No. 843

Waiting Time Analysis for $M^X/G/1$ Priority Queues with/without Vacations under Ramdom Order of Service Discipline

by

Norikazu Kawasaki, Hideaki Takagi,
Yutaka Takahashi, Sung-Jo Hong
and Toshiharu Hasegawa

November 1999

Waiting Time Analysis for $M^X/G/1$ Priority Queues with/without Vacations under Random Order of Service Discipline

Norikazu Kawasaki *
Hideaki Takagi †
Yutaka Takahashi ‡
Sung-Jo Hong §
Toshiharu Hasegawa ¶

Abstract

We study $M^X/G/1$ nonpreemptive and preemptive-resume priority queues with/without vacations under random order of service (ROS) discipline within each class. By considering the conditional waiting times given the states of the system which an arbitrary message observes upon arrival, we derive the Laplace-Stieltjes transforms of the waiting time distributions and explicitly obtain the first two moments. The relationship for the second moments under ROS and first-come first-served disciplines extends that found by Takács and Fuhrmann for non-priority single arrival queues.

1 Introduction

An M/G/1 queue is a typical model in the fundamental queueing theory. Messages arrive at the buffer of infinite capacity according to a Poisson process, each being served for a generally distributed service time. A single server works continuously until the system becomes empty. So far many variants of the M/G/1 queue have been studied (Kleinrock [18, 19], Takagi [26]).

An $M^X/G/1$ priority queue extends the arrival process as follows; there are P classes of messages indexed as $p = 1, 2, \cdots, P$. Messages arrive in groups whose sizes are generally distributed; groups

*Second Engineering Department, Systems & Electronics Division, Sumitomo Electric Industries, Ltd., Japan
†Institute of Policy and Planning Sciences, University of Tsukuba, Japan
‡Department of Applied Systems Science, Kyoto University, Japan
§Department of Industrial Engineering, Dongguk University, Korea
¶Department of Information Systems and Quantitative Sciences, Nanzan University, Japan

of class p messages arrive according to a Poisson process at rate λ_p. Messages of class p have *priority* over those of class q iff $p < q$. We assume that the service times for each class are independent and identically distributed (i.i.d.).

In this paper we consider two types of priority scheduling. In a *non-preemptive* priority queue, once the service to a message is started, it is not interrupted until it is complete, while in a *preemptive-resume* priority queue, the service to a message of any priority is immediately preempted by the arrival of a batch of a higher priority. The service of the preempted message is resumed from the preempted point when there are no messages of higher priorities.

When the server finishes serving a message and finds the system empty, he waits for the first batch to arrive at the system (*non-vacation model*), or he takes a *vacation* [5] (*vacation model*). We assume that the length of a vacation is i.i.d. We consider two vacation models. If the server returns from a vacation to find no messages waiting, in the *multiple vacation* case, he begins another vacation immediately, and in the *single vacation* case, he waits for the first batch to arrive while keeping the system idle.

Various (single arrival) M/G/1 priority queues under *first-come first-served* (FCFS) discipline within each class have been studied so far by many authors. Cobham [1], Holley [12], Kesten and Runnenburg [16], Miller [21], Welch [29], Takács [25], Jaiswal [13], and Fujiki and Gambe [9] studied models without vacations. Conway, Maxwell and Miller [2], Kella and Yechiali [15] and Shanthikumar [23] studied models with vacations. Takagi and Takahashi [28] treated batch arrival models with/without vacations, which are extensions of the above single arrival models. On the other hand, Durr [6] studied an M/G/1 priority queue without vacations under *last-come first-served* (LCFS) discipline.

Under *random order of service* (ROS) discipline, the next message for service is selected at random from the messages of the highest priority class waiting in the queue. M/G/1 non-priority non-vacation models under ROS were studied by Kingman [17], Takács [24], Conolly [3], and Takagi and Kudoh [27]. Scholl and Kleinrock [22] studied a model with multiple vacations. Kawasaki, Takagi, Takahashi and Hasegawa [14] extended them to batch arrival models, while Durr [7] analyzed a two-class M/M/1 (exponentially distributed service times) priority queue without vacations under ROS. The results in this paper include all these with ROS discipline as special cases. Namely, we study the following six models in a unified manner:

	non-preemptive priority queue	preemptive-resume priority queue
without vacations	**NPNV**	**PRNV**
with multiple vacations	**NPMV**	**PRMV**
with single vacations	**NPSV**	**PRSV**

Our objective is the derivation of the first two moments for the waiting time of an arbitrary message of class p ($p = 1, 2, \cdots, P$) in the above six cases. First, in Section 2 we derive the queue

size distribution for the messages of class p at the beginning of service to a message of class p. In Section 3, we consider the waiting time distributions conditioned on the system state when an arbitrary message of class p arrives. They are used in Section 4 to calculate the first two moments of the waiting time for the arbitrary message of class p. The results are compared for different models in Section 5, and numerical examples are presented in Section 6. In Section 7, we summarize the work, and remark on further results that can be derived straightforwardly from the present results.

Throughout this paper we assume that the system for each case is *unsaturated* (sec. 3.1 in Takagi [26]), namely the existence of the steady state for all classes in the system. (This assumption is removed in a remark made in Section 7.) Furthermore, for convenience's sake, we call the priority classes higher than class p *H-class*, those lower than class p *L-class*, and a set of messages included in a batch a *supermessage*. We define the following notation:

λ_p	arrival rate of batches of class p $(p = 1, \cdots, P)$,
λ	$= \sum_{p=1}^{P} \lambda_p$,
λ_p^+	arrival rate of batches of H-class and class p $(= \sum_{k=1}^{p} \lambda_k)$,
λ_p^-	arrival rate of batches of L-class $(= \sum_{k=p+1}^{P} \lambda_k)$,
V	length of a vacation,
$V^*(s)$	Laplace-Stieltjes transform (LST) of the distribution function (DF) for V,
I	length of an idle period,
$I^*(s)$	LST of the DF for I,
$g_{p,n}$	probability that the batch size of class p is n,
$G_p(z)$	generating function (GF) for $g_{p,n}$,
$G_p^{(1)}(z)$	first derivative of $G_p(z)$,
g_p	mean batch size of class p,
$g_p^{(i)}$	ith factorial moment of the batch size of class p,
$B_p^*(s)$	LST of the DF for the service time of a message of class p,
b_p	mean service time of a message of class p,
$b_p^{(i)}$	ith moment of the service time of a message of class p,
$B_{g,p}^*(s)$	LST of the DF for the service time of a supermessage of class p $(\equiv G_p[B_p^*(s)])$,
$b_{g,p}$	mean service time of a supermessage of class p $(= g_p b_p)$,
$b_{g,p}^{(i)}$	ith moment of the service time of a supermessage of class p,
$b_{g,p}^{(2)}$	$= g_p^{(2)} b_p^2 + g_p b_p^{(2)}$,
$b_{g,p}^{(3)}$	$= g_p^{(3)} b_p^3 + 3 g_p^{(2)} b_p b_p^{(2)} + g_p b_p^{(3)}$,
$B_{g,p}^+(s)$	$\equiv \frac{1}{\lambda_p^+} \sum_{k=1}^{p} \lambda_k B_{g,k}^*(s)$,
$b_{g,p}^+$	$= \frac{1}{\lambda_p^+} \sum_{k=1}^{p} \lambda_k b_{g,k}$,
$b_{g,p}^{+(i)}$	$= \frac{1}{\lambda_p^+} \sum_{k=1}^{p} \lambda_k b_{g,k}^{(i)}$,
$B_g^*(s)$	$\equiv B_{g,P}^+(s) = \frac{1}{\lambda} \sum_{p=1}^{P} \lambda_p B_{g,p}^*(s)$,

b_g	$= b^+_{g,P} = \frac{1}{\lambda}\sum_{p=1}^{P} \lambda_p b_{g,p}$,
ρ_p	traffic intensity of messages of class p $(= \lambda_p b_{g,p})$,
ρ	total traffic intensity $(= \sum_{p=1}^{P} \rho_p)$,
ρ_p^+	$= \sum_{k=1}^{p} \rho_k < 1$,
ρ_p^-	$= \sum_{k=p+1}^{P} \rho_k < 1$,
$\Theta^+_{g,p-1}$	length of the delay cycle generated by messages of H-class, whose initial delay is the service time of a batch of messages of H-class,
$\Theta^+_{g,p-1}(x)$	DF for $\Theta^+_{g,p-1}$,
$\Theta^+_{g,p-1}(s)$	LST of $\Theta^+_{g,p-1}(x)$,
$\Theta^*_g(s)$	$\equiv \Theta^+_{g,P}(s)$,
$E[\cdot]$	expected value of a random variable,
$W^*_p(s)$	LST of the DF for the waiting time of an arbitrary message of class p.

We note that the LST $\Theta^+_{g,p-1}(s)$ satisfies the equation

$$\Theta^+_{g,p-1}(s) = B^+_{g,p-1}[s + \lambda^+_{p-1} - \lambda^+_{p-1}\Theta^+_{g,p-1}(s)] \tag{1}$$

and that the first three moments of $\Theta^+_{g,p-1}$ are given by

$$E[\Theta^+_{g,p-1}] = \frac{b^+_{g,p-1}}{1-\rho^+_{p-1}} \tag{2a}$$

$$E[(\Theta^+_{g,p-1})^2] = \frac{b^{+(2)}_{g,p-1}}{(1-\rho^+_{p-1})^3} \tag{2b}$$

$$E[(\Theta^+_{g,p-1})^3] = \frac{b^{+(3)}_{g,p-1}}{(1-\rho^+_{p-1})^4} + \frac{3\lambda^+_{p-1}(b^{+(2)}_{g,p-1})^2}{(1-\rho^+_{p-1})^5}. \tag{2c}$$

2 Queue Size at Service Start Points

In this section, we derive the probability generating function (PGF) for the queue size of messages of class p at the beginning of service to a message of class p in the steady state, denoted by $\Phi_p(z)$, which will be needed in Section 3.7 when the waiting time of an arbitrary message of class p is considered. We can apply the same approach to all our models. Since the queue size distribution is invariant as long as the service discipline is impartial (sec. 3.4 in Kleinrock [19]), we can utilize the results for the FCFS system given by Takagi and Takahashi [28].

