No. **613**

Analysis of Preemptive Loss Priority Queues

by

Hideaki Takagi and Yoshio Kodera

December 1994

Analysis of Preemptive Loss Priority Queues

Hideaki Takagi
Institute of Socio-Economic Planning, University of Tsukuba
1-1-1 Tennoudai, Tsukuba-shi, Ibaraki 305, Japan
and
Yoshio Kodera
College of Socio-Economic Planning, University of Tsukuba
1-1-1 Tennoudai, Tsukuba-shi, Ibaraki 305, Japan

In a queueing system with preemptive loss priority discipline, customers disappear from the system immediately when their service is preempted by the arrival of another customer with higher priority. Such a system can model a case in which old requests of low priority are not worthy of service. This paper is concerned with preemptive loss priority queues in which customers of each priority class arrive in a Poisson process and have general service time distribution. By extending the existing analysis, we obtain the distribution of the number of customers of each class in the system (queue size) at an arbitrary time, and the distribution of the time that a customer of each class stays in the system (sojourn time). While the mean queue size and the mean sojourn time are given explicitly, their second moments can be recursively calculated. We also consider systems with server vacations.

Key Words: Queues, priority queues, preemptive loss priority discipline, queue size, waiting time, sojourn time, completion time, busy period.

1. Introduction

Most chapters in textbooks and papers on priority queues treat systems with nonpreemptive, preemptive resume, and preemptive repeat (identical and different) priority disciplines [Conway et al. 1967 (chap. 8); Gaver 1962; Jaiswal 1968; Takagi 1991 (chap. 3)]. As far as the authors know, only Gnedenko and König [1984 (sec. 6.3)] and Klimow [1977 (sec. 3.2)] show brief results of analysis for *preemptive loss* priority queues. In such queues, customers disappear from the system immediately when their service is preempted by the arrival of another customer with higher priority. They can model a situation in which old requests of low priority are not worthy of service when a new request of service with high priority arrives. It is our observation that the analysis of preemptive loss priority queues is not widely known, partly because the above-mentioned references are in German.

In this paper, we are concerned with preemptive loss priority queues in which customers of each priority class arrive in a Poisson process and have general service time distribution. We elaborate the existing analysis and extend it in several ways. Our new results include (1) the marginal distribution, the mean, and the second moment of the number of customers of each class present in the system (*queue size*) at an arbitrary time, (2) the second moment of the time from the arrival to the service start (*waiting time*) of a customer of each class (the mean waiting time is given in Gnedenko and König [1984 (sec. 6.3)] and Klimow [1977 (sec. 3.2)]), and (3) the distribution and the mean of the waiting time for several server vacation models. We also correct the expression for the distribution of the time that a customer of each class spends in the system (*sojourn time*) given in Gnedenko and König [1984 (sec.6.3.3)].

2. Model and Notation

Specifically, our model is as follows. A queueing system with customers of multiple classes is served by a single server. Customer classes 1 through P are priority classes such that class p has priority over class q if $p < q$. Customers of class p arrive in a Poisson process at rate λ_p, where $p = 1, 2, \ldots, P$. The aggregate arrival rate of customers of class $1, 2, \ldots, p$ is denoted by

$$\lambda_p^+ := \sum_{k=1}^{p} \lambda_k \tag{2.1}$$

The distribution function (DF) and its Laplace-Stieltjes transform (LST) for the service time of a customer of class p are denoted by $B_p(x)$ and $B_p^*(s)$, respectively. The service discipline is preemptive loss as defined in Section 1.

Our objective is to obtain the distribution for the number L_p of customers of class p present in the system (queue size) at an arbitrary time, and the distributions for the waiting time W_p and the sojourn time T_p of a customer of class p. The waiting time W_p is the interval from the arrival to the service start of a customer of class p. The sojourn time T_p is the interval from the arrival to the service termination (either by completion or by preemption) of a customer of class p. As preliminaries, we analyze the *actual service time* $\overline{x}_p$ for a customer of class p, the *completion time* C_p for a customer of class p, and the length Θ_p^+ of a *busy period* generated by customers of class 1 though p. Note the relation

$$T_p = W_p + \overline{x}_p \tag{2.2}$$

where $W + p$ and $\overline{x}_p$ are independent. The completion time C_p is the interval from the start of service to a customer of class p to the first moment after the service termination at which no customers of class 1 through $p-1$ are present in the system.

We consider only the steady state. Because of the preemption mechanism, the arrival of customers of class $p+1$ through P does not influence the queue size of customers of class p and the waiting and sojourn times for a customer of class p. The stability condition for a subsystem consisting of customers of class 1 through p is therefore given by

$$\sum_{k=1}^{p} \lambda_k E[\overline{x}_k] < 1 \tag{2.3}$$

In Section 10, we consider systems in which the server takes vacations when there are no customers in the system at the end of a service completion (*server vacation models*). The vacation generically represents a period in which the server is not available even when there are customers in the queue. In systems with server vacations, the arrival of customers of class $p+1$ through P *does* affect the queue of customers of class p because the vacation process depends on all customers. See, for example, Takagi [1991 (chap. 2 and chap. 3)] for a treatment of server vacation models of nonpriority, nonpreemptive priority, preemptive resume priority, and preemptive repeat priority systems.

