$$\kappa_{ipq}(\text{subseqt}) = \lambda_{ip}(w_{ip} + h_{ip}).$$

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

With probability ρ_{iq} the tagged customer found the server busy with a class-$(i.q)$ customer upon the tagged customer's arrival, but this class-(i,q) customer leaves from the system before the tagged customer's service. Also, this case (with probability ρ_{iq}) no longer contributes to κ_{ipq}(subseqt). Summing up these two equations above gives

$$\kappa_{ipq} = \lambda_{iq}(w_{ip} + w_{iq} + h_{ip}) \text{ (for } p < q, i \in 1l). \tag{3.16}$$

Substituting (3.14) through (3.16) into (3.12) and using (2.3), (3.11), and (3.12), we obtain (3.9). This completes the proof of the former part.

The $i \in 1d$ case. We will generalize the argument of Fournier & Rosberg [8] for our batch input system. We first evaluate κ_{ipq} for $p = q$ as in the $i \in 1l$ case.

Let n_{ip} be the mean number of class-(i,p) customers left behind by a class-(i,p) departing [tagged] customer. The derivation for (3.15) can be similarly applied to give

$$n_{ip} = \lambda_{ip}(w_{ip} + h_{ip}) + \frac{\lambda_{ip}^{(2)} - \lambda_{ip}}{2\lambda_{ip}}. \tag{3.17}$$

The 1-decrementing service strategy gives

$$n_{ip} = \kappa_{ipp} + [\frac{\lambda_{ip}^2 h_{ip}^{(2)}}{2(1-\rho_{ip})} + \frac{\lambda_{ip}^{(2)} - \lambda_{ip}^2 - \lambda_{ip}}{2(1-\rho_{ip})} h_{ip} + \rho_{ip}]. \tag{3.18}$$

See Chiarawongse and Srinivasan [5] for (3.18). Substituting (3.17) into (3.18) gives

$$\kappa_{ipp} = \lambda_{ip} w_{ip} + \frac{\lambda_{ip}^{(2)} - \lambda_{ip}}{2\lambda_{ip}} - \frac{\lambda_{ip}^2 h_{ip}^{(2)}}{2(1-\rho_{ip})} - \frac{\lambda_{ip}^{(2)} - \lambda_{ip}^2 - \lambda_{ip}}{2(1-\rho_{ip})} h_{ip} (\text{for } p = q, i \in 1d). \tag{3.19}$$

It remains for us to evaluate κ_{ipq} for $p < q$. If we denote by l_{ipq} the mean number of class-(i,q) customers at the beginning of the visit period T_{ip}, it is straightforward that

$$\kappa_{ipq} = l_{ipq} + \lambda_{iq} t_{ip} \text{ (for p < q)}. \tag{3.20}$$

The quantity l_{ipq} can be expressed as

$$l_{ipq} = k_{ipq} - j_{ipq} \text{ (for } p < q), \tag{3.21}$$

where k_{ipq} is the mean number of class-(i,q) customers at the beginning of the tagged customer's service, and j_{ipq} is the mean number of class-(i,q) customers who arrive from the beginning of the visit period T_{ip} until the beginning of the tagged customer's service.

We are now going to evaluate k_{ipq} and j_{ipq}. We decompose k_{ipq} into the following two terms:

k_{ipq} (senior): the mean number of class-(i,q) customers who were already present upon the arrival of the tagged customer and who remain until the beginning of the tagged customer's service, and

k_{ipq} (subseqt): the mean number of class-(i,q) customers who arrive during the waiting time of the tagged customer and who remain until the beginning of the tagged customer's service.

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

The quantity k_{ipq}(senior) is obtained from

$$k_{ipq}(\text{senior})h_{iq} = L_{iq}h_{iq} + \rho_{iq}[\frac{h_{iq}^{(2)}}{2h_{iq}} + \frac{1}{2}] - \rho_{iq}\{[\frac{\lambda_{iq}h_{iq}^{(2)}}{2(1-\rho_{iq})} + \frac{\lambda_{iq}^{(2)} - \lambda_{iq}^2 - \lambda_{iq}}{2(1-\rho_{iq})\lambda_{iq}}h_{iq}]\frac{1}{\rho_{iq}}\} (\text{for } p < q). \tag{3.22}$$

The left-hand side is the mean amount of work for class-(i,q) customers who were already present upon the arrival of the tagged customer. The first two terms on the right-hand side represent the mean amount of work seen by the tagged customer upon its arrival. With probability ρ_{iq}, however, the tagged customer arrived during the class-(i,q) visit period. In this case, these two terms include the amount of class-(i,q) work that will leave from the system until the tagged customer's service begins. The braced term on the right-hand side thus represents this expected leaving amount, which corresponds to the mean amount of class-(i,q) work at an arbitrary time of the busy period for the discrete-time $Geom^X/GI/1$ queue with batch size X_{iq} and service time H_{iq} for the 1-decrementing service strategy. This observation validates (3.22).