In order to derive $\Phi_p(z)$, we consider a *tagged* message of class p, denoted by M, and the supermessage, denoted by SM, which M belongs to. At the beginning of service to M, the following three types of messages of class p may exist in the queue (Figure 1).

(a) Messages that arrive during the waiting time of SM.

(b) Messages that arrive during the waiting time of M while the SM is in service.

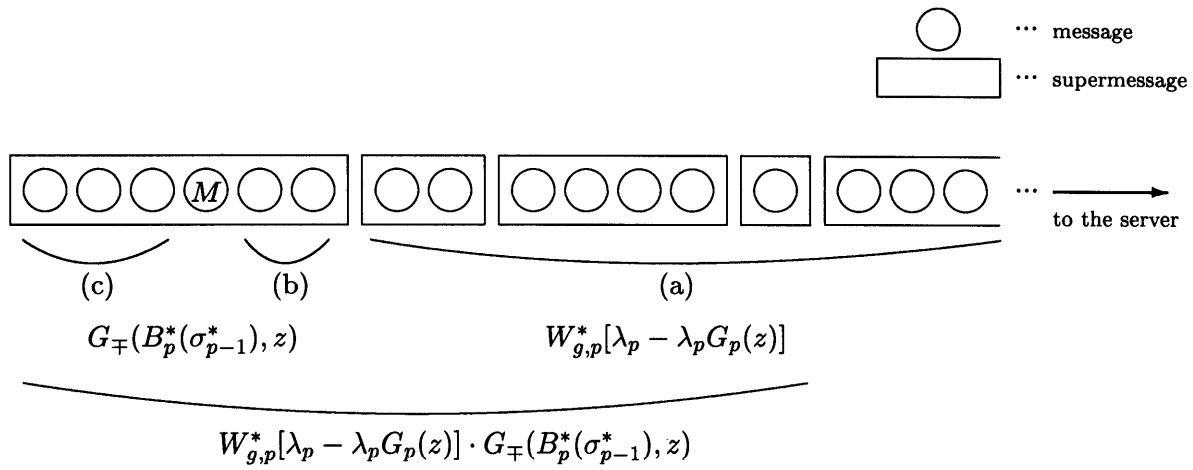

Figure 1: The components of $\Phi_p(z)$.

(c) Messages that belong to SM but have not been served by the time of service to M.

Let $W_{g,p}^*(s)$ be the LST of the waiting time of the SM which M belongs to. Then the PGF for the number of (a)-messages is given by $W_{g,p}^*[\lambda_p - \lambda_p G_p(z)]$ (sec. 5.5 in Kleinrock [18]). Let $D_p(z)$ be the PGF for the sum of the numbers of (b) and (c)-messages. $D_p(z)$ is derived in the same way for all our models. First we place the condition that the batch size G of SM is n. It then occurs with probability $1/n$ that the number G_- of messages which belong to SM and are served before M is i and that the number G_+ of messages which are served after M is j, where $i + j + 1 = n$. Since the probability that $G = n$ is given by $ng_{p,n}/g_p$, we have

$$\text{Prob}[G_- = i, G_+ = j] = \frac{g_{p,i+j+1}}{g_p}, \qquad i + j = 0, 1, \ldots.$$

Therefore we obtain

$$\begin{aligned} G_{\mp}(z_-, z_+) &:= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} z_-^i z_+^j \text{Prob}[G_- = i, G_+ = j] = \sum_{i=0}^{\infty}\sum_{n=i+1}^{\infty} z_-^i z_+^{n-i-1}\frac{g_{p,n}}{g_p} \\ &= \sum_{n=1}^{\infty}\frac{g_{p,n}z_+^n}{g_p z_+}\sum_{i=0}^{n-1}\left(\frac{z_-}{z_+}\right)^i = \sum_{n=1}^{\infty}\frac{g_{p,n}(z_-^n - z_+^n)}{g_p(z_- - z_+)} = \frac{G_p(z_-) - G_p(z_+)}{g_p(z_- - z_+)}. \end{aligned} \tag{3}$$

Since (b)-messages arrive during the services for the G_- messages, we obtain

$$D_p(z) = G_{\mp}(B_p^*(\sigma_{p-1}^*), z) = \frac{B_{g,p}^*(\sigma_{p-1}^*) - G_p(z)}{g_p[B_p^*(\sigma_{p-1}^*) - z]}, \tag{4}$$

where

$$\sigma_{p-1}^* := \lambda_p^+ - \lambda_p G_p(z) - \lambda_{p-1}^+ \Theta_{g,p-1}^+[\lambda_p - \lambda_p G_p(z)]. \tag{5}$$

Therefore we have $\Phi_p(z)$ as

$$\Phi_p(z) = W_{g,p}^*[\lambda_p - \lambda_p G_p(z)] D_p(z). \tag{6}$$

From the expressions for $W_{g,p}^*(s)$ given in [26], [28] for the FCFS systems and $D_p(z)$ in (4), the expressions for $\Phi_p(z)$ in our models are derived as follows.

NPNV

$$\Phi_p(z) = \frac{(1-\rho)\sigma^*_{p-1} + \sum_{k=p+1}^{P} \lambda_k g_k [1 - B^*_k(\sigma^*_{p-1})]}{\lambda_p g_p [B^*_p(\sigma^*_{p-1}) - z]} \tag{7}$$

PRNV

$$\Phi_p(z) = \frac{(1-\rho^+_p)\sigma^*_{p-1}}{\lambda_p g_p [B^*_p(\sigma^*_{p-1}) - z]} \tag{8}$$

NPMV

$$\Phi_p(z) = \frac{(1-\rho)\frac{1-V^*(\sigma^*_{p-1})}{E[V]} + \sum_{k=p+1}^{P} \lambda_k g_k [1 - B^*_k(\sigma^*_{p-1})]}{\lambda_p g_p [B^*_p(\sigma^*_{p-1}) - z]} \tag{9}$$

PRMV

$$\Phi_p(z) = \frac{(1-\rho)\frac{1-V^*(\sigma^*_{p-1})}{E[V]} + \rho^-_p \sigma^*_{p-1}}{\lambda_p g_p [B^*_p(\sigma^*_{p-1}) - z]} \tag{10}$$

NPSV

$$\begin{aligned} \Phi_p(z) \;=\; & \left(\frac{(1-\rho)\{V^*(\lambda)\sigma^*_{p-1} + \lambda[1 - V^*(\sigma^*_{p-1})]\}}{V^*(\lambda) + \lambda E[V]} + \sum_{k=p+1}^{P} \lambda_k g_k [1 - B^*_k(\sigma^*_{p-1})] \right) \\ & \times \frac{1}{\lambda_p g_p [B^*_p(\sigma^*_{p-1}) - z]} \end{aligned} \tag{11}$$

PRSV

$$\Phi_p(z) = \frac{(1-\rho)\lambda[1 - V^*(\sigma^*_{p-1})] + \{(1-\rho^+_p)V^*(\lambda) + \rho^-_p \lambda E[V]\}\sigma^*_{p-1}}{(V^*(\lambda) + \lambda E[V])\lambda_p g_p [B^*_p(\sigma^*_{p-1}) - z]} \tag{12}$$

3 Conditional Waiting Times

In this section we show preliminary results for the analysis of the waiting time W_p of a tagged message M of class p, which is defined as the time interval from its arrival to the service start. We first divide the time axis into several periods of system states for the non-preemptive models in Section 3.1, and for the preemptive-resume models in Section 3.2. Then we derive the LST and the first two moments of the DF for the conditional waiting time when the tagged message arrives during each of these periods in Sections 3.3 through 3.7.

3.1 Classification of the system states for non-preemptive models

In non-preemptive models we call the duration in which the server is neither busy nor taking a vacation an idle period. If M arrives as a member of a batch during an idle period, it has a chance to be selected for service (called *eligible* hereafter) immediately. Because the length I of an idle period is exponentially distributed with mean $1/\lambda$, $I^*(s)$ and $E[I]$ are given by

$$I^*(s) = \frac{\lambda}{s+\lambda} \quad , \quad E[I] = \frac{1}{\lambda}. \tag{13}$$

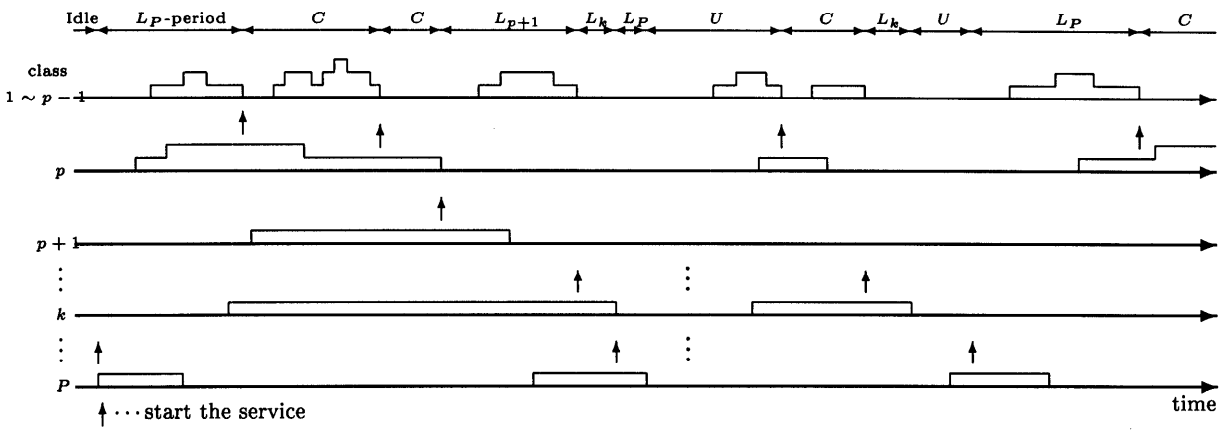

Figure 2: The service periods for the non-preemptive models.