3. Actual Service Time

We first consider the service time $\overline{x}_p$ actually received by a customer of class p. Let x_p be the original service time of a customer of class p. If no customers of class $1, 2, \ldots, p-1$

arrive during x_p, then $\bar{x}_p$ equals x_p. Once any customer of class $1, 2, \ldots, p-1$ arrives during x_p, the service is terminated. Therefore, we have

$$\begin{aligned} E[e^{-s\bar{x}_p}|x_p] &= e^{-\lambda_{p-1}^{+}x_p}e^{-sx_p} + \int_0^{x_p} \lambda_{p-1}^{+} e^{-\lambda_{p-1}^{+}x}e^{-sx}dx \\ &= \frac{\lambda_{p-1}^{+} + se^{-(s+\lambda_{p-1}^{+})x_p}}{s+\lambda_{p-1}^{+}} \end{aligned} \tag{3.1}$$

Removing the condition on x_p in (3.1), we get the LST $\overline{X}_p^*(s)$ of the DF for $\bar{x}_p$ as

$$\overline{X}_p^*(s) = \frac{\lambda_{p-1}^{+} + sB_p^*(s+\lambda_{p-1}^{+})}{s+\lambda_{p-1}^{+}} \tag{3.2}$$

From (3.2), we have

$$E[\bar{x}_p] = \frac{1 - B_p^*(\lambda_{p-1}^{+})}{\lambda_{p-1}^{+}} \tag{3.3a}$$

$$E[\bar{x}_p{}^2] = \frac{2[1 - B_p^*(\lambda_{p-1}^{+})]}{(\lambda_{p-1}^{+})^2} - \frac{2E[x_p e^{-\lambda_{p-1}^{+}x_p}]}{\lambda_{p-1}^{+}} \tag{3.3b}$$

$$E[\bar{x}_p{}^3] = \frac{6[1 - B_p^*(\lambda_{p-1}^{+})]}{(\lambda_{p-1}^{+})^3} - \frac{6E[x_p e^{-\lambda_{p-1}^{+}x_p}]}{(\lambda_{p-1}^{+})^2} - \frac{3E[x_p^2 e^{-\lambda_{p-1}^{+}x_p}]}{\lambda_{p-1}^{+}} \tag{3.3c}$$

Note that $\overline{X}_1^*(s) = B_1^*(s)$, because the service of customers of class 1 is never preempted. Accordingly, $E[\bar{x}_1] = b_1$, $E[\bar{x}_1{}^2] = b_1^{(2)}$, and $E[\bar{x}_1{}^3] = b_1^{(3)}$, where b_1, $b_1^{(2)}$, and $b_1^{(3)}$ are the mean, the second moment, and the third moment of the service time for a customer of class 1.

We can condition the actual service time on the disposal of customers. Namely, the LST of the DF for $\bar{x}_p$ for a customer of class p whose service is completed is given by

$$\overline{X}_p^*(s|\text{completed}) = \frac{\displaystyle\int_0^\infty e^{-\lambda_{p-1}^{+}x}e^{-sx}dB_p(x)}{\displaystyle\int_0^\infty e^{-\lambda_{p-1}^{+}x}dB_p(x)} = \frac{B_p^*(s+\lambda_{p-1}^{+})}{B_p^*(\lambda_{p-1}^{+})} \tag{3.4}$$

where $B_p^*(\lambda_{p-1}^{+})$ is the probability that the service time x_p is not preempted. From (3.4), we have

$$E[\bar{x}_p|\text{completed}] = \frac{E[x_p e^{-\lambda_{p-1}^{+}x_p}]}{B_p^*(\lambda_{p-1}^{+})} \tag{3.5}$$

The LST of the DF for $\bar{x}_p$ for a customer whose service is preempted is given by

$$\begin{aligned} \overline{X}_p^*(s|\text{preempted}) &= \frac{\displaystyle\int_0^\infty dB_p(x)\int_0^{x_p} \lambda_{p-1}^{+} e^{-\lambda_{p-1}^{+}x}e^{-sx}dx}{\displaystyle\int_0^\infty dB_p(x)\int_0^{x_p} \lambda_{p-1}^{+} e^{-\lambda_{p-1}^{+}x}dx} \\ &= \frac{\lambda_{p-1}^{+}[1 - B_p^*(s+\lambda_{p-1}^{+})]}{(s+\lambda_{p-1}^{+})[1 - B_p^*(\lambda_{p-1}^{+})]} \end{aligned} \tag{3.6}$$

from which we have

$$E[\overline{x}_p|\text{preempted}] = \frac{1}{\lambda_{p-1}^+} - \frac{E[x_p e^{-\lambda_{p-1}^+ x_p}]}{1 - B_p^*(\lambda_{p-1}^+)} \tag{3.7}$$

4. Completion Time

Let us also consider the completion time C_p for a customer of class p. If no customers of class $1, 2, \ldots, p-1$ arrive during the service time x_p of a customer of class p, C_p equals x_p. If any customer of class $1, 2, \ldots, p-1$ arrives during x_p, C_p is extended by the length equal to Θ_{p-1}^+. Therefore, we have

$$\begin{aligned} E[e^{-sC_p}|x_p] &= e^{-\lambda_{p-1}^+ x_p} e^{-sx_p} + \Theta_{p-1}^+(s) \int_0^{x_p} \lambda_{p-1}^+ e^{-\lambda_{p-1}^+ x} e^{-sx} dx \\ &= e^{-(s+\lambda_{p-1}^+)x_p} + \frac{\lambda_{p-1}^+}{s + \lambda_{p-1}^+} \Theta_{p-1}^+(s)[1 - e^{-(s+\lambda_{p-1}^+)x_p}] \end{aligned} \tag{4.1}$$

Removing the condition on x_p in (4.1), we get

$$C_p^*(s) = B_p^*(s + \lambda_{p-1}^+) + \frac{\lambda_{p-1}^+}{s + \lambda_{p-1}^+} \Theta_{p-1}^+(s)[1 - B_p^*(s + \lambda_{p-1}^+)] \tag{4.2}$$

From (4.2), we have

$$E[C_p] = \left[\frac{1}{\lambda_{p-1}^+} + E[\Theta_{p-1}^+]\right][1 - B_p^*(\lambda_{p-1}^+)] \tag{4.3a}$$

$$\begin{aligned} E[C_p^2] &= \left[\frac{2}{(\lambda_{p-1}^+)^2} + \frac{2E[\Theta_{p-1}^+]}{\lambda_{p-1}^+} + E[(\Theta_{p-1}^+)^2]\right][1 - B_p^*(\lambda_{p-1}^+)] \\ &\quad -2\left[\frac{1}{\lambda_{p-1}^+} + E[\Theta_{p-1}^+]\right] E[x_p e^{-\lambda_{p-1}^+ x_p}] \end{aligned} \tag{4.3b}$$

$$\begin{aligned} E[C_p^3] &= \left[\frac{6}{(\lambda_{p-1}^+)^3} + \frac{6E[\Theta_{p-1}^+]}{(\lambda_{p-1}^+)^2} + \frac{3E[(\Theta_{p-1}^+)^2]}{\lambda_{p-1}^+} + E[(\Theta_{p-1}^+)^3]\right][1 - B_p^*(\lambda_{p-1}^+)] \\ &\quad -3\left[\frac{2}{(\lambda_{p-1}^+)^2} + \frac{2E[\Theta_{p-1}^+]}{\lambda_{p-1}^+} + E[(\Theta_{p-1}^+)^2]\right] E[x_p e^{-\lambda_{p-1}^+ x_p}] \\ &\quad -3\left[\frac{1}{\lambda_{p-1}^+} + E[\Theta_{p-1}^+]\right] E[x_p^2 e^{-\lambda_{p-1}^+ x_p}] \end{aligned} \tag{4.3c}$$