The quantity k_{ipq}(subseqt) is obtained as

$$k_{ipq}(\text{subseqt}) = \lambda_{iq}w_{ip} - \rho_{iq}\lambda_{iq}[\frac{b_{iq}^{(2)}}{2b_{iq}} - \frac{1}{2}] \ (\text{for } p < q), \tag{3.23}$$

where b_{iq} and $b_{iq}^{(2)} (1 \leq q < P)$ denote the first two moments of the busy period for the discrete-time $Geom^X/GI/1$ queue with batch size X_{iq} and service time H_{iq} initiated by one class-(i,q) customer, i. e.,

$$b_{iq} = \frac{h_{iq}}{1-\rho_{iq}},$$

and

$$b_{iq}^{(2)} = \frac{h_{iq}^{(2)} + h_{iq}^3[\lambda_{iq}^{(2)} - \lambda_{iq}^2 - \lambda_{iq}]}{(1-\rho_{iq})^3} (1 \leq q \leq P). \tag{3.24}$$

The number of class-(i,q) customers who arrive during the waiting time of the tagged customer will be given by the first term on the right-hand side of (3.23) unless the server serves class-(i,q) customers. With probability ρ_{iq}, however, the tagged customer arrived during a class-(i,q) visit period. In this case, those class-(i,q) customers who arrive during the interval I_{iq} from the tagged customer's arrival epoch to the end of the class-(i,q) visit period will leave from the system (and so those customers should be removed). The interval I_{iq} corresponds the backward recurrence time of the class-(i,q) busy period, so that

$$E[I_{iq}] = \frac{b_{iq}^{(2)}}{2b_{iq}} - \frac{1}{2}.$$

The expected number of those class-(i,q) customers to be removed is then given by the second negative term on the right-hand side, validating (3.23).

Hence, it follows from (3.22) through (3.24) that

$$k_{ipq} = \lambda_{iq}\{w_{ip} + w_{iq} + [\frac{h_{ip}^{(2)}}{2h_{ip}} + \frac{1}{2}] - \rho_{iq}[\frac{h_{iq}^{(2)} + h_{iq}^3(\lambda_{iq}^{(2)} - \lambda_{ip}^2 - \lambda_{iq})}{2h_{iq}(1-\rho_{iq})^2} - \frac{1}{2}]\}$$
$$- [\frac{\lambda_{iq}h_{iq}^{(2)}}{2(1-\rho_{iq})h_{iq}} + \frac{\lambda_{iq}^{(2)} - \lambda_{iq}^2 - \lambda_{iq}}{2(1-\rho_{iq})\lambda_{iq}}] \ (\text{for } p < q). \tag{3.25}$$

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

Recalling (3.20) and (3.21), it remains for us to evaluate the quantity j_{ipq}. To simplify the argument, we assume a non-preemptive LIFO rule within class (i,p). This assumption does not change the queue length nor the work load for class (i,p). If the tagged customer does not arrive during the visit period T_{ip}, it will be served first during T_{ip} under the 1-decrementing service strategy. This is because a class-(i,p) customer who arrives after the tagged customer and who finds that the server is serving other classes or that the server is switching stations has to initiate a class-(i,p) visit period. This class-(i,p) visit period comes before T_{ip} for the LIFO rule. Hence, the elapsed time between the beginning of the visit period T_{ip} and the beginning of the tagged customer's service, denoted as Γ_{ip}, is zero unless the tagged customer arrives during T_{ip}. We thus have

$$\begin{aligned} j_{ipq} &= \lambda_{iq} E[\Gamma_{ip}] \\ &= \lambda_{iq}\{\rho_{ip}\{[\frac{h_{ip}^{(2)} + h_{ip}^{3}(\lambda_{ip}^{(2)} - \lambda_{ip}^{2} - \lambda_{ip})}{2h_{ip}(1-\rho_{ip})^{2}} - \frac{1}{2}] + [\frac{h_{ip}^{(2)}}{2h_{ip}} + \frac{1}{2}] + \lambda_{ip} b_{ip}[\frac{h_{ip}^{(2)}}{2h_{ip}} + \frac{1}{2}]\}\} \\ &(p < q). \end{aligned} \tag{3.26}$$