If M arrives when the server is busy (or taking a vacation), it must wait at least until the server finishes the current service (or the vacation) and all messages of H-class leave the system. Such a period is called a *delay cycle* [26]. Consider a delay cycle of length T, called a T-period, with its initial delay denoted by T_0. If $T^*(s)$ and $T_0^*(s)$ denote the LSTs of the DFs for T and T_0, respectively, we have [26]

$$T^*(s) = T_0^*[s + \lambda_{p-1}^+ - \lambda_{p-1}^+ \Theta_{g,p-1}^+(s)], \tag{14a}$$

$$E[T] = \frac{E[T_0]}{1-\rho_{p-1}^+}, \tag{14b}$$

$$E[T^2] = \frac{E[T_0^2]}{(1-\rho_{p-1}^+)^2} + \frac{E[T_0]\lambda_{p-1}^+ b_{g,p-1}^{+(2)}}{(1-\rho_{p-1}^+)^3}, \tag{14c}$$

$$E[T^3] = \frac{E[T_0^3]}{(1-\rho_{p-1}^+)^3} + \frac{3E[T_0^2]\lambda_{p-1}^+ b_{g,p-1}^{+(2)}}{(1-\rho_{p-1}^+)^4} + \frac{E[T_0]\lambda_{p-1}^+ b_{g,p-1}^{+(3)}}{(1-\rho_{p-1}^+)^4} + \frac{3E[T_0](\lambda_{p-1}^+ b_{g,p-1}^{+(2)})^2}{(1-\rho_{p-1}^+)^5}. \tag{14d}$$

A non-idle period is divided into the following disjoint sets of T periods (Figure 2).

U-period which begins with a vacation and ends when the server has exhaustively served messages of H-class after the vacation.

H-period which begins when a batch of messages of H-class arrives to find the server idle, and ends when the server has exhaustively served messages of H-class.

L_k-period which is initiated by the service to a message of class k ($k = p+1, \cdots, P$), and terminated when the server has exhaustively served messages of H-class.

C-period which begins with the service to a message of class p, and ends when the server has exhaustively served messages of H-class.

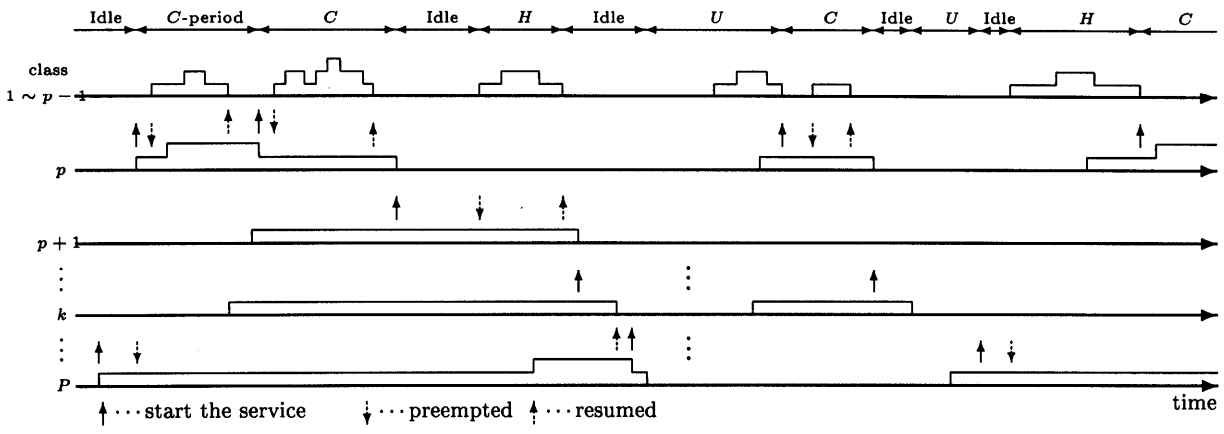

Figure 3: The service periods for the preemptive-resume models.

The LSTs of the DFs for the initial delays for the above periods are respectively given by

U-period $V^*(s)$,

H-period $B^+_{g,p-1}(s)$,

L$_k$-period $B^*_k(s)$ $(k = p+1, \cdots, P)$,

C-period $B^*_p(s)$.

We denote the LSTs of the DFs for the above periods by $U^*(s)$, $H^*(s)$, $L^*_k(s)$ $(k = p+1, \cdots, P)$, and $C^*_p(s)$, respectively. Note that $H^*(s)$ equals $\Theta^+_{g,p-1}(s)$ and that $C^*_p(s)$ represents the LST of the DF for a *completion time* C_p (Gaver [10]) of a message of class p. These are obtained by substituting the LSTs for the corresponding initial delays into $T^*_0(s)$ in (14a).

3.2 Classification of the system states for preemptive-resume models

In preemptive-resume models, the service to a message is preempted upon the arrival of a batch of a higher priority. Since a message M of a given class is never delayed by the service to any lower-class message, we can neglect lower-class messages in the analysis for M. Thus there are no L$_k$-periods in the system.

We define the following periods for the preemptive-resume models (Figure 3).

Idle period which begins when there are no messages of class p and H and continues while the server is neither busy nor taking a vacation, or serving a message of L-class.

U-period which begins with a vacation and ends when the server has exhaustively served messages of H-class after the vacation.

H-period which begins when a batch of messages of H-class arrives in an idle period, and ends when the server has exhaustively served messages of H-class.

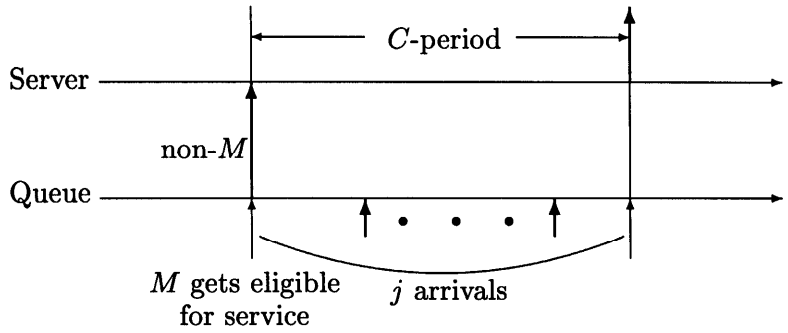

Figure 4: The conditional waiting time when M is not selected with prob. $m/(m+1)$.

C-period which begins with the service to a message of class p, and ends when the server has exhaustively served messages of H-class.

Note that each of a U-period, an H-period and a C-period has the same distribution as each of those in the non-preemptive models. For the idle period in this case, $I^*(s)$ and $E[I]$ are given by

$$I^*(s) = \frac{\lambda_p^+}{s+\lambda_p^+} \quad , \quad E[I] = \frac{1}{\lambda_p^+}. \tag{15}$$

3.3 Conditional waiting time when M arrives during an idle period

When M arrives during an idle period, M gets eligible for service upon arrival. Let $W_{p,m}$ be the waiting time of M from the epoch when M gets eligible for service, on the condition that there are m messages of class p, excluding M, in the system at that epoch. M is selected for next service immediately with probability $1/(m+1)$, or is delayed with probability $m/(m+1)$ until a later chance, which occurs after a completion time of class p (Figure 4). By conditioning that j messages of class p arrive during the completion time, we have the following recurrence relation for the LST $W^*_{p,m}(s)$ of the DF for $W_{p,m}$:

$$W^*_{p,m}(s) = \frac{1}{m+1} + \frac{m}{m+1}\sum_{j=0}^{\infty} C^*_{p,j}(s) W^*_{p,m+j-1}(s), \tag{16a}$$

where

$$\sum_{j=0}^{\infty} C^*_{p,j}(s) z^j = C^*_p[s+\lambda_p - \lambda_p G_p(z)]. \tag{16b}$$

which is an extension of Kingman's result [17] for the non-priority model. By following Takács [24], we obtain the first two moments of $W_{p,m}$ as follows.

$$E[W_{p,m}] = \frac{mE[C_p]}{2-\lambda_p g_p E[C_p]} = \frac{mb_p}{2-\rho_{p-1}^+ - \rho_p^+}, \tag{16c}$$

$$E[W^2_{p,m}] = \frac{2E[C_p]^2 m(m-1)}{(2-\lambda_p g_p E[C_p])(3-2\lambda_p g_p E[C_p])} + \frac{m\left\{(6-\lambda_p g_p E[C_p])E[C_p^2] + 2\lambda_p g_p^{(2)}(E[C_p])^3\right\}}{(2-\lambda_p g_p E[C_p])^2(3-2\lambda_p g_p E[C_p])}$$

$$= \frac{2b_p^2 m(m-1)}{(2-\rho_{p-1}^+-\rho_p^+)(3-2\rho_p^+-\rho_{p-1}^+)} + \frac{m(6-5\rho_{p-1}^+-\rho_p^+)b_p\lambda_{p-1}^+ b_{g,p-1}^{+(2)}}{(1-\rho_{p-1}^+)(2-\rho_{p-1}^+-\rho_p^+)^2(3-2\rho_p^+-\rho_{p-1}^+)}$$
$$+ \frac{m[(6-5\rho_{p-1}^+-\rho_p^+)b_p^{(2)}+2\lambda_p g_p^{(2)}b_p^3]}{(2-\rho_{p-1}^+-\rho_p^+)^2(3-2\rho_p^+-\rho_{p-1}^+)}. \tag{16d}$$

Next we derive the PGF $\Pi_p^{\rm I}(z)$ for the number $\Pi_p^{\rm I}$ of messages of class p, other than M, that arrive in the same batch as M. Since

$$\pi_{p,m}^{\rm I} := \text{Prob}[\Pi_p^{\rm I} = m] = \frac{(m+1)g_{p,m+1}}{\sum_{j=0}^{\infty}(j+1)g_{p,j+1}} = \frac{(m+1)g_{p,m+1}}{g_p},$$

we have

$$\Pi_p^{\rm I}(z) = E[z^{\Pi_p^{\rm I}}] = \sum_{m=0}^{\infty}\pi_{p,m}^{\rm I} z^m = \frac{G_p^{(1)}(z)}{g_p}, \tag{17a}$$

which yields

$$\sum_{m=1}^{\infty} m\pi_{p,m}^{\rm I} = E[\Pi_p^{\rm I}] = \frac{g_p^{(2)}}{g_p}, \tag{17b}$$

$$\sum_{m=2}^{\infty} m(m-1)\pi_{p,m}^{\rm I} = E[(\Pi_p^{\rm I})^2] - E[\Pi_p^{\rm I}] = \frac{g_p^{(3)}}{g_p}. \tag{17c}$$