From (3.3) and (4.3), we note that

$$E[C_p] = (1 + \lambda_{p-1}^+ E[\Theta_{p-1}^+])E[\overline{x}_p] \tag{4.4a}$$

$$E[C_p^2] = (1 + \lambda_{p-1}^+ E[\Theta_{p-1}^+])E[\overline{x}_p{}^2] + \lambda_{p-1}^+ E[(\Theta_{p-1}^+)^2]E[\overline{x}_p] \tag{4.4b}$$

$$E[C_p^3] = (1 + \lambda_{p-1}^+ E[\Theta_{p-1}^+])E[\overline{x}_p{}^3] + \frac{3}{2}\lambda_{p-1}^+ E[(\Theta_{p-1}^+)^2]E[\overline{x}_p{}^2] + \lambda_{p-1}^+ E[(\Theta_{p-1}^+)^3]E[\overline{x}_p] \tag{4.4c}$$

5. Busy Period

We next consider $\Theta_p^+(s)$, the LST of the DF for the length Θ_p^+ of a busy period generated by customers of class $1,2,\ldots,p$. The arguments for $\Theta_p^+(s)$ for the preemptive loss priority queues is formally the same as those for the preemptive repeat priority queues given by Chang [1965] and Jaiswal [1968 (sec. IV.7-2)].

Because of the preemptive discipline, the presence of customers of class $p+1,\ldots,P$ does not influence Θ_p^+. It is initiated by a customer of class $1,2,\ldots,p$ that arrives when the server is either idle or serving a customer of class $p+1,\ldots,P$. Furthermore, the set of customers of class $1,2,\ldots,p-1$ can be grouped to obtain the distribution of Θ_p^+. If the arriving customer is of class p (which occurs with probability λ_p/λ_p^+), the busy period Θ_p^+ is equal to a busy period, denoted by Θ_p, for customers of class p that have the completion time C_p as the service time. The LST $\Theta_p^*(s)$ of the DF for Θ_p satisfies the equation

$$\Theta_p^*(s) = C_p^*[s + \lambda_p - \lambda_p\Theta_p^*(s)] \tag{5.1}$$

If the arriving customer is of class $1,2,\ldots,p-1$ (which occurs with probability $\lambda_{p-1}^+/\lambda_p^+$), the busy period Θ_p^+ is equal to the delay cycle with initial delay Θ_{p-1}^+ generated by customers of class p. Therefore, we have

$$\Theta_p^+(s) = \frac{\lambda_p}{\lambda_p^+}\Theta_p^*(s) + \frac{\lambda_{p-1}^+}{\lambda_p^+}\Theta_{p-1}^+[s + \lambda_p - \lambda_p\Theta_p^*(s)] \tag{5.2}$$

From (5.1) and (5.2), we get

$$E[\Theta_p^+] = \frac{\lambda_p E[C_p] + \lambda_{p-1}^+ E[\Theta_{p-1}^+]}{\lambda_p^+(1 - \lambda_p E[C_p])} \tag{5.3a}$$

$$E[(\Theta_p^+)^2] = \frac{\lambda_p(1 + \lambda_{p-1}^+ E[\Theta_{p-1}^+])E[C_p^2]}{\lambda_p^+(1 - \lambda_p E[C_p])^3} + \frac{\lambda_{p-1}^+ E[(\Theta_{p-1}^+)^2]}{\lambda_p^+(1 - \lambda_p E[C_p])^2} \tag{5.3b}$$

$$\begin{aligned} E[(\Theta_p^+)^3] &= \frac{\lambda_p(1 + \lambda_{p-1}^+ E[\Theta_{p-1}^+])}{\lambda_p^+}\left[\frac{E[C_p^3]}{(1 - \lambda_p E[C_p])^4} + \frac{3\lambda_p(E[C_p^2])^2}{(1 - \lambda_p E[C_p])^5}\right] \\ &+ \frac{\lambda_{p-1}^+}{\lambda_p^+}\left[\frac{E[(\Theta_{p-1}^+)^3]}{(1 - \lambda_p E[C_p])^3} + \frac{3\lambda_p E[(\Theta_{p-1}^+)^2]E[C_p^2]}{(1 - \lambda_p E[C_p])^4}\right] \end{aligned} \tag{5.4c}$$

By combining (5.3) with (4.3), we can find the moments for Θ_p^+ recursively.

From (4.4a) and (5.3a), we get the relation

$$\frac{1}{1 + \lambda_{p-1}^+ E[\Theta_{p-1}^+]} - \frac{1}{1 + \lambda_p^+ E[\Theta_p^+]} = \lambda_p E[\overline{x}_p] \tag{5.4}$$

which is a recurrence relation with respect to p. Thus we get

$$E[\Theta_p^+] = \frac{\rho_p^+}{\lambda_p^+(1 - \rho_p^+)} \tag{5.5}$$

where

$$\rho_p^+ := \sum_{k=1}^{p} \lambda_k E[\overline{x}_k] = \lambda_1 b_1 + \sum_{k=2}^{p} \frac{\lambda_k}{\lambda_{k-1}^+}[1 - B_k^*(\lambda_{k-1}^+)] \tag{5.6}$$