The tagged customer arrives during T_{ip} (and in this case Γ_{ip} is positive) with probability ρ_{ip}. The first term in the inner braces on the right-hand side represents the mean elapsed time between the beginning of T_{ip} and the arrival epoch of the tagged customer, corresponding to the mean backward recurrence time of the busy period for the discrete-time $Geom^{X}/GI/1$ queue with batch size X_{ip} and service time H_{ip} initiated by one class-(i,p) customer. The second term in the inner braces is the mean residual service time seen by the tagged customer (because of the non-preemptive rule). The third term in the braces is the waiting time of the tagged customer, corresponding to the mean busy period initiated by those customers who arrive during this residual service time (because of the LIFO rule). This validates (3.26).

Using (3.21), (3,25), and (3.26), (3,20) then gives

$$\begin{aligned} \kappa_{ipq} &= \lambda_{iq}\{w_{ip} + w_{iq} + [\frac{h_{ip}^{(2)}}{2h_{ip}} + \frac{1}{2}] - \rho_{iq}[\frac{h_{iq}^{(2)} + h_{iq}^{3}(\lambda_{iq}^{(2)} - \lambda_{iq}^{2} - \lambda_{iq})}{2h_{iq}(1-\rho_{iq})^{2}} - \frac{1}{2}]\} \\ &- [\frac{\lambda_{iq} h_{iq}^{(2)}}{2(1-\rho_{ip})h_{iq}} + \frac{\lambda_{iq}^{(2)} - \lambda_{iq}^{2} - \lambda_{iq}}{2(1-\rho_{iq})\lambda_{iq}}] \\ &- \lambda_{iq}\{\rho_{ip}\{[\frac{h_{ip}^{(2)} + h_{ip}^{3}(\lambda_{ip}^{(2)} - \lambda_{ip}^{2} - \lambda_{ip})}{2h_{ip}(1-\rho_{ip})^{2}} - \frac{1}{2}] + [\frac{h_{ip}^{(2)}}{2h_{ip}} + \frac{1}{2}] + \lambda_{ip} b_{ip}[\frac{h_{ip}^{(2)}}{2h_{ip}} + \frac{1}{2}]\}\} \\ &+ \lambda_{iq} t_{ip} \ (\text{for } p < q, i \in 1d). \end{aligned} \tag{3.27}$$

As in the $i \in 1l$ case, substituting (3.14), (3.19), and (3.27) into (3.12), and using (2.5), (3.11) through (3.13), we obtain (3.10). This completes the proof.□

Lemma 3.1, together with Lemmas 3.2 and 3.3, gives the following pseudo-conservation law.

Theorem 3.1 (Pseudo-conservation law) For a discrete-time $Geom^{X}/GI/1$ type multi-queue priority system with mixed exhaustive, gated, 1-limited, and 1-decrementing services stations, we have

$$\sum_{i \in e \cup g} \sum_{p=1}^{P} \rho_{ip} w_{ip} + \sum_{i \in 1l} \sum_{p=1}^{P} [\rho_{ip} - c\{\lambda_{ip} \sum_{u=p}^{P} \rho_{iu} + \rho_{ip} \sum_{u=1}^{p-1} \lambda_{iu}\}] w_{ip}$$

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

$$+\sum_{i\in 1d}\sum_{p=1}^{P}[\rho_{ip}-c\{\lambda_{ip}(1-\rho_{ip})\sum_{u=p}^{P}\rho_{iu}+\rho_{ip}\sum_{u=1}^{p-1}\lambda_{iu}(1-\rho_{iu})\}]w_{ip}$$

$$=\frac{\rho}{2(1-\rho)}\sum_{i=1}^{N}\sum_{p=1}^{P}\lambda_{ip}h_{ip}^{(2)}+\sum_{i=1}^{N}\sum_{p=1}^{P}\frac{\lambda_{ip}^{(2)}-\lambda_{ip}^{2}-\lambda_{ip}}{2(1-\rho)}h_{ip}^{2}+\rho[\frac{s^{(2)}}{2s}-\frac{1}{2}]$$

$$+c\{\frac{1}{2}[\rho^{2}-\sum_{i=1}^{N}\rho_{i}^{2}]+\sum_{i\in g\cup 1l\cup 1d}\rho_{i}^{2}+\sum_{i\in 1l}\sum_{p=1}^{P}\frac{\lambda_{ip}^{(2)}-\lambda_{ip}}{2\lambda_{ip}}\rho_{ip}$$