We thus obtain the LST of the DF for the conditional waiting time of M when it arrives during an idle period:

$$E[e^{-sW_p}|\text{I}] = \sum_{m=0}^{\infty}\pi_{p,m}^{\rm I} W_{p,m}^*(s). \tag{18}$$

3.4 Conditional waiting time when M arrives during a U-period

The tagged message M must wait until the end of the U-period during which it arrives. Let x be the length of the U-period. First, we derive the waiting time for given x. It consists of the time until the end of the U-period whose LST of the DF is denoted by $W_p^1(s|\text{U},x)$, and the time thereafter until the start of service to M whose LST of the DF is denoted by $W_p^2(s|\text{U},x)$ (Figure 5). They are independent of each other being conditioned on x. Note that $W_p^1(s|\text{U},x)$ is the LST of the DF for the remaining time of a U-period of length x (sec. 5.7 in Cooper [4] and sec. 5.2 in Kleinrock [18]). Thus it is given by

$$W_p^1(s|\text{U},x) = \int_0^x e^{-sy}\frac{dy}{x} = \frac{1-e^{-sx}}{sx}. \tag{19}$$

When the U-period ends, M gets eligible for service. Messages of class p in the system at this epoch consist of those messages that have arrived together with M in a batch and those arriving during x. Let $\Pi_p^{\rm II}(x)$ be the number of messages of class p, excluding M, in the system when the U-period of length x ends. Then the GF $\Pi_p^{\rm II}(z|x)$ for $\pi_{p,m}^{\rm II}(x) := \text{Prob}[\Pi_p^{\rm II}(x) = m]$ is given by

$$\Pi_p^{\rm II}(z|x) = E[z^{\Pi_p^{\rm II}(x)}] = \sum_{m=0}^{\infty}\pi_{p,m}^{\rm II}(x)z^m = e^{-\lambda_p[1-G_p(z)]x}\cdot\frac{G_p^{(1)}(z)}{g_p}, \tag{20a}$$

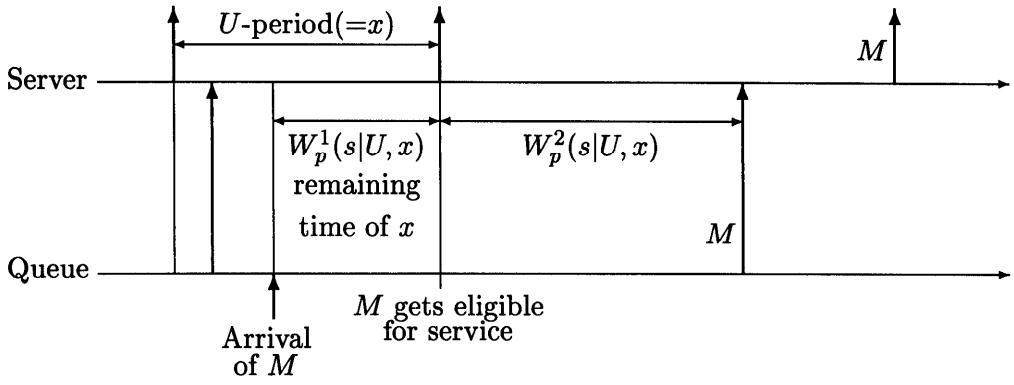

Figure 5: The conditional waiting time when M arrives during a U-period of length x.

which gives

$$\sum_{m=1}^{\infty} m\pi_{p,m}^{\mathrm{II}}(x) = E[\Pi_p^{\mathrm{II}}(x)] = \lambda_p g_p x + \frac{g_p^{(2)}}{g_p}, \tag{20b}$$

$$\sum_{m=2}^{\infty} m(m-1)\pi_{p,m}^{\mathrm{II}}(x) = E[(\Pi_p^{\mathrm{II}}(x))^2] - E[\Pi_p^{\mathrm{II}}(x)] = \lambda_p^2 g_p^2 x^2 + 3\lambda_p g_p^{(2)} x + \frac{g_p^{(3)}}{g_p}. \tag{20c}$$

From (16) and (20), we obtain

$$W_p^2(s|\mathrm{U},x) = \sum_{m=0}^{\infty} \pi_{p,m}^{\mathrm{II}}(x) W_{p,m}^*(s). \tag{21}$$

The product of (19) and (21) yields $E[e^{-sW_p}|\mathrm{U},x]$. Since the probability that a message arrives during the U-period of length x is proportional to x as well as to the relative frequency of such a length given by $dU(x)$ (sec. 5.2 in Kleinrock [18]), after normalizing properly, we obtain

$$E[e^{-sW_p}|\mathrm{U}] = \int_0^{\infty} \frac{x dU(x)}{E[U]} \cdot \frac{1-e^{-sx}}{sx} \sum_{m=0}^{\infty} \pi_{p,m}^{\mathrm{II}}(x) W_{p,m}^*(s). \tag{22}$$

3.5 Conditional waiting time when M arrives during an H-period

By an argument similar to the one that led to (22), we can derive $E[e^{-sW_p}|\mathrm{H}]$ as

$$E[e^{-sW_p}|\mathrm{H}] = \int_0^{\infty} \frac{x d\Theta_{g,p-1}^{+}(x)}{E[\Theta_{g,p-1}^{+}]} \cdot \frac{1-e^{-sx}}{sx} \sum_{m=0}^{\infty} \pi_{p,m}^{\mathrm{II}}(x) W_{p,m}^*(s), \tag{23}$$

where $\pi_{p,m}^{\mathrm{II}}(x)$ is given in (20a).

3.6 Conditional waiting time when M arrives during an L_k-period

We can also derive $E[e^{-sW_p}|\mathrm{L}_k]$ as

$$E[e^{-sW_p}|\mathrm{L}_k] = \int_0^{\infty} \frac{x dL_k(x)}{E[L_k]} \cdot \frac{1-e^{-sx}}{sx} \sum_{m=0}^{\infty} \pi_{p,m}^{\mathrm{II}}(x) W_{p,m}^*(s). \tag{24}$$

where $\pi_{p,m}^{\mathrm{II}}(x)$ is given in (20a).

3.7 Conditional waiting time when M arrives during a C-period

We can apply the same argument to derive $E[e^{-sW_p}|\mathrm{C}]$ as

$$E[e^{-sW_p}|\mathrm{C}] = \int_0^\infty \frac{xdC_p(x)}{E[C_p]} \cdot \frac{1-e^{-sx}}{sx} \sum_{m=0}^{\infty} \pi_{p,m}^{\mathrm{C}}(x) W_{p,m}^*(s), \tag{25}$$

where $\pi_{p,m}^{\mathrm{C}}(x)$ denotes the probability that there are m messages of class p, excluding M, in the system when the C-period of length x ends. Let $\Pi_p^{\mathrm{C}}(z|x)$ be its GF. $\Pi_p^{\mathrm{C}}(z|x)$ is given by the product of the following three PGFs. The first is $\Phi_p(z)$, the PGF for the number of messages of class p in the system when the C-period starts, namely, when the service to a message of class p is started; it is given in Section 2 for the individual models. The second is $e^{-\lambda_p[1-G_p(z)]x}$, which is the PGF for the number of messages of class p arriving during the C-period of length x, excluding the batch which M belongs to. The third is $G_p^{(1)}(z)/g_p$, which is the PGF for the number of messages arriving with M in a batch, excluding M. Thus we have

$$\Pi_p^{\mathrm{C}}(z|x) = \sum_{m=0}^{\infty} \pi_{p,m}^{\mathrm{C}}(x) z^m = \Phi_p(z) \Pi_p^{\mathrm{II}}(z|x), \tag{26}$$

where $\Pi_p^{\mathrm{II}}(z|x)$ is given in (20a).

4 Waiting Times

In the following subsections, we derive the unconditional LST $W_p^*(s)$ of the DF for the waiting time W_p of an arbitrary message of class p and its first two moments for each model. To do so, we first obtain the probabilities that the system is at a random point in time during each period of the state defined in Section 3. Because of *PASTA* (*Poisson arrivals see time averages*, see sec. 11.2 of Heyman and Sobel [11]), we can then derive the unconditional waiting time from the conditional waiting times for each model.