Substituting (5.5) into (4.3a), we get

$$E[C_p] = \frac{E[\overline{x}_p]}{1-\rho_{p-1}^+} \tag{5.7}$$

According to Gnedenko and König [1984 (sec. 6.3)] and Klimow [1977 (sec. 3.2)], we have

$$E[(\Theta_p^+)^2] = \frac{\lambda_1 b_1^{(2)} + \sum_{k=2}^{p} \lambda_k (1-\rho_{k-1}^+) E[\overline{x}_k{}^2]}{\lambda_p^+ (1-\rho_p^+)^3} \tag{5.8}$$

Using (5.5) in (4.4b), we get

$$\begin{aligned} E[C_p^2] &= \lambda_{p-1}^+ E[(\Theta_{p-1}^+)^2] E[\overline{x}_p] + \frac{E[\overline{x}_p{}^2]}{1-\rho_{p-1}^+} \\ &= \frac{\left(\lambda_1 b_1^{(2)} + \sum_{k=2}^{p-1} \lambda_k (1-\rho_{k-1}^+) E[\overline{x}_k{}^2]\right) E[\overline{x}_p]}{(1-\rho_{p-1}^+)^3} + \frac{E[\overline{x}_p{}^2]}{1-\rho_{p-1}^+} \end{aligned} \tag{5.9}$$

The third moments $E[(\Theta_p^+)^3]$ and $E[C_p^3]$ can be calculated recursively by (4.4c) and (5.3c).

6. Queue Size

We proceed to find the queue size distribution of customers of class p, following the treatment of Chang [1965] for the preemptive repeat priority queues. Let $\Pi_p'(z)$ be the PGF of the number of customers of class p present in the system at the Markov points embedded at service completion times of customers of class p and at the ending times of busy periods generated by special customers of class $1, 2, \ldots, p-1$. These special customers of class $1, 2, \ldots, p-1$ are those that arrive and find no customers of class p in the system and thus initiate a busy period Θ_{p-1}^+. Considering the events that occur between two successive Markov points, we have

$$\Pi_p'(z) = \left[\frac{\Pi_p'(z) - \Pi_p'(0)}{z} + \frac{\lambda_p}{\lambda_p^+}\Pi_p'(0)\right] C_p^*(\lambda_p - \lambda_p z) + \frac{\lambda_{p-1}^+}{\lambda_p^+}\Pi_p'(0)\Theta_{p-1}^+(\lambda_p - \lambda_p z) \tag{6.1}$$

Solving (6.1) for $\Pi_p'(z)$, we get

$$\Pi_p'(z) = \Pi_p'(0) \frac{(\lambda_p z - \lambda_p^+) C_p^*(\lambda_p - \lambda_p z) + \lambda_{p-1}^+ z \Theta_{p-1}^+(\lambda_p - \lambda_p z)}{\lambda_p^+ [z - C_p^*(\lambda_p - \lambda_p z)]} \tag{6.2}$$

From the normalization condition $\Pi_p'(1) = 1$, we can determine the unknown $\Pi_p'(0)$ as

$$\Pi_p'(0) = \frac{1 - \lambda_p E[C_p]}{1 - \frac{\lambda_{p-1}^+}{\lambda_p^+}\lambda_p E[C_p] + \frac{\lambda_p}{\lambda_p^+}\lambda_{p-1}^+ E[\Theta_{p-1}^+]} \tag{6.3}$$

Let $\Pi_p(z)$ be the PGF for the number of customers of class p in the system (queue size) at the departure time of a customer of class p. By considering only those cases when the

Markov point is the departure of a customer of class p in the events that led to the r.h.s. of (6.1), we can express $\Pi_p(z)$ as

$$\Pi_p(z) = \frac{\left[\dfrac{\Pi_p'(z) - \Pi_p'(0)}{z} + \dfrac{\lambda_p}{\lambda_p^+}\Pi_p'(0)\right]\overline{X}_p^*(\lambda_p - \lambda_p z)}{1 - \Pi_p'(0) + \dfrac{\lambda_p}{\lambda_p^+}\Pi_p'(0)} \tag{6.4}$$

Substituting (6.2) and (6.3) into (6.4), we get

$$\begin{aligned} \Pi_p(z) &= \frac{(1-\lambda_p E[C_p])[\lambda_p z - \lambda_p^+ + \lambda_{p-1}^+ \Theta_{p-1}^+(\lambda_p - \lambda_p z)]\overline{X}_p^*(\lambda_p - \lambda_p z)}{\lambda_p(1 + \lambda_{p-1}^+ E[\Theta_{p-1}^+])[z - C_p^*(\lambda_p - \lambda_p z)]} \\ &= \frac{(1-\rho_p^+)[\lambda_p z - \lambda_p^+ + \lambda_{p-1}^+ \Theta_{p-1}^+(\lambda_p - \lambda_p z)]\overline{X}_p^*(\lambda_p - \lambda_p z)}{\lambda_p[z - C_p^*(\lambda_p - \lambda_p z)]} \end{aligned} \tag{6.5}$$

Note that $\Pi_p(z)$ in (6.5) is also the PGF for the number L_p of customers of class p in the system at an arbitrary time. This comes from PASTA (Poisson arrivals see time averages) and Burke's theorem on the process with unit jumps (Cooper [1981 (sec. 3.2 and sec. 5.3)]) applied to the number of customers of class p present in the system.

The mean queue size of customers of class p at an arbitrary time is given by

$$\begin{aligned} E[L_p] &= \frac{\lambda_p^2 E[C_p^2]}{2(1-\lambda_p E[C_p])} + \frac{\lambda_p \lambda_{p-1}^+ E[(\Theta_{p-1}^+)^2]}{2(1+\lambda_{p-1}^+ E[\Theta_{p-1}^+])} + \lambda_p E[\overline{x}_p] \\ &= \frac{\lambda_p \lambda_p^+ (1-\rho_p^+)^2 E[(\Theta_p^+)^2]}{2(1-\rho_{p-1}^+)} + \lambda_p E[\overline{x}_p] \end{aligned} \tag{6.6}$$

where we have used (5.3b) and (5.7). Substituting (5.8) into (6.6) for $E[(\Theta_p^+)^2]$, we get

$$E[L_p] = \frac{\lambda_p\left(\lambda_1 b_1^{(2)} + \sum_{k=2}^{p} \lambda_k (1-\rho_{k-1}^+) E[\overline{x}_k{}^2]\right)}{2(1-\rho_{p-1}^+)(1-\rho_p^+)} + \lambda_p E[\overline{x}_p] \tag{6.7}$$

The second moment of the queue size of customers of class p at an arbitrary time is given by

$$E[L_p^2] = \lambda_p^2 E[T_p^2] \tag{6.8}$$

where $E[T_p^2]$ is the second moment of the sojourn time of a customer of class p, to be given in (8.8).