$$-\sum_{i\in 1d}\sum_{p=1}^{P}[\rho_{ip}^{2}+\frac{\lambda_{ip}^{2}h_{ip}^{(2)}}{2(1-\rho_{ip})}\{\sum_{u=p+1}^{P}\rho_{iu}(3-\rho_{ip})+(1-\rho_{ip})\rho_{ip}\}$$

$$+\lambda_{ip}(1-\rho_{ip})\sum_{u=p+1}^{P}\frac{\lambda_{iu}\rho_{iu}h_{iu}^{(2)}}{2(1-\rho_{iu})^{2}}(2-\rho_{iu})+R_{ip}(\mathrm{disc})]\},\qquad(3.28)$$

where c and $R_{ip}(\mathrm{disc})$ are given in (2.1) and (3.10).

Remark 3.2 a) For a discrete-time non-preemptive priority system with zero switch-over times, if we set $s=0$ and $s^{(2)}/s=1$, (3.28) together with (3.3) reduces to the conservation-law result in Takahashi & Hashida [20]. For a single-class (non-priority) discrete-time system with non-zero switch- over times, if we set $P=1$, (3.28) corresponds to Eq. (4.22) of Boxma & Groenedijk [3], correcting their error. b) In the busy-period second-moment (3.24), we have corrected a typographical error in Eq. (26) of Klimko & Neuts [13] where cubing of the service time in the numerator is missing.□

4. The continuous-time result as a special case

So far we have expressed all quantities in slots with the slot length equal to unity. If instead, we assume a slot to be of length Δ, and if we let the length of a slot go to zero ($\Delta\to 0$) as in the discrete-time literature [3, 16], we can obtain the continuous-time pseudo-conservation law.

Even if we assume that the slot length is Δ, the results in Section 3 are still valid. To be more exact, in this case, all the quantities are measured in Δ units. We have to distinguish between a quantity measured in Δ units and the corresponding quantity measured in time units. Here, we will attach a tilde ($\sim$) to quantities measured in Δ units, while we will use the notation in Sections 2 and 3 for quantities measured in time units. The mean waiting time, for example, is expressed as

$$w_{ip}=\tilde{w}_{ip}\Delta,$$

where $\tilde{w}_{ip}$ denotes the class-(i,p) waiting time in Δ units and w_{ip} the corresponding waiting time in time units. Similarly, we have

$$h_{ip}=\tilde{h}_{ip}\Delta \text{ and } h_{ip}^{(2)}=\tilde{h}_{ip}^{(2)}\Delta^{2}.$$

It should be noted that all the results in Section 3 are valid for quantities with these tildes when the slot length is Δ.

Let $\tilde{X}_{ip}(z)$ be the pgf of class-(i,p) batch size $\tilde{X}_{ip}$ in a slot with length Δ, and $X_{ip}(z)$ be the pgf of the total number of class-(i,p) customers during a time unity. Since $1/\Delta$ is the number of slots per time unity and since we are assuming a batch Bernoulli process (where

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

the batch size arriving at a slot is statistically independent of the one at another slot), we have

$$\tilde{X}_{ip}(z) = X_{ip}(z)^{1/\Delta},$$

which yields

$$\lambda_{ip} = \frac{\tilde{\lambda}_{ip}}{\Delta}, \text{ and } \lambda_{ip}^{(2)} = \frac{\tilde{\lambda}_{ip}^{(2)}}{\Delta} + \frac{1}{\Delta}(\frac{1}{\Delta} - 1)\tilde{\lambda}_{ip}^2.$$

Traffic intensity is invariant regardless of the slot length, i. e.,

$$\rho_{ip} = \lambda_{ip} h_{ip} = \tilde{\lambda}_{ip}\tilde{h}_{ip} = \tilde{\rho}_{ip}.$$