4.1 NPNV

In the NPNV model, the system is in an idle period, an H-period, an L_k-periods or a C-period. Note that a whole *busy period* consists of H-periods, L_k-periods and C-periods. Those epochs when the system becomes empty are *regenerative points* (sec. 6.4 in Heyman and Sobel [11]), and a pair of an idle period and a busy period appears exactly once between any two successive regenerative points. Hence, from the ratio of the mean lengths of the both periods, we have

$$\mathrm{Prob[idle]} = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda} + \frac{gb}{1-\rho}} = 1-\rho. \tag{27a}$$

An H-period appears once per busy period if a batch of H-class arrives with probability λ_{p-1}^+/λ when the system is idle. Thus

$$\text{Prob[H]} = \frac{\frac{\lambda_{p-1}^+}{\lambda} \cdot \frac{b_{g,p-1}^+}{1-\rho_{p-1}^+}}{\frac{1}{\lambda} + \frac{gb}{1-\rho}} = \frac{\rho_{p-1}^+(1-\rho)}{1-\rho_{p-1}^+}. \tag{27b}$$

From (27a) and (27b), the remaining probabilities should sum to

$$\text{Prob[C]} + \sum_{k=p+1}^{P} \text{Prob[L}_k\text{]} = 1 - \rho - \frac{\rho_{p-1}^+(1-\rho)}{1-\rho_{p-1}^+} = \frac{\rho_{p-1}^-}{1-\rho_{p-1}^+}.$$

Noting that each group of messages of class p or L-class starts a C-period or an L_k-period, we have the ratio

$$\begin{aligned} \text{Prob[C]} : \text{Prob[L}_{p+1}\text{]} : \cdots : \text{Prob[L}_P\text{]} &= \lambda_p g_p \cdot b_p : \lambda_{p+1} g_{p+1} \cdot b_{p+1} : \cdots : \lambda_P g_P \cdot b_P \\ &= \rho_p : \rho_{p+1} : \cdots : \rho_P. \end{aligned}$$

Thus we obtain

$$\text{Prob[L}_k\text{]} = \frac{\rho_k}{1-\rho_{p-1}^+}, \qquad k = p+1, \cdots, P, \tag{27c}$$

$$\text{Prob[C]} = \frac{\rho_p}{1-\rho_{p-1}^+}. \tag{27d}$$

From (18), (23), (24), (25) and (27), we can compute $W_p^*(s)$ and the first two moments as follows.

$$\begin{aligned} W_p^*(s) &= (1-\rho)E[e^{-sW_p}|\text{I}] + \frac{\rho_{p-1}^+(1-\rho)}{1-\rho_{p-1}^+} E[e^{-sW_p}|\text{H}] \\ &\quad + \sum_{k=p+1}^{P} \frac{\rho_k}{1-\rho_{p-1}^+} E[e^{-sW_p}|\text{L}_\text{k}] + \frac{\rho_p}{1-\rho_{p-1}^+} E[e^{-sW_p}|\text{C}], \end{aligned} \tag{28a}$$

$$E[W_p] = \frac{g_p^{(2)} b_p}{2g_p(1-\rho_p^+)} + \frac{\sum_{k=p}^{P} \lambda_k g_k b_k^{(2)} + \lambda_{p-1}^+ b_{g,p-1}^{+(2)}}{2(1-\rho_p^+)(1-\rho_{p-1}^+)}, \tag{28b}$$

$$\begin{aligned} E[W_p^2] &= \frac{2g_p^{(3)} b_p^2}{3(1-\rho_p^+)(2-\rho_{p-1}^+-\rho_p^+)g_p} + \frac{g_p^{(2)} b_p \lambda_{p-1}^+ b_{g,p-1}^{+(2)}}{(1-\rho_p^+)^2(2-\rho_{p-1}^+-\rho_p^+)g_p} \\ &\quad + \frac{(1-\rho_{p-1}^+) g_p^{(2)} b_p^{(2)}}{(1-\rho_p^+)^2(2-\rho_{p-1}^+-\rho_p^+)g_p} + \frac{\lambda_p (g_p^{(2)})^2 b_p^3}{(1-\rho_p^+)^2(2-\rho_{p-1}^+-\rho_p^+)g_p} \\ &\quad + \frac{2}{3(1-\rho_p^+)(1-\rho_{p-1}^+)(2-\rho_{p-1}^+-\rho_p^+)} \left(\sum_{k=p}^{P} \lambda_k g_k b_k^{(3)} + \lambda_{p-1}^+ b_{g,p-1}^{+(3)} \right) \\ &\quad + \frac{1}{(1-\rho_p^+)^2} \left(\sum_{k=p}^{P} \lambda_k g_k b_k^{(2)} + \lambda_{p-1}^+ b_{g,p-1}^{+(2)} \right) \\ &\quad \times \left(\frac{g_p^{(2)} b_p}{(2-\rho_{p-1}^+-\rho_p^+)g_p} + \frac{\lambda_{p-1}^+ b_{g,p-1}^{+(2)}}{(1-\rho_{p-1}^+)^2} + \frac{\lambda_p g_p b_p^{(2)}}{(1-\rho_{p-1}^+)(2-\rho_{p-1}^+-\rho_p^+)} \right). \end{aligned} \tag{28c}$$

4.2 PRNV

In the PRNV model, the system is in an idle period, an H-period, or a C-period. The probability that the server is not busy at an arbitrary epoch is $1-\rho$, and the probability that a message of L-class is in service at an arbitrary epoch is ρ_p^-. Hence we have

$$\text{Prob[idle]} = (1-\rho) + \rho_p^- = 1 - \rho_p^+. \tag{29a}$$

The probability that the system is in a C-period equals that in the NPNV model, thus

$$\text{Prob[C]} = \frac{\rho_p}{1-\rho_{p-1}^+}. \tag{29b}$$

It follows that

$$\text{Prob[H]} = 1 - (1-\rho_p^+) - \frac{\rho_p}{1-\rho_{p-1}^+} = \frac{\rho_{p-1}^+(1-\rho_p^+)}{1-\rho_{p-1}^+}. \tag{29c}$$

From (18), (23), (25) and (29), we get

$$W_p^*(s) = (1-\rho_p^+)E[e^{-sW_p}|\text{I}] + \frac{\rho_{p-1}^+(1-\rho_p^+)}{1-\rho_{p-1}^+}E[e^{-sW_p}|\text{H}] + \frac{\rho_p}{1-\rho_{p-1}^+}E[e^{-sW_p}|\text{C}], \tag{30a}$$

$$E[W_p] = \frac{g_p^{(2)}b_p}{2g_p(1-\rho_{p-1}^+)} + \frac{\lambda_p^+ b_{g,p}^{+(2)}}{2(1-\rho_p^+)(1-\rho_{p-1}^+)}, \tag{30b}$$

$$\begin{aligned}
E[W_p^2] = {} & \frac{2g_p^{(3)}b_p^2}{3(1-\rho_p^+)(2-\rho_{p-1}^+-\rho_p^+)g_p} + \frac{g_p^{(2)}b_p\lambda_{p-1}^+ b_{g,p-1}^{+(2)}}{(1-\rho_p^+)^2(2-\rho_{p-1}^+-\rho_p^+)g_p} \\
& + \frac{(1-\rho_{p-1}^+)g_p^{(2)}b_p^{(2)}}{(1-\rho_p^+)^2(2-\rho_{p-1}^+-\rho_p^+)g_p} + \frac{\lambda_p(g_p^{(2)})^2 b_p^3}{(1-\rho_p^+)^2(2-\rho_{p-1}^+-\rho_p^+)g_p} \\
& + \frac{2}{3(1-\rho_p^+)(1-\rho_{p-1}^+)(2-\rho_{p-1}^+-\rho_p^+)}\left(\lambda_{p-1}^+ b_{g,p-1}^{+(3)} + \lambda_p g_p b_p^{(3)}\right) \\
& + \frac{1}{(1-\rho_p^+)^2}\left(\lambda_{p-1}^+ b_{g,p-1}^{+(2)} + \lambda_p g_p b_p^{(2)}\right) \\
& \times \left(\frac{g_p^{(2)}b_p}{(2-\rho_{p-1}^+-\rho_p^+)g_p} + \frac{\lambda_{p-1}^+ b_{g,p-1}^{+(2)}}{(1-\rho_{p-1}^+)^2} + \frac{\lambda_p g_p b_p^{(2)}}{(1-\rho_{p-1}^+)(2-\rho_{p-1}^+-\rho_p^+)}\right).
\end{aligned} \tag{30c}$$

4.3 NPMV

In multiple vacation models, if the server returns from a vacation to find no messages waiting, it begins another vacation immediately. A regenerative point in such systems is the epoch at which the system is empty and a vacation begins. The time interval between two such successive regenerative points is called a *regeneration cycle* (sec. 2.2 in Takagi [26]), whose length is denoted by V_c. The LST $V_c^*(s)$ of the DF and the mean for V_c are given by

$$V_c^*(s) = V^*[s+\lambda-\lambda\Theta_g^*(s)], \tag{31a}$$

$$E[V_c] = \frac{E[V]}{1-\rho}, \tag{31b}$$

where $\Theta_g^*(s) \equiv \Theta_{g,P}^+(s)$ is the LST of the DF for the length Θ_g of a busy period generated by all messages, and it satisfies the equation

$$\begin{aligned}\Theta_g^*(s) &= B_g^*[s+\lambda-\lambda\Theta_g^*(s)],\\ E[\Theta_g] &= \frac{b_g}{1-\rho}.\end{aligned}$$

In the NPMV model the system is in a U-period, an L_k-period or a C-period. Since a U-period appears exactly once in a regeneration cycle, we get

$$\text{Prob[U]} = \frac{E[V]/(1-\rho_{p-1}^+)}{E[V_c]} = \frac{1-\rho}{1-\rho_{p-1}^+}, \tag{32a}$$

which leaves

$$\text{Prob[C]} + \sum_{k=p+1}^{P} \text{Prob[L}_k\text{]} = \frac{\rho_{p-1}^-}{1-\rho_{p-1}^+}.$$

By the similar argument as in Section 4.1, we obtain

$$\text{Prob[L}_k\text{]} = \frac{\rho_k}{1-\rho_{p-1}^+}, \quad k=p+1,\cdots,P \tag{32b}$$

$$\text{Prob[C]} = \frac{\rho_p}{1-\rho_{p-1}^+}. \tag{32c}$$

The results in (22), (24), (25) and (32) yield $W_p^*(s)$ and the first two moments as follows.