7. Waiting Time

If $W_p^*(s)$ denotes the LST of the DF for the waiting time of a customer of class p, we have the relation

$$\Pi_p(z) = W_p^*(\lambda_p - \lambda_p z)\overline{X}_p^*(\lambda_p - \lambda_p z) \tag{7.1}$$

The r.h.s. of this equation is the PGF for the number of customers of class p that arrived while the customer of class p was in the system. From (6.5) and (7.1), we get

$$\begin{aligned} W_p^*(s) &= \frac{(1-\lambda_p E[C_p])[s+\lambda_{p-1}^+ - \lambda_{p-1}^+ \Theta_{p-1}^+(s)]}{(1+\lambda_{p-1}^+ E[\Theta_{p-1}^+])[s-\lambda_p+\lambda_p C_p^*(s)]} \\ &= \frac{(1-\rho_p^+)[s+\lambda_{p-1}^+ - \lambda_{p-1}^+ \Theta_{p-1}^+(s)]}{s-\lambda_p+\lambda_p C_p^*(s)} \end{aligned} \tag{7.2}$$

which has the same form for the preemptive resume and repeat priority systems [Chang 1965; Jaiswal 1968 (sec. IV.7-4)]. We note that $W_p^*(s)$ in (7.2) can also be derived from

$$\begin{aligned} W_p^*(s) &= 1-\rho_p^+ + \rho_{p-1}^+ \cdot \frac{(1-\lambda_p E[C_p])[1-\Theta_{p-1}^+(s)]}{E[\Theta_{p-1}^+][s-\lambda_p+\lambda_p C_p^*(s)]} \\ &+\lambda_p E[\overline{x}_p] \cdot \frac{(1-\lambda_p E[C_p])[1-C_p^*(s)]}{E[C_p][s-\lambda_p+\lambda_p C_p^*(s)]} \end{aligned} \tag{7.3}$$

From (7.2), we have the mean waiting time

$$E[W_p] = \frac{\lambda_p E[C_p^2]}{2(1-\lambda_p E[C_p])} + \frac{\lambda_{p-1}^+ E[(\Theta_{p-1}^+)^2]}{2(1+\lambda_{p-1}^+ E[\Theta_{p-1}^+])} \tag{7.4}$$

By comparing (5.3b) and (7.4), we get

$$E[W_p] = \frac{\lambda_p^+ (1-\lambda_p E[C_p])^2 E[(\Theta_p^+)^2]}{2(1+\lambda_{p-1}^+ E[\Theta_{p-1}^+])} = \frac{\lambda_p^+ (1-\rho_p^+)^2 E[(\Theta_p^+)^2]}{2(1-\rho_{p-1}^+)} \tag{7.5}$$

From (7.4) and (7.5), we obtain a recursive relation [Conway et al. 1967 (sec. 8-7)]

$$E[W_p] = \frac{\lambda_p E[C_p^2]}{2(1-\lambda_p E[C_p])} + \frac{E[W_{p-1}]}{1-\lambda_{p-1} E[C_{p-1}]} \tag{7.6}$$

By substituting (5.8) into (7.5), we get

$$E[W_p] = \frac{\lambda_1 b_1^{(2)} + \sum_{k=2}^{p} \lambda_k (1-\rho_{k-1}^+) E[\overline{x}_k{}^2]}{2(1-\rho_{p-1}^+)(1-\rho_p^+)} \tag{7.7}$$

The second moment of the waiting time of a customer of class p is given by

$$\begin{aligned} E[W_p^2] &= \frac{\lambda_p E[C_p^3]}{3(1-\lambda_p E[C_p])} + \frac{(\lambda_p E[C_p^2])^2}{2(1-\lambda_p E[C_p])^2} \\ &+\frac{\lambda_{p-1}^+ E[(\Theta_{p-1}^+)^3]}{3(1+\lambda_{p-1}^+ E[\Theta_{p-1}^+])} + \frac{\lambda_p \lambda_{p-1}^+ E[C_p^2] E[(\Theta_{p-1}^+)^2]}{2(1-\lambda_p E[C_p])(1+\lambda_{p-1}^+ E[(\Theta_{p-1}^+])} \end{aligned} \tag{7.8}$$

8. Sojourn Time

The LST of the DF for the sojourn time T_p of a customer of class p whose service is completed is given by

$$\begin{aligned} T_p^*(s|\text{completed}) &= W_p^*(s)\overline{X}_p^*(s|\text{completed}) \\ &= \frac{(1-\rho_p^+)[s+\lambda_{p-1}^+ - \lambda_{p-1}^+\Theta_{p-1}^+(s)]}{s-\lambda_p+\lambda_p C_p^*(s)} \cdot \frac{B_p^*(s+\lambda_{p-1}^+)}{B_p^*(\lambda_{p-1}^+)} \end{aligned} \tag{8.1}$$

which yields

$$E[T_p|\text{completed}] = \frac{\lambda_1 b_1^{(2)} + \sum_{k=2}^p \lambda_k (1-\rho_{k-1}^+)E[\overline{x}_k{}^2]}{2(1-\rho_{p-1}^+)(1-\rho_p^+)} + \frac{E[x_p e^{-\lambda_{p-1}^+ x_p}]}{B_p^*(\lambda_{p-1}^+)} \tag{8.2}$$