Equation (3.28) with tildes leads to

$$\begin{aligned}
&\sum_{i\in e\cup g}\sum_{p=1}^{P} \rho_{ip} w_{ip}\frac{1}{\Delta} + \sum_{i\in 1l}\sum_{p=1}^{P}[\rho_{ip} - c\{\lambda_{ip}\sum_{u=p}^{P}\rho_{iu} + \rho_{ip}\sum_{u=1}^{p-1}\lambda_{iu}\}]w_{ip}\frac{1}{\Delta} \\
&+ \sum_{i\in 1d}\sum_{p=1}^{P}[\rho_{ip} - c\{\lambda_{ip}(1-\rho_{ip})\sum_{u=p}^{P}\rho_{iu} + \rho_{ip}\sum_{u=1}^{p-1}\lambda_{iu}(1-\rho_{iu})\}]w_{ip}\frac{1}{\Delta} \\
&= \frac{\rho}{2(1-\rho)}\sum_{i=1}^{N}\sum_{p=1}^{P}\lambda_{ip}h_{ip}^{(2)}\frac{1}{\Delta} + \sum_{i=1}^{N}\sum_{p=1}^{P}\frac{\lambda_{ip}^{(2)} - \lambda_{ip}^2 - \lambda_{ip}}{2(1-\rho)}h_{ip}^2\frac{1}{\Delta} + \rho[\frac{s^{(2)}}{2s}\cdot\frac{1}{\Delta} - \frac{1}{2}] \\
&+ c\{\frac{1}{2}[\rho^2 - \sum_{i=1}^{N}\rho_i^2] + \sum_{i\in g\cup 1l\cup 1d}\rho_i^2 + \sum_{i\in 1l}\sum_{p=1}^{P}\frac{\lambda_{ip}^{(2)} - (1-\Delta)\lambda_{ip}^2 - \lambda_{ip}}{2\lambda_{ip}}\rho_{ip} \\
&- \sum_{i\in 1d}\sum_{p=1}^{P}[\rho_{ip}^2 + \frac{\lambda_{ip}^2 h_{ip}^{(2)}}{2(1-\rho_{ip})}\{\sum_{u=p+1}^{P}\rho_{iu}(3-\rho_{ip}) + (1-\rho_{ip})\rho_{ip}\} \\
&+ \lambda_{ip}(1-\rho_{ip})\sum_{u=p+1}^{P}\frac{\lambda_{iu}\rho_{iu}h_{iu}^{(2)}}{2(1-\rho_{iu})^2}(2-\rho_{iu}) + R_{ip}(\text{disc})]\}\frac{1}{\Delta},
\end{aligned} \tag{4.1}$$

where

$$c = \frac{s}{(1-\rho)} \text{ and}$$

$$\begin{aligned}
R_{ip}(\text{disc}) = \frac{1}{2}\{&\frac{\rho_{ip}h_{ip}}{1-\rho_{ip}}(\lambda_{ip}^{(2)} - \lambda_{ip}^2 - \lambda_{ip})(1 - \rho_{ip} + \rho_{ip}\sum_{u=p+1}^{P}\rho_{iu}) \\
&- \frac{\lambda_{ip}^{(2)} - (1-\Delta)\lambda_{ip}^2 - \lambda_{ip}}{\lambda_{ip}}(1-\rho_{ip})\rho_{ip} \\
&+ \lambda_{ip}(1-\rho_{ip})[\sum_{u=p+1}^{P}(\lambda_{iu}^{(2)} - \lambda_{iu}^2 - \lambda_{iu})(1 - \rho_{iu} + \rho_{iu}^2)\frac{h_{ip}}{\lambda_{iu}(1-\rho_{iu})^2}] \\
&+ \lambda_{ip}\sum_{u=p+1}^{P}\rho_{iu}[\rho_{ip}^2 - (1-\rho_{ip})(1+\rho_{iu})]\Delta\}.
\end{aligned}$$

We are now in a position to consider a continuous-time batch Poisson input multi-queue priority system. In this case, the discretization of the input (batch Poisson) process forms a

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

batch Bernoulli process for any slot with length Δ; see Powell & Avi-Itzhak [16]. Multiplying both sides of (4.1) by Δ and taking the limit as Δ tends to 0, we obtain the following theorem.

Theorem 4.1 For a continuous-time $M^X/GI/1$ type multi-queue priority system with mixed exhaustive, gated, 1-limited, and 1-decrementing service stations, we have