$$W_p^*(s) = \frac{1-\rho}{1-\rho_{p-1}^+}E[e^{-sW_p}|\text{U}] + \sum_{k=p+1}^{P}\frac{\rho_k}{1-\rho_{p-1}^+}E[e^{-sW_p}|\text{L}_k] + \frac{\rho_p}{1-\rho_{p-1}^+}E[e^{-sW_p}|\text{C}], \tag{33a}$$

$$E[W_p] = \frac{g_p^{(2)}b_p}{2g_p(1-\rho_p^+)} + \frac{\sum_{k=p}^{P}\lambda_k g_k b_k^{(2)} + \lambda_{p-1}^+ b_{g,p-1}^{+(2)}}{2(1-\rho_p^+)(1-\rho_{p-1}^+)} + \frac{(1-\rho)E[V^2]}{2(1-\rho_p^+)(1-\rho_{p-1}^+)E[V]}, \tag{33b}$$

$$\begin{aligned}E[W_p^2] = {}& \frac{2g_p^{(3)}b_p^2}{3(1-\rho_p^+)(2-\rho_{p-1}^+-\rho_p^+)g_p} + \frac{g_p^{(2)}b_p\lambda_{p-1}^+ b_{g,p-1}^{+(2)}}{(1-\rho_p^+)^2(2-\rho_{p-1}^+-\rho_p^+)g_p}\\ &+ \frac{(1-\rho_{p-1}^+)g_p^{(2)}b_p^{(2)}}{(1-\rho_p^+)^2(2-\rho_{p-1}^+-\rho_p^+)g_p} + \frac{\lambda_p(g_p^{(2)})^2 b_p^3}{(1-\rho_p^+)^2(2-\rho_{p-1}^+-\rho_p^+)g_p}\\ &+ \frac{2}{3(1-\rho_p^+)(1-\rho_{p-1}^+)(2-\rho_{p-1}^+-\rho_p^+)}\left(\sum_{k=p}^{P}\lambda_k g_k b_k^{(3)} + \lambda_{p-1}^+ b_{g,p-1}^{+(3)} + \frac{(1-\rho)E[V^3]}{E[V]}\right)\\ &+ \frac{1}{(1-\rho_p^+)^2}\left(\frac{(1-\rho)E[V^2]}{E[V]} + \sum_{k=p}^{P}\lambda_k g_k b_k^{(2)} + \lambda_{p-1}^+ b_{g,p-1}^{+(2)}\right)\\ &\times\left(\frac{g_p^{(2)}b_p}{(2-\rho_{p-1}^+-\rho_p^+)g_p} + \frac{\lambda_{p-1}^+ b_{g,p-1}^{+(2)}}{(1-\rho_{p-1}^+)^2} + \frac{\lambda_p g_p b_p^{(2)}}{(1-\rho_{p-1}^+)(2-\rho_{p-1}^+-\rho_p^+)}\right).\end{aligned} \tag{33c}$$

4.4 PRMV

In the PRMV model, the system is in an idle period, a U-period, an H-period or a C-period. The probabilities that the system is in each of a U-period and a C-period equal those in the NPMV model given in Section 4.3:

$$\text{Prob[U]} = \frac{1-\rho}{1-\rho_{p-1}^{+}}, \tag{34a}$$

$$\text{Prob[C]} = \frac{\rho_p}{1-\rho_{p-1}^{+}}. \tag{34b}$$

Since an idle period corresponds to the time for service to messages of L-class, we have

$$\text{Prob[idle]} = \rho_p^{-}, \tag{34c}$$

which yields

$$\text{Prob[H]} = \frac{\rho_{p-1}^{+}\rho_p^{-}}{1-\rho_{p-1}^{+}}. \tag{34d}$$

From (18), (22), (23), (25) and (34), we can obtain $W_p^*(s)$ and the first two moments as follows.

$$\begin{aligned} W_p^*(s) &= \rho_p^{-} E[e^{-sW_p}|\text{I}] + \frac{\rho_{p-1}^{-}\rho_p^{-}}{1-\rho_{p-1}^{+}} E[e^{-sW_p}|\text{H}] \\ &\quad + \frac{1-\rho}{1-\rho_{p-1}^{+}} E[e^{-sW_p}|\text{U}] + \frac{\rho_p}{1-\rho_{p-1}^{+}} E[e^{-sW_p}|\text{C}], \end{aligned} \tag{35a}$$

$$E[W_p] = \frac{g_p^{(2)} b_p}{2g_p(1-\rho_{p-1}^{+})} + \frac{\lambda_p^{+} b_{g,p}^{+(2)}}{2(1-\rho_p^{+})(1-\rho_{p-1}^{+})} + \frac{(1-\rho)E[V^2]}{2(1-\rho_p^{+})(1-\rho_{p-1}^{+})E[V]}, \tag{35b}$$

$$\begin{aligned} E[W_p^2] &= \frac{2g_p^{(3)} b_p^2}{3(1-\rho_p^{+})(2-\rho_{p-1}^{+}-\rho_p^{+})g_p} + \frac{g_p^{(2)} b_p \lambda_{p-1}^{+} b_{g,p-1}^{+(2)}}{(1-\rho_p^{+})^2(2-\rho_{p-1}^{+}-\rho_p^{+})g_p} \\ &\quad + \frac{(1-\rho_{p-1}^{+}) g_p^{(2)} b_p^{(2)}}{(1-\rho_p^{+})^2(2-\rho_{p-1}^{+}-\rho_p^{+})g_p} + \frac{\lambda_p (g_p^{(2)})^2 b_p^3}{(1-\rho_p^{+})^2(2-\rho_{p-1}^{+}-\rho_p^{+})g_p} \\ &\quad + \frac{2}{3(1-\rho_p^{+})(1-\rho_{p-1}^{+})(2-\rho_{p-1}^{+}-\rho_p^{+})} \left(\lambda_{p-1}^{+} b_{g,p-1}^{+(3)} + \lambda_p g_p b_p^{(3)} + \frac{(1-\rho)E[V^3]}{E[V]} \right) \\ &\quad + \frac{1}{(1-\rho_p^{+})^2} \left(\lambda_{p-1}^{+} b_{g,p-1}^{+(2)} + \lambda_p g_p b_p^{(2)} + \frac{(1-\rho)E[V^2]}{E[V]} \right) \\ &\quad \times \left(\frac{g_p^{(2)} b_p}{(2-\rho_{p-1}^{+}-\rho_p^{+})g_p} + \frac{\lambda_{p-1}^{+} b_{g,p-1}^{+(2)}}{(1-\rho_{p-1}^{+})^2} + \frac{\lambda_p g_p b_p^{(2)}}{(1-\rho_{p-1}^{+})(2-\rho_{p-1}^{+}-\rho_p^{+})} \right). \end{aligned} \tag{35c}$$

4.5 NPSV

In single vacation models, if the server returns from a vacation to find no messages waiting, the system becomes idle. A regenerative point in this system is again the epoch at which the system is empty and a vacation begins. The LST $V_c^*(s)$ of the DF and the mean for the length V_c of a

regeneration cycle are given by

$$V_c^*(s) = V^*(s+\lambda)I^*(s)\Theta_g^*(s) + V^*[s+\lambda-\lambda\Theta_g^*(s)] - V^*(s+\lambda), \tag{36a}$$

$$E[V_c] = \frac{V^*(\lambda)+\lambda E[V]}{\lambda(1-\rho)}. \tag{36b}$$

In the NPSV model, the system is in an idle period, a U-period, an H-period, an L_k-period or a C-period. Since a U-period appears exactly once in a regeneration cycle, we have

$$\text{Prob[U]} = \frac{E[V]/(1-\rho_{p-1}^+)}{E[V_c]} = \frac{(1-\rho)\lambda E[V]}{(1-\rho_{p-1}^+)(V^*(\lambda)+\lambda E[V])}. \tag{37a}$$

The system enters an idle period whose mean length is $1/\lambda$ if no messages arrive during a vacation, which occurs with probability $V^*(\lambda)$. Thus we have

$$\text{Prob[idle]} = \frac{V^*(\lambda)/\lambda}{E[V_c]} = \frac{(1-\rho)V^*(\lambda)}{V^*(\lambda)+\lambda E[V]}. \tag{37b}$$

An H-period appears once in a regeneration cycle if a batch of H-class arrives during an idle period. Therefore

$$\text{Prob[H]} = \frac{V^*(\lambda)\frac{\lambda_{p-1}^+}{\lambda}\cdot\frac{b_{g,p-1}^+}{1-\rho_{p-1}^+}}{E[V_c]} = \frac{(1-\rho)\rho_{p-1}^+ V^*(\lambda)}{(1-\rho_{p-1}^+)(V^*(\lambda)+\lambda E[V])}, \tag{37c}$$

which leaves

$$\text{Prob[C]} + \sum_{k=p+1}^{P}\text{Prob[L}_k\text{]} = \frac{\rho_{p-1}^-}{1-\rho_{p-1}^+}.$$

By the similar argument as in Section 4.1, we obtain

$$\text{Prob[L}_k\text{]} = \frac{\rho_k}{1-\rho_{p-1}^+}, \quad k = p+1,\cdots,P \tag{37d}$$

$$\text{Prob[C]} = \frac{\rho_p}{1-\rho_{p-1}^+}. \tag{37e}$$

From (18), (22), (23), (24), (25) and (37), we get

$$\begin{aligned} W_p^*(s) = {} & \frac{(1-\rho)V^*(\lambda)}{V^*(\lambda)+\lambda E[V]}E[e^{-sW_p}|\text{I}] + \frac{(1-\rho)\rho_{p-1}^+ V^*(\lambda)}{(1-\rho_{p-1}^+)(V^*(\lambda)+\lambda E[V])}E[e^{-sW_p}|\text{H}] \\ & + \frac{(1-\rho)\lambda E[V]}{(1-\rho_{p-1}^+)(V^*(\lambda)+\lambda E[V])}E[e^{-sW_p}|\text{U}] + \sum_{k=p+1}^{P}\frac{\rho_k}{1-\rho_{p-1}^+}E[e^{-sW_p}|\text{L}_k] \\ & + \frac{\rho_p}{1-\rho_{p-1}^+}E[e^{-sW_p}|\text{C}], \end{aligned} \tag{38a}$$

$$E[W_p] = \frac{g_p^{(2)}b_p}{2g_p(1-\rho_p^+)} + \frac{\sum_{k=p}^{P}\lambda_k g_k b_k^{(2)} + \lambda_{p-1}^+ b_{g,p-1}^{+(2)} + \frac{(1-\rho)\lambda E[V^2]}{V^*(\lambda)+\lambda E[V]}}{2(1-\rho_p^+)(1-\rho_{p-1}^+)}, \tag{38b}$$