The LST of the DF for the sojourn time T_p of a customer of class p whose service is preempted is given by

$$\begin{aligned} T_p^*(s|\text{preempted}) &= W_p^*(s)\overline{X}_p^*(s|\text{preempted}) \\ &= \frac{(1-\rho_p^+)[s+\lambda_{p-1}^+ - \lambda_{p-1}^+\Theta_{p-1}^+(s)]}{s-\lambda_p+\lambda_p C_p^*(s)} \cdot \frac{\lambda_{p-1}^+[1-B_p^*(s+\lambda_{p-1}^+)]}{(s+\lambda_{p-1}^+)[1-B_p^*(\lambda_{p-1}^+)]} \end{aligned} \tag{8.3}$$

which yields

$$E[T_p|\text{preempted}] = \frac{\lambda_1 b_1^{(2)} + \sum_{k=2}^p \lambda_k (1-\rho_{k-1}^+)E[\overline{x}_k{}^2]}{2(1-\rho_{p-1}^+)(1-\rho_p^+)} + \frac{1}{\lambda_{p-1}^+} - \frac{E[x_p e^{-\lambda_{p-1}^+ x_p}]}{1-B_p^*(\lambda_{p-1}^+)} \tag{8.4}$$

The LST $T_p^*(s)$ of the DF for the unconditional sojourn time T_p of a customer of class p is given by †

$$T_p^*(s) = W_p^*(s)\overline{X}_p^*(s) \tag{8.5}$$

Hence we get

$$E[T_p] = \frac{\lambda_1 b_1^{(2)} + \sum_{k=2}^p \lambda_k (1-\rho_{k-1}^+)E[\overline{x}_k{}^2]}{2(1-\rho_{p-1}^+)(1-\rho_p^+)} + E[\overline{x}_p] \tag{8.6}$$

which confirms Little's theorem

$$E[L_p] = \lambda_p E[T_p] \tag{8.7}$$

The second moment of the sojourn time is given by

$$E[T_p^2] = E[W_p^2] + 2E[W_p]E[\overline{x}_p] + E[\overline{x}_p{}^2] \tag{8.8}$$

9. Numerical Examples

For numerical examples, we consider a system with $P = 5$ classes of customers. We assume that the parameters of the arrival and service processes are identical for all classes, and that the service times are exponentially distributed with unit mean. Namely,

$$\lambda_p = \frac{\lambda}{P} \quad ; \quad B_p^*(s) = \frac{1}{1+s} \qquad p = 1, 2, \dots, P \tag{9.1}$$

†Gnedenko and Kōnig [1984 (sec. 6.3.3)] give $W_p^*(s)C_p^*(s)$ as the LST of the DF for T_p, which is incorrect, because the completion time C_p includes a busy period Θ_{p-1}^+ after the service for a customer of class p is preempted and therefore it has disappeared.

where λ is the total arrival rate.

In Figure 1, we plot the mean actual service time $E[\overline{x}_p]$ for $p = 1, 2, \dots, P$ against λ. As noted in Section 2., we see that $E[\overline{x}_1] = 1$, independent of λ, because the service to customers of class 1 is never preempted. For $p \geq 2$, each $E[\overline{x}_p]$ is monotonously decreasing as λ grows. This is because the service is more likely to be preempted at a greater arrival rate. In this example, we also have the relation

$$1 = E[\overline{x}_1] > E[\overline{x}_2] > \cdots > E[\overline{x}_P] \tag{9.2}$$

because the service to customers of lower priority classes are more likely to be preempted by the arrival of a customer of high priority.

In Figure 2, we show the mean waiting time $E[W_p]$ for $p = 1, 2, \dots, P$ against λ. We see that each $E[W_p]$ is monotonously increasing as λ grows as a result of queueing and by the effects from customers of higher priority. We also note the relation

$$E[W_1] < E[W_2] < \cdots < E[W_P] \tag{9.3}$$

which discriminates the customers of different classes.

In Figure 3, we show the mean sojourn time $E[T_p]$ for $p = 1, 2, \dots, P$ against λ. We see again that each $E[T_p]$ is monotonously increasing as λ grows and that

$$E[T_1] < E[T_2] < \cdots < E[T_P] \tag{9.4}$$

The discrimination in the mean sojourn times for customers of different classes is milder than that in the mean waiting times, because the mean sojourn times are favorable (small) for customers of low priority.

10. Vacation Models

Server vacation models of the systems with preemptive loss priority discipline can be handled similarly. We present only the results for the LST $W_p^*(s)$ of the DF and the mean $E[W_p]$ for the waiting time W_p for a customer of class p for three vacation models. The LST $T_p^*(s)$ of the DF for the sojourn time T_p for a customer of class p in these systems is given by (8.5) with $\overline{X}_p^*(s)$ given in (3.2). For the convenience of notation, we define

$$\rho := \sum_{k=1}^{P} \lambda_k E[\overline{x}_k] \quad ; \quad \rho_p^- := \sum_{k=p+1}^{P} \lambda_k E[\overline{x}_k] \tag{10.1}$$

and

$$\sigma_{p-1} := s + \lambda_{p-1}^+ - \lambda_{p-1}^+ \Theta_{p-1}^+(s) \tag{10.2}$$

In the *multiple vacation model*, the server takes vacations if the system is empty at the end of a service. If the server returns from a vacation to find the queue not empty, it starts to work immediately and continues service until the system becomes empty again. If the server returns from a vacation to find no customers waiting, it begins another vacation immediately, and repeat vacations in this manner until it finds at least one customer waiting upon returning from a vacation. The lengths of successive vacations are assumed to be independent and identically distributed, and also independent of the arrival and service processes.