$$\begin{aligned}
&\sum_{i\in e\cup g}\sum_{p=1}^{P}\rho_{ip}w_{ip}+\sum_{i\in 1l}\sum_{p=1}^{P}[\rho_{ip}-c\{\lambda_{ip}\sum_{u=p}^{P}\rho_{iu}+\rho_{ip}\sum_{u=1}^{p-1}\lambda_{iu}\}]w_{ip}\\
&+\sum_{i\in 1d}\sum_{p=1}^{P}[\rho_{ip}-c\{\lambda_{ip}(1-\rho_{ip})\sum_{u=p}^{P}\rho_{iu}+\rho_{ip}\sum_{u=1}^{p-1}\lambda_{iu}(1-\rho_{iu})\}]w_{ip}\\
&=\frac{\rho}{2(1-\rho)}\sum_{i=1}^{N}\sum_{p=1}^{P}\lambda_{ip}h_{ip}^{(2)}+\sum_{i=1}^{N}\sum_{p=1}^{P}\frac{\lambda_{ip}^{(2)}-\lambda_{ip}^2-\lambda_{ip}}{2(1-\rho)}h_{ip}^2+\rho\frac{s^{(2)}}{2s}+c\{\frac{1}{2}[\rho^2-\sum_{i=1}^{N}\rho_i^2]\\
&+\sum_{i\in g\cup 1l\cup 1d}\rho_i^2+\sum_{i\in 1l}\sum_{p=1}^{P}\frac{\lambda_{ip}^{(2)}-\lambda_{ip}^2-\lambda_{ip}}{2\lambda_{ip}}\rho_{ip}\\
&-\sum_{i\in 1d}\sum_{p=1}^{P}[\rho_{ip}^2+\frac{\lambda_{ip}^2h_{ip}^{(2)}}{2(1-\rho_{ip})}\{\sum_{u=p+1}^{P}\rho_{iu}(3-\rho_{ip})+(1-\rho_{ip})\rho_{ip}\}\\
&+\lambda_{ip}(1-\rho_{ip})\sum_{u=p+1}^{P}\frac{\lambda_{iu}\rho_{iu}h_{iu}^{(2)}}{2(1-\rho_{iu})^2}(2-\rho_{iu})+R_{ip}(\mathrm{cont})]\},
\end{aligned}\tag{4.2}$$

where

$$c=\frac{s}{(1-\rho)}\ \text{and}$$

$$\begin{aligned}
R_{ip}(\mathrm{cont}):=\frac{1}{2}\{&\frac{\rho_{ip}h_{ip}}{1-\rho_{ip}}(\lambda_{ip}^{(2)}-\lambda_{ip}^2-\lambda_{ip})(1-\rho_{ip}+\rho_{ip}\sum_{u=p+1}^{P}\rho_{iu})\\
&-\frac{\lambda_{ip}^{(2)}-\lambda_{ip}^2-\lambda_{ip}}{\lambda_{ip}}(1-\rho_{ip})\rho_{ip}\\
&+\lambda_{ip}(1-\rho_{ip})[\sum_{u=p+1}^{P}(\lambda_{iu}^{(2)}-\lambda_{iu}^2-\lambda_{iu})(1-\rho_{iu}+\rho_{iu}^2)\frac{h_{ip}}{\lambda_{iu}(1-\rho_{iu})^2}]\}.
\end{aligned}$$

Remark 4.1 Equation (4.2) is a slight extension of the Poisson input results by Fournier & Rosberg [8] and Shimogawa & Takahashi [17]. For a Poisson input ($\lambda_{ip}^{(2)}=\lambda_{ip}^2+\lambda_{ip}, 1\le p\le P$) system, (4.2) reduces to the main result in [17]. For a Poisson input single-class (non-priority) system, (4.2) agrees with the result of Boxma & Groenendijk [1] if we set $P=1$, and $\lambda_{i1}^{(2)}=\lambda_{i1}^2+\lambda_{i1}$. However, there is a discrepancy between our result and that in Fournier & Rosberg [8]. We understand that a calculation error is involved in [8], because Eq. (28) in [8] (should but) does not reduce to Eq. (3.21) in [1] for the Poisson input single-class system.□

5. Concluding remarks

We have derived the pseudo-conservation law for a discrete-time $Geom^X/GI/1$ type multi-queue priority system with mixed exhaustive, gated, 1-limited and 1-decrementing service stations. Taking the discrete-time result as the slot length tends to zero has enabled

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

us to obtain the continuous-time result for an $M^X/GI/1$ type multi-queue priority system. It is left for future work to derive a pseudo-conservation law for a more general (e. g., Markov-modulated batch Bernoulli process [21]) input system, since mean performance measures were shown to be influenced more by variances of the input processes than by those of the service times in the literature [11, 12, 15]. It is also worthwhile to study a distributional form of the pseudo-conservation law (a distributional relationship between the waiting time and the input random variables batch size, service time and switch-over time).

Appendix.