$$E[W_p^2] = \frac{2g_p^{(3)}b_p^2}{3(1-\rho_p^+)(2-\rho_{p-1}^+-\rho_p^+)g_p} + \frac{g_p^{(2)}b_p\lambda_{p-1}^+ b_{g,p-1}^{+(2)}}{(1-\rho_p^+)^2(2-\rho_{p-1}^+-\rho_p^+)g_p}$$

$$+\frac{(1-\rho_{p-1}^+)g_p^{(2)}b_p^{(2)}}{(1-\rho_p^+)^2(2-\rho_{p-1}^+-\rho_p^+)g_p}+\frac{\lambda_p(g_p^{(2)})^2b_p^3}{(1-\rho_p^+)^2(2-\rho_{p-1}^+-\rho_p^+)g_p}$$
$$+\frac{2}{3(1-\rho_p^+)(1-\rho_{p-1}^+)(2-\rho_{p-1}^+-\rho_p^+)}\left(\sum_{k=p}^{P}\lambda_k g_k b_k^{(3)}+\lambda_{p-1}^+ b_{g,p-1}^{+(3)}+\frac{(1-\rho)\lambda E[V^3]}{V^*(\lambda)+\lambda E[V]}\right)$$
$$+\frac{1}{(1-\rho_p^+)^2}\left(\frac{(1-\rho)\lambda E[V^2]}{V^*(\lambda)+\lambda E[V]}+\sum_{k=p}^{P}\lambda_k g_k b_k^{(2)}+\lambda_{p-1}^+ b_{g,p-1}^{+(2)}\right)$$
$$\times\left(\frac{g_p^{(2)}b_p}{(2-\rho_{p-1}^+-\rho_p^+)g_p}+\frac{\lambda_{p-1}^+ b_{g,p-1}^{+(2)}}{(1-\rho_{p-1}^+)^2}+\frac{\lambda_p g_p b_p^{(2)}}{(1-\rho_{p-1}^+)(2-\rho_{p-1}^+-\rho_p^+)}\right). \tag{38c}$$

4.6 PRSV

In the PRSV model, the system is in an idle period, a U-period, an H-period or a C-period. The probabilities that the system is in a U-period and a C-period equal those in the NPSV model, thus we have

$$\text{Prob[U]}=\frac{(1-\rho)\lambda E[V]}{(1-\rho_{p-1}^+)(V^*(\lambda)+\lambda E[V])}, \tag{39a}$$

$$\text{Prob[C]}=\frac{\rho_p}{1-\rho_{p-1}^+}. \tag{39b}$$

Since an idle period consists of the time when the server is idle and the time for service to messages of L-class, we have

$$\text{Prob[idle]}=\frac{(1-\rho)V^*(\lambda)}{V^*(\lambda)+\lambda E[V]}+\rho_p^-=\frac{(1-\rho_p^+)V^*(\lambda)+\rho_p^-\lambda E[V]}{V^*(\lambda)+\lambda E[V]}, \tag{39c}$$

which yields

$$\text{Prob[H]}=\frac{\rho_{p-1}^+[(1-\rho_p^+)V^*(\lambda)+\rho_p^-\lambda E[V]]}{(1-\rho_{p-1}^+)(V^*(\lambda)+\lambda E[V])}. \tag{39d}$$

From (18), (22), (23), (25) and (39), we get

$$\begin{aligned}W_p^*(s) &= \frac{(1-\rho_p^+)V^*(\lambda)+\rho_p^-\lambda E[V]}{V^*(\lambda)+\lambda E[V]}E[e^{-sW_p}|\text{I}]\\ &+\frac{\rho_{p-1}^+[(1-\rho_p^+)V^*(\lambda)+\rho_p^-\lambda E[V]]}{(1-\rho_{p-1}^+)(V^*(\lambda)+\lambda E[V])}E[e^{-sW_p}|\text{H}]\\ &+\frac{(1-\rho)\lambda E[V]}{(1-\rho_{p-1}^+)(V^*(\lambda)+\lambda E[V])}E[e^{-sW_p}|\text{U}]+\frac{\rho_p}{1-\rho_{p-1}^+}E[e^{-sW_p}|\text{C}],\end{aligned} \tag{40a}$$

$$E[W_p] = \frac{g_p^{(2)}b_p}{2g_p(1-\rho_{p-1}^+)}+\frac{\lambda_p^+ b_{g,p}^{+(2)}+\frac{(1-\rho)\lambda E[V^2]}{V^*(\lambda)+\lambda E[V]}}{2(1-\rho_p^+)(1-\rho_{p-1}^+)}, \tag{40b}$$

$$E[W_p^2]=\frac{2g_p^{(3)}b_p^2}{3(1-\rho_p^+)(2-\rho_{p-1}^+-\rho_p^+)g_p}+\frac{g_p^{(2)}b_p\lambda_{p-1}^+ b_{g,p-1}^{+(2)}}{(1-\rho_p^+)^2(2-\rho_{p-1}^+-\rho_p^+)g_p}$$
$$+\frac{(1-\rho_{p-1}^+)g_p^{(2)}b_p^{(2)}}{(1-\rho_p^+)^2(2-\rho_{p-1}^+-\rho_p^+)g_p}+\frac{\lambda_p(g_p^{(2)})^2b_p^3}{(1-\rho_p^+)^2(2-\rho_{p-1}^+-\rho_p^+)g_p}$$

$$+\frac{2}{3(1-\rho_p^+)(1-\rho_{p-1}^+)(2-\rho_{p-1}^+-\rho_p^+)}\left(\lambda_{p-1}^+ b_{g,p-1}^{+(3)}+\lambda_p g_p b_p^{(3)}+\frac{(1-\rho)\lambda E[V^3]}{V^*(\lambda)+\lambda E[V]}\right)$$
$$+\frac{1}{(1-\rho_p^+)^2}\left(\lambda_{p-1}^+ b_{g,p-1}^{+(2)}+\lambda_p g_p b_p^{(2)}+\frac{(1-\rho)\lambda E[V^2]}{V^*(\lambda)+\lambda E[V]}\right)$$
$$\times\left(\frac{g_p^{(2)} b_p}{(2-\rho_{p-1}^+-\rho_p^+)g_p}+\frac{\lambda_{p-1}^+ b_{g,p-1}^{+(2)}}{(1-\rho_{p-1}^+)^2}+\frac{\lambda_p g_p b_p^{(2)}}{(1-\rho_{p-1}^+)(2-\rho_{p-1}^+-\rho_p^+)}\right). \tag{40c}$$

5 Comparison of the Moments

In this section we compare the results obtained for the individual models in Section 4.

5.1 Comparison between ROS and FCFS systems

For each model, the mean waiting time under ROS equals that under FCFS; this is obvious from *Little's formula* (Little [20]) and the fact that the queue size distribution is invariant.

We can also derive the following relationship on the second moments between ROS and FCFS disciplines for each priority class common to all models:

$$E[W_p^2]_{\mathrm{ROS}}=\frac{2(1-\rho_{p-1}^+)}{2-\rho_{p-1}^+-\rho_p^+}E[W_p^2]_{\mathrm{FCFS}}\geq E[W_p^2]_{\mathrm{FCFS}}. \tag{41}$$

This relation extends the result for the non-priority, single arrival model, which was originally derived by Takács [24] and later interpreted by Fuhrmann [8]. We note that Fuhrmann's argument does not apply to batch arrival models. Therefore, the relation in (41) is established for batch arrival priority models for the first time in this paper.

5.2 Comparison between non-preemptive and preemptive-resume systems

Comparing (28) with (30), the results for the PRNV model can be derived by setting $\lambda_k = 0$ ($k = p+1, \cdots, P$) in the results for the NPNV model; this is because the existence of L-class messages has no influence on the waiting time of a message of class p. However, the above never holds in the vacation models, because a sequence of vacations or an idle period may be terminated by the arrival of L-class messages. These observations are also made under FCFS [28].

5.3 Comparison between systems without vacations and with vacations

The moments of the waiting times in vacation models have some terms common to those in non-vacation models. Although the service to a message which finds the system idle starts upon arrival in non-vacation models, it does so after the residual time of a vacation in vacation models. This explains the difference in the moments for the two models. Therefore as ρ gets closer to 1, namely as the probability that a message arrives during a vacation gets smaller, the waiting time distribution

gets closer to that of the model without vacations. The similar argument is given by Kella and Yechiali [15].

6 Numerical Examples

In this section, we present some numerical examples. First, Figures 6 and 7 show the mean and the coefficient of variation of the waiting time as a function of ρ for three non-preemptive models, where the ratio of the arrival rates among different classes is fixed. These figures show the behavior concerning class 1 and class 4 so that we can clearly see the difference among classes. We obtain the numerical results under the following scenario.

number of classes	4
ratio of arrival rates	$\lambda_1 : \lambda_2 : \lambda_3 : \lambda_4 = 1 : 1 : 1 : 1$
service time for messages of class 1	3-stage Erlang distribution with mean 0.5
service time for messages of class 2	constant of length 0.5
service time for messages of class 3	2-stage Erlang distribution with mean 0.5
service time for messages of class 4	exponential distribution with mean 0.5
batch size for messages of class 1	geometric distribution with mean 2
batch size for messages of class 2	constant of size 2
batch size for messages of class 3	uniform distribution with mean 2
batch size for messages of class 4	constant of size 2
vacation time	2-stage Erlang distribution with mean 1.0

In Figure 6, we observe the following relationship

$$E[W_p]_{\mathrm{NV}} \leq E[W_p]_{\mathrm{SV}} \leq E[W_p]_{\mathrm{MV}},$$

while in Figure 7, we get the reverse relationship about the coefficient of variation of the waiting time. These relationships also hold for preemptive-resume models.

Next, Figures 8 and 9 show the mean and the coefficient of variation of the waiting time as a function of ρ for the NPNV model, where we assume the same scenario as in Figures 6 and 7. Figures 10 and 11 show the mean and the coefficient of variation of the waiting time as a function of g_2 for the NPNV model, where we assume that $\rho_1 = \rho_2 = \rho_3 = \rho_4 = 0.1$, that the batch size of class 2 is constant, and that the scenario is otherwise the same as in Figures 6 and 7. From these figures, we find the following interesting behavior. In the case where ρ is small and the mean batch size of a higher class is larger than that of a lower class, the mean waiting time of the higher class can be larger than that of the lower class. This is because the service to a tagged message may be delayed by other messages which belong to the same supermessage. Note that this never occurs in single arrival models. A similar phenomenon is also observed for the case where ρ is small and the mean service time of a higher class is larger than that of a lower class.