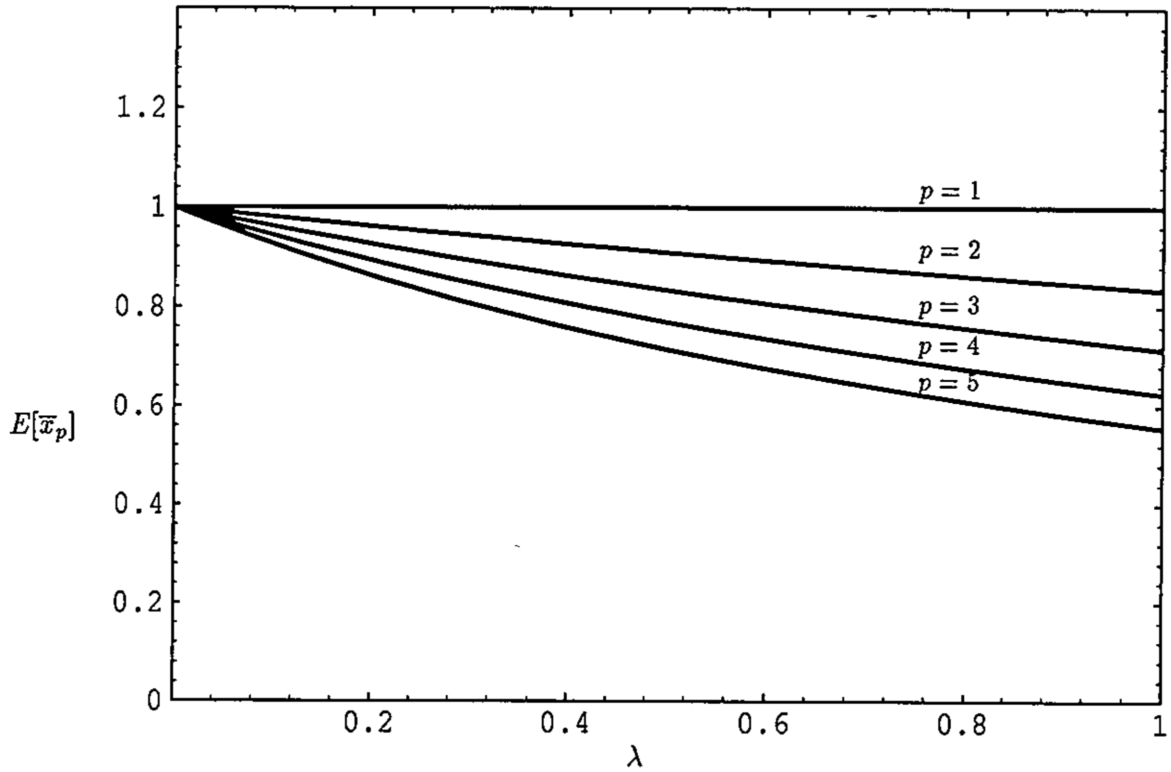

Figure 1. Mean actual service times in a preemptive loss priority queue.

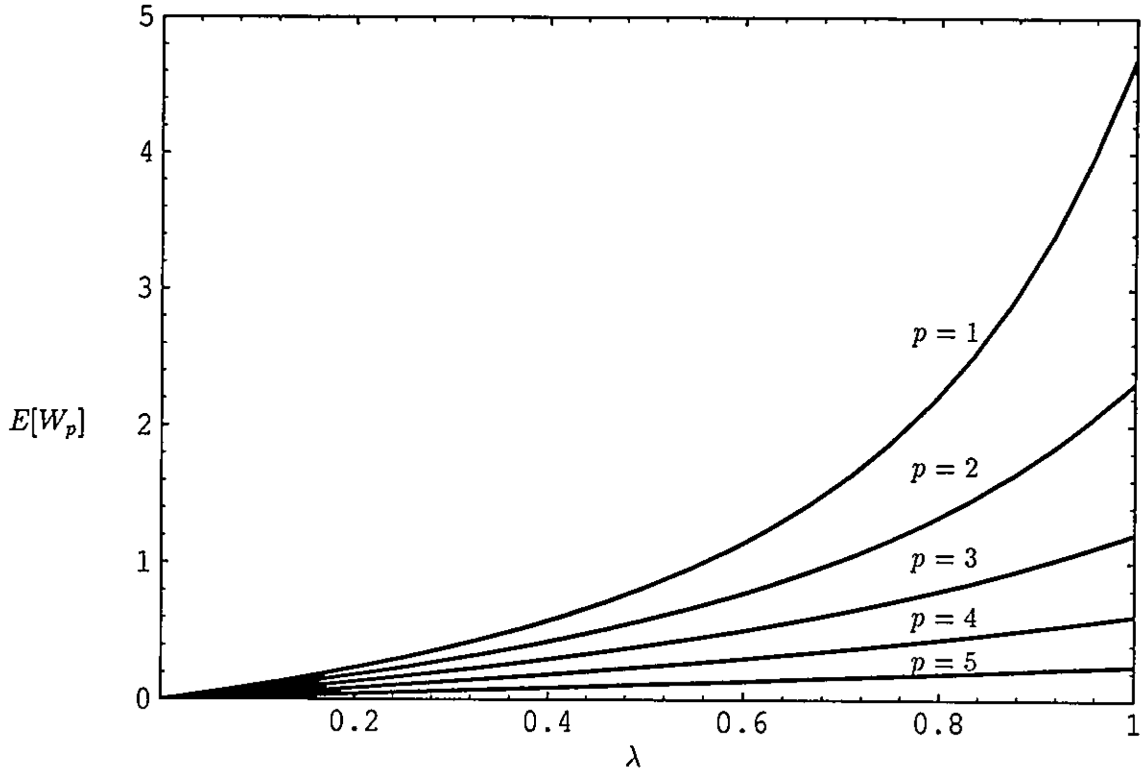

Figure 2. Mean waiting times in a preemptive loss priority queue.