We will treat more general (stationary and ergodic) discrete-time $G^X/GI/1$ type multi-queue priority system than the $Geom^X/GI/1$ type system described in Section 2. The batch inter-arrival times are generally distributed (but can be correlated). The service time and switch-over time are respectively assumed to be independent and identically distributed (i. i. d.) and mutually independent of the arrival processes. For any $(i,p)(1 \leq i \leq N; 1 \leq p \leq P)$, we consider a class (i,p) queue in the $G^X/GI/1$ type multi-queue priority system. We introduce the following notation.

$\{\tau_i(k) \mid k = 0,1,2,\cdots\}$: sequence of the server arrival time points at station i,

$$\text{where } \tau_i(0) = 0$$

$C_i(k) := \tau_i(k+1) - \tau_i(k); k = 0,2,\cdots$ (cycle-time sequence)

$\Xi_i(k) := \sum_{u=0}^{k-1} C_i(u)$. We assume the cycle-time sequence $\{C_i(k)\}$ is stationary and ergodic. This assumption is valid for the (batch-renewal input [12]) $GI^X/GI/1$ type and the (Markov-modulated batch Bernoulli process input [21]) $MBBP/GI/1$ type systems. Applying the ergodic theorem gives

$$\lim_{k\to\infty} \frac{\Xi_i(k)}{k} = \lim_{k\to\infty} \frac{\sum_{u=0}^{k-1} C_i(u)}{k} = c_i, \tag{A.1}$$

where c_i denotes the mean cycle time at station i. This c_i is independent of station index, as shown below.

A) Proof of equation (2.1)

We also introduce the following notation.

$TS_i(k)$: total switch-over time of the server during $C_i(k)$

$\Xi B_i(k)$: subset of time interval $[0, \Xi_i(k)]$ during which the server is busy and

$\Xi I_i(k)$: subset of time interval $[0, \Xi_i(k)]$ during which the server switches stations.

The sequence $\{TS_i(k)\}$ is independent, and hence, the strong law of large numbers gives

$$\lim_{k\to\infty} \frac{\Xi I_i(k)}{k} = \lim_{k\to\infty} \frac{\sum_{u=0}^{k-1} TS_i(u)}{k} = s, \tag{A.2}$$

where s is the mean total switch-over time of the server during a cycle.

Now that the server either serves customers or switches stations at an arbitrary time, we have

$$\Xi_i(k) = \Xi B_i(k) + \Xi I_i(k). \tag{A.3}$$

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

Applying Little's law to the subsystem composed of only the server facility (without any waiting room for an individual class), we have

$$\lim_{k\to\infty}\frac{\Xi B_i(k)}{\Xi_i(k)} = Pr[\text{the server is busy}] = \rho. \tag{A.4}$$

Multiplying both sides of (A.3) by $1/k$, and taking the limit as k tends to infinity, we have from equations (A.1), (A.2), and (A.4)

$$\begin{aligned} c_i &= \lim_{k\to\infty}\frac{\Xi_i(k)}{k} = \lim_{k\to\infty}\frac{\Xi_i(k)}{k}\frac{\Xi B_i(k)}{\Xi_i(k)} + \lim_{k\to\infty}\frac{\Xi I_i(k)}{k} \\ &= c_i\rho + s. \end{aligned} \tag{A.5}$$

Equation (A.5) yields

$$c_i = \frac{s}{1-\rho},$$

showing that mean cycle time c_i does not depend on station index. We have denoted this mean cycle time by c in Section 2. This completes the proof of equation (2.1).□

B) Proof of equation (2.3)

We also need the following notation.

$Q_{ip}(n)$: number of class-(i,p) customers at time point $n(n = 0, \pm1, \pm2, \cdots)$
$\Sigma_{ip}(k)$: number of class-(i,p) customers who were served during $[0, \Xi_i(k)]$
$\Psi_{ip}(k)$: number of class-(i,p) customers who arrived during $[0, \Xi_i(k)]$

and

$\Theta_{ip}(k) := Q_{ip}(\tau_i(k)+1)$ (queue length found by the server upon its k-th arrival)