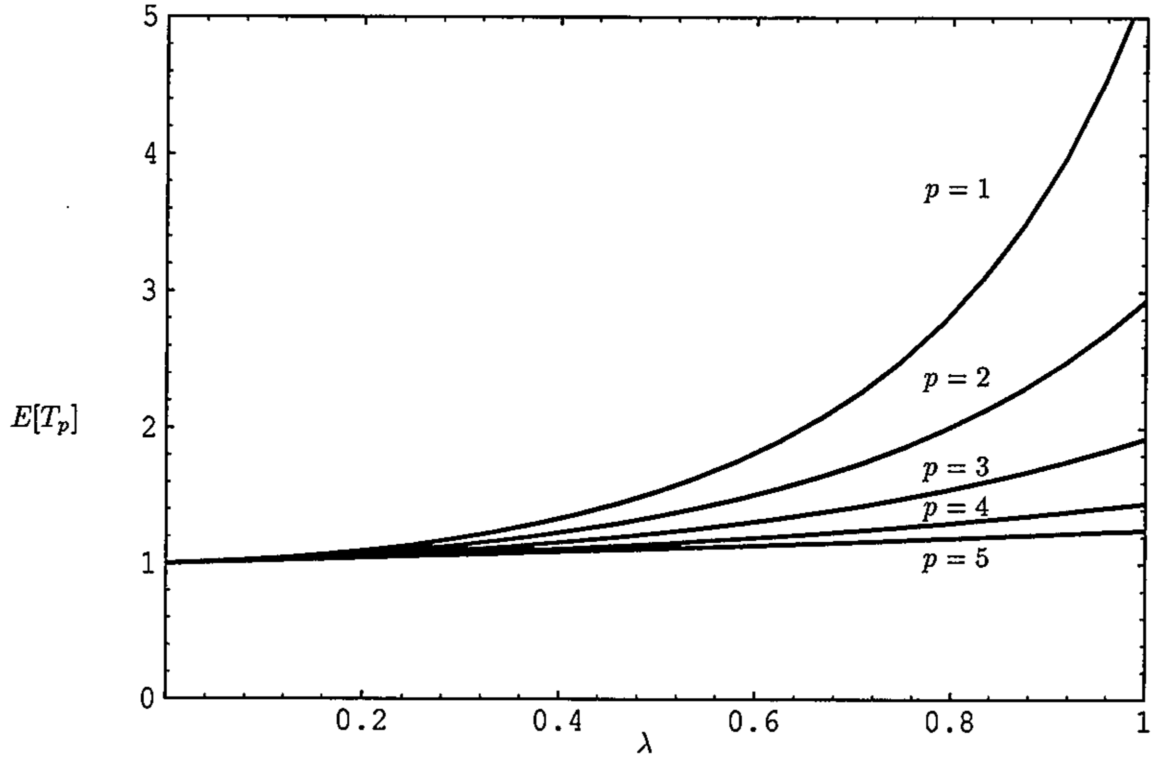

Figure 3. Mean sojourn times in a preemptive loss priority queue.

Let $V^*(s)$ be the LST of the DF for the length V of each vacation. The LST $W_p^*(s)$ of the DF for the waiting time W_p in the preemptive loss priority system is given by

$$W_p^*(s) = \frac{(1-\rho)\dfrac{1 - V^*(\sigma_{p-1})}{E[V]} + \rho_p^- \sigma_{p-1}}{s - \lambda_p + \lambda_p C_p^*(s)} \tag{10.3}$$

The mean waiting time for a customer of class p is given by

$$E[W_p] = \frac{\dfrac{(1-\rho)E[V^2]}{E[V]} + \lambda_1 b_1^{(2)} + \sum_{k=2}^{p} \lambda_k (1 - \rho_{k-1}^+) E[\overline{x}_k{}^2]}{2(1-\rho_{p-1}^+)(1-\rho_p^+)} \tag{10.4}$$

In the *single vacation model*, the server takes exactly one vacation if the system is empty at the end of the service. If the server returns from a vacation to find the queue not empty, it starts to work immediately and continues service until the system becomes empty again. If the server returns from a vacation to find no customers waiting, it becomes idle until a customer arrives, and the starts service immediately. For the single vacation model, we have

$$W_p^*(s) = \frac{(1-\rho)\lambda[1 - V^*(\sigma_{p-1})] + \{(1-\rho_p^+)V^*(\lambda) + \rho_p^- \lambda E[V]\}\sigma_{p-1}}{(V^*(\lambda) + \lambda E[V])[s - \lambda_p + \lambda_p C_p^*(s)]} \tag{10.5}$$

and

$$E[W_p] = \frac{\dfrac{(1-\rho)\lambda E[V^2]}{V^*(\lambda) + \lambda E[V]} + \lambda_1 b_1^{(2)} + \sum_{k=2}^{p} \lambda_k (1 - \rho_{k-1}^+) E[\overline{x}_k{}^2]}{2(1-\rho_{p-1}^+)(1-\rho_p^+)} \tag{10.6}$$

For systems with *setup times*, we assume that a setup time S is needed before starting the service to the first customer that arrives when the system is empty. Let $S^*(s)$ be the LST of the DF for S. We then have

$$W_p^*(s) = \left(\frac{1-\overline{\rho}}{1+\lambda E[S]}\right) \frac{\lambda + (\sigma_{p-1} - \lambda)S^*(\sigma_{p-1})}{s - \lambda_p + \lambda_p C_p^*(s)} + \frac{\rho_p^- \sigma_{p-1}}{s - \lambda_p + \lambda_p C_p^*(s)} \tag{10.7}$$

and

$$E[W_p] = \frac{\dfrac{(1-\rho)(\lambda E[S^2] + 2E[S])}{1+\lambda E[S]} + \lambda_1 b_1^{(2)} + \sum_{k=2}^{p} \lambda_k (1 - \rho_{k-1}^+) E[\overline{x}_k{}^2]}{2(1-\rho_{p-1}^+)(1-\rho_p^+)} \tag{10.8}$$

Acknowledgment

This work is supported in part by the grant for the fiscal year 1994 from the Telecommunications Advancement Foundation.

References

[1] Chang, W. 1965. Preemptive priority queues. *Operations Research,* Vol.13, No.5 (September-October), pp.820-827.

[2] Conway, R. W., Maxwell, W. L., and Miller, L. W. 1967. *Theory of Scheduling.* Addison-Wesley, Reading, Massachusetts.

[3] Cooper, R. B. 1981. *Introduction to Queueing Theory.* Second edition, North-Holland Publishing Company, New York

[4] Gaver, D. P., Jr. 1962. A waiting line with interrupted service, including priorities. *Journal of the Royal Statistical Society,* Series B (Methodological), Vol.24, No.1, pp.73-90.

[5] Gnedenko, B. V., and König, D. (editors) 1984. *Handbuch der Bedienungstheorie, Volume II. Formeln und andere Ergebnisse.* Akademie-Verlag, Berlin.

[6] Jaiswal, N. K. 1968. *Priority Queues.* Academic Press, New York.

[7] Klimow, G. P. 1979. *Bedienungsprozesse.* Birkhäuser Verlag, Basel.

[8] Takagi, H. 1991. *Queueing Analysis, Volume 1: Vacation and Priority Systems.* Elsevier, Amsterdam.