Noting that for the 1-limited service station, the maximum number of customers that can be served per cycle is only one, we have

$$\Theta_{ip}(k) = \Theta_{ip}(0) + \Psi_{ip}(k) - \Sigma_{ip}(k), \tag{A.6}$$

and

$$Pr[A_{ip}] = \lim_{k\to\infty}\frac{\Sigma_{ip}(k)}{k}. \tag{A.7}$$

Assume the stationary condition:

$$\lim_{k\to\infty}\frac{\Theta_{ip}(k)}{k} = 0. \tag{A.8}$$

Multiplying both sides of (A.6) by $1/k$, taking the limit as k tends to infinity, and applying the ergodic theorem, we have from equations (A.1) and (A.6) through (A.8)

$$\begin{aligned} Pr[A_{ip}] &= \lim_{k\to\infty}\frac{\Sigma_{ip}(k)}{k} = \lim_{k\to\infty}\frac{\Psi_{ip}(k)}{k} = \lim_{k\to\infty}\frac{\Psi_{ip}(k)}{\Xi_i(k)}\frac{\Xi_i(k)}{k} \\ &= \lambda_{ip}c, \end{aligned}$$

proving (2.3).□

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

References

[1] C. Bisdikian, A note on the conservation law for queues with batch arrivals, *IEEE Trans. Commun.*, **41** (1993) 832–835.

[2] O. J. Boxma and W. P. Groenendijk, Pseudo-conservation laws in cyclic service systems, *J. Appl. Prob.*, **24** (1987) 949–964.

[3] O. J. Boxma and W. P. Groenendijk, Waiting times, in discrete-time cyclic service systems, *IEEE Trans. Commun.*, **36** (1988) 164–170.

[4] K. Change and D. Sandhu, Pseudo-conservation laws in cyclic server, multi-queue systems with a class of limited service policies, *IEEE INFOCOM* (1990) 260–267.

[5] J. Chiarawongse and M. M. Srinivasan, On pseudo-conservation laws for the cyclic server system with compound Poisson arrivals, *Operations Res. Letters*, **10** (1991) 453–459.

[6] D. Everitt, A note on pseudo-conservation laws for cyclic service systems with limited service disciplines, *IEEE Trans. Commun.*, **37** (1989) 781–783.

[7] Y. Fukagawa, S. Murakami, and S. Yoshida, An approximate analysis for a multiqueue with a non-preemptive priority and cyclic service, *Trans. IEICE*, **J70-A** (1987) 1351–1354 (in Japanese).

[8] L. Fournier and Z. Rosberg, Expected waiting times in polling systems under priority disciplines, *Queueing Systems*, **9** (1991) 419–440.

[9] S. Halfin: "Batch delays versus customer delays," *Bell Syst. Tech. J.*, **62** (1983) 2011–2015.

[10] D. Karvelas and A. Leon-Garcia, Performance of integrated packet voice/data token-passing rings, *IEEE J. Selected Areas in Commun.* **SAC-4** (1986) 823–832.

[11] G. Kimura and Y. Takahashi, Diffusion approximation for a token ring system with non-exhaustive service, *IEEE J. Selected Areas in Commun.*, **SAC-4** (1986) 794–801.

[12] G. Kimura and Y. Takahashi, An approximation for a token ring system with priority classes of messages, *J. Information Processing*, **10** (1987) 86–91.

[13] E. M. Klimko and M. F. Neuts, The single-server queue in discrete time — Numerical analysis II, *Naval Res. Logist. Quart.*, **20** (1973) 305–317.

[14] L. Kleinrock, *Queueing Systems, Vol. 2* (Wiley, New York 1976).

[15] P. J. Kuehn, Multi-queue systems with non-exhaustive cyclic service, *Bell Syst. Tech. J.*, **58** (1979) 671–698.

[16] B. A. Powell and B. Avi-Itzhak, Queueing systems with enforced idle time, *Operat. Res.*, **15** (1967) 1145–1156.

[17] S. Shimogawa and Y. Takahashi, A note on the pseudo-conservation law for a multi-queue with local priority, *Queueing Systems*, **11** (1992) 145–151.

[18] H. Takagi, *Analysis, of Polling Systems* (The MIT Press, Cambridge, Mass, 1986).

[19] H. Takagi, *Queueing Analysis, Vol. 1: Vacation and Priority Systems, Part 1* (North-Holland, Amsterdam, 1991).

[20] Y. Takahashi and O. Hashida, Delay analysis of discrete-time priority queue with structured inputs, *Queueing Systems*, **8** (1991) 149–164.

[21] T. Tsuchiya and Y. Takahashi, On discrete-time single-server queues with Markov modulated batch Bernoulli input and finite capacity, *JORSJ*, **36** (1993) 29–45.

Yoshitaka Takahashi
Performance Evaluation Research Group
Network Traffic Laboratory
NTT Telecommunication Networks Laboratories
3–9–11 Midori-cho, Musashino-shi
Tokyo 180, Japan

B. Krishna Kumar
Department of Mathematics
College of Engineering
Anna University, Guindy
Madras-600025, India

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.