7 Concluding Remarks

In this paper we have analyzed $M^X/G/1$ priority queues with/without vacations under ROS. By considering the waiting times under various conditions, we have explicitly derived the first two moments for the waiting time distribution of an arbitrary message, which have revealed some noteworthy new results, especially the one in (41). We have also noted some interesting observations from the numerical examples for the mean waiting time and the coefficient of variation of the waiting time.

We remark that we can further derive the *response time* distribution for each model from our results. The LST $R_p^*(s)$ of the DF for the time that a message of class p spends in the system is given by

$$\begin{array}{rcll} R_p^*(s) & = & W_p^*(s)B_p^*(s) & \text{for the non-preemptive models,} \\ R_p^*(s) & = & W_p^*(s)B_p^*(\sigma_{p-1}) & \text{for the preemptive-resume models,} \end{array} \tag{42}$$

where $\sigma_{p-1} := s + \lambda_{p-1}^+ - \lambda_{p-1}^+ \Theta_{g,p-1}^+(s)$.

Although we have assumed that our systems are unsaturated ($\rho < 1$) throughout the paper, we can easily extend our results to *saturated* systems ($\rho \geq 1$). Consider q such that $\rho_{q-1}^+ < 1$ and $\rho_q^+ \geq 1$. The steady-state probability that the server is on a vacation is zero in a saturated system. The steady-state probability that a message of class $q+1, \cdots, P$ is in service also becomes zero. Messages of class q are served partially. In a non-preemptive priority model, service times for class q can be regarded as vacations when we are concerned with messages of class $1, 2, \cdots, q-1$. Therefore $W_p^*(s)$ $(p = 1, \cdots, q-1)$ and the first two moments for the non-preemptive model are given by (33) in which $V^*(s)$ is replaced by $B_q^*(s)$ and P is replaced by $q-1$. The $W_p^*(s)$ $(p = 1, \cdots, q-1)$ and the first two moments for preemptive-resume model are still given by (35) (see sec. 3.3 in [26] and [28]).

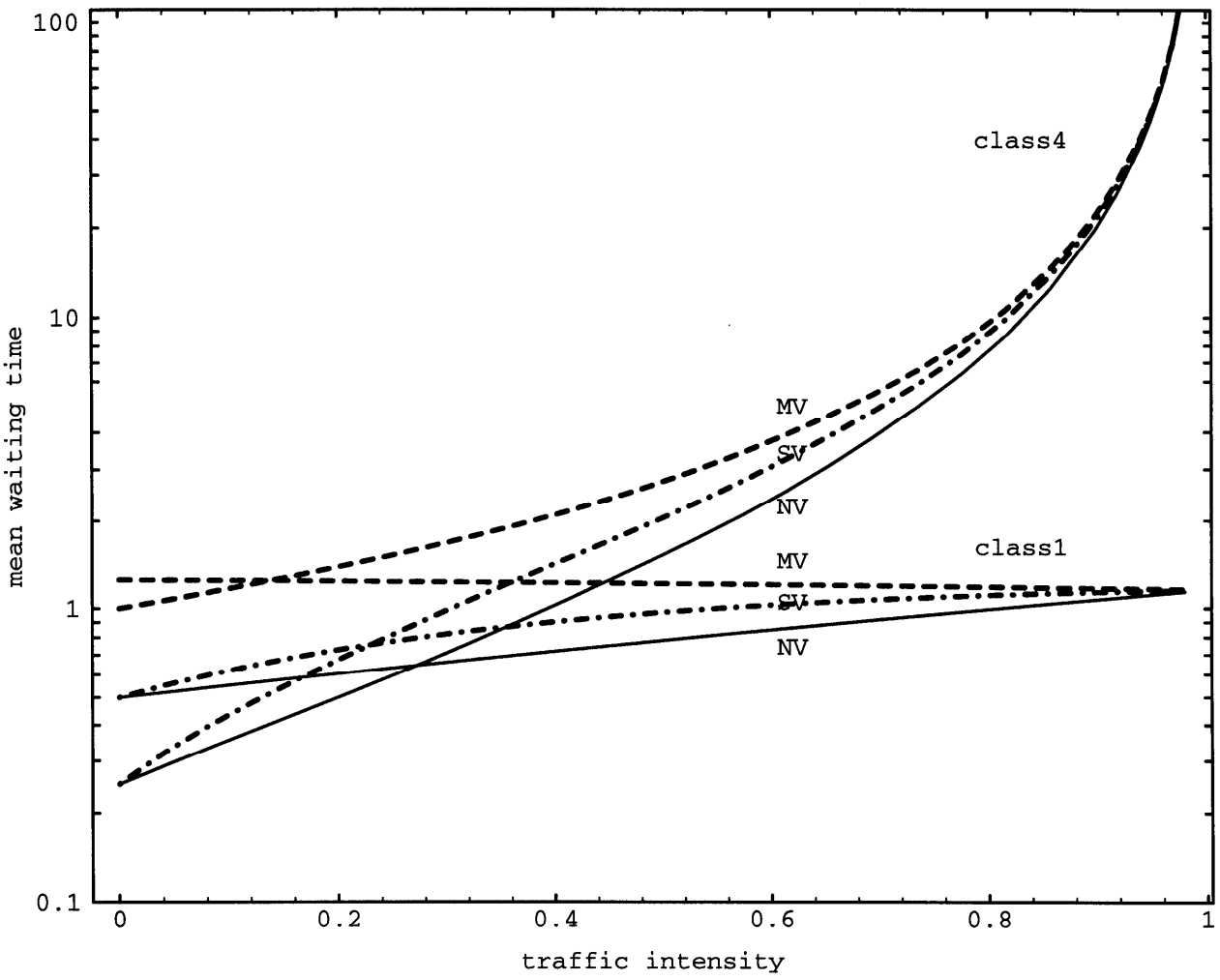

Figure 6: The mean waiting time in non-preemptive models.

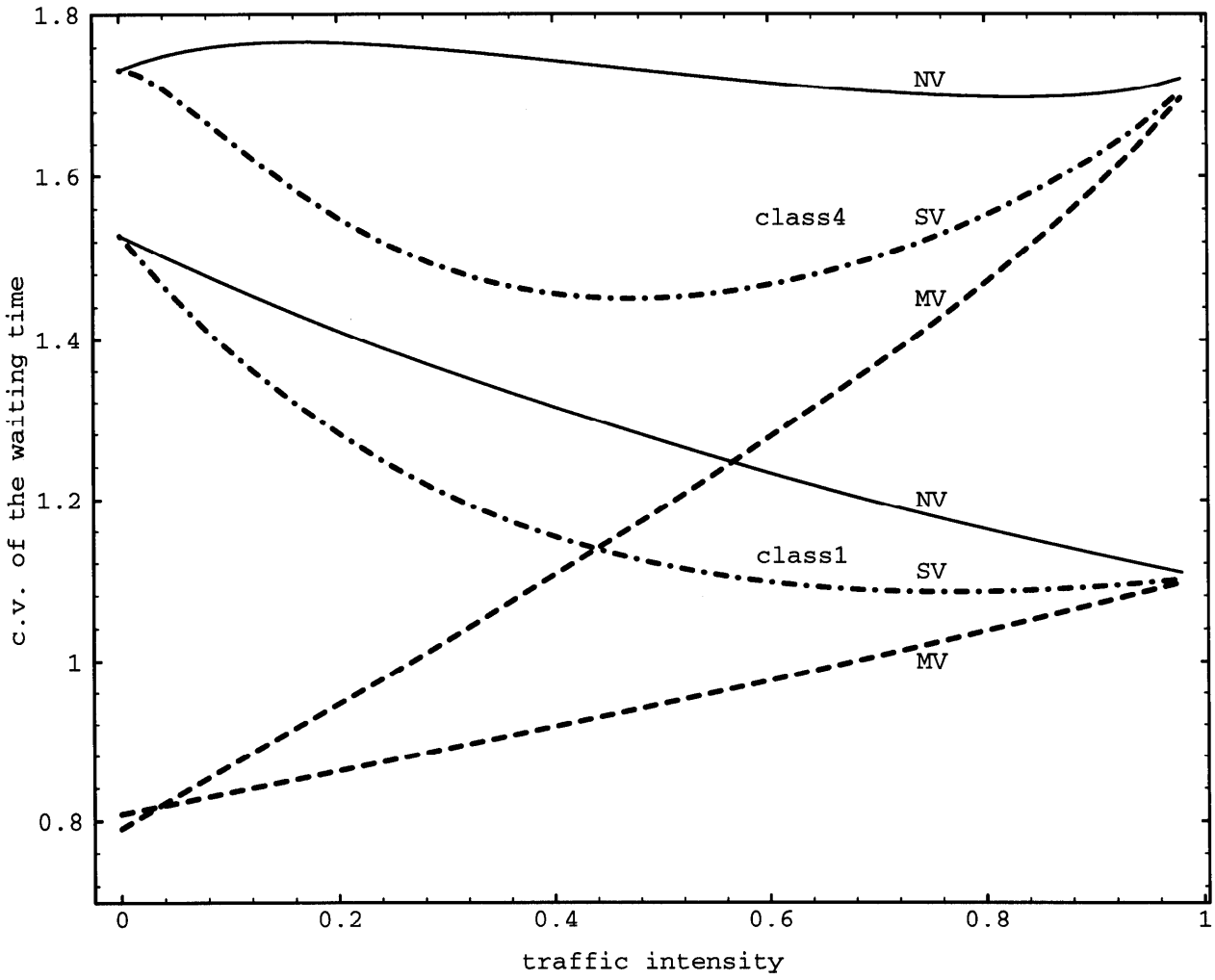

Figure 7: The coefficient of variation of the waiting time in non-preemptive models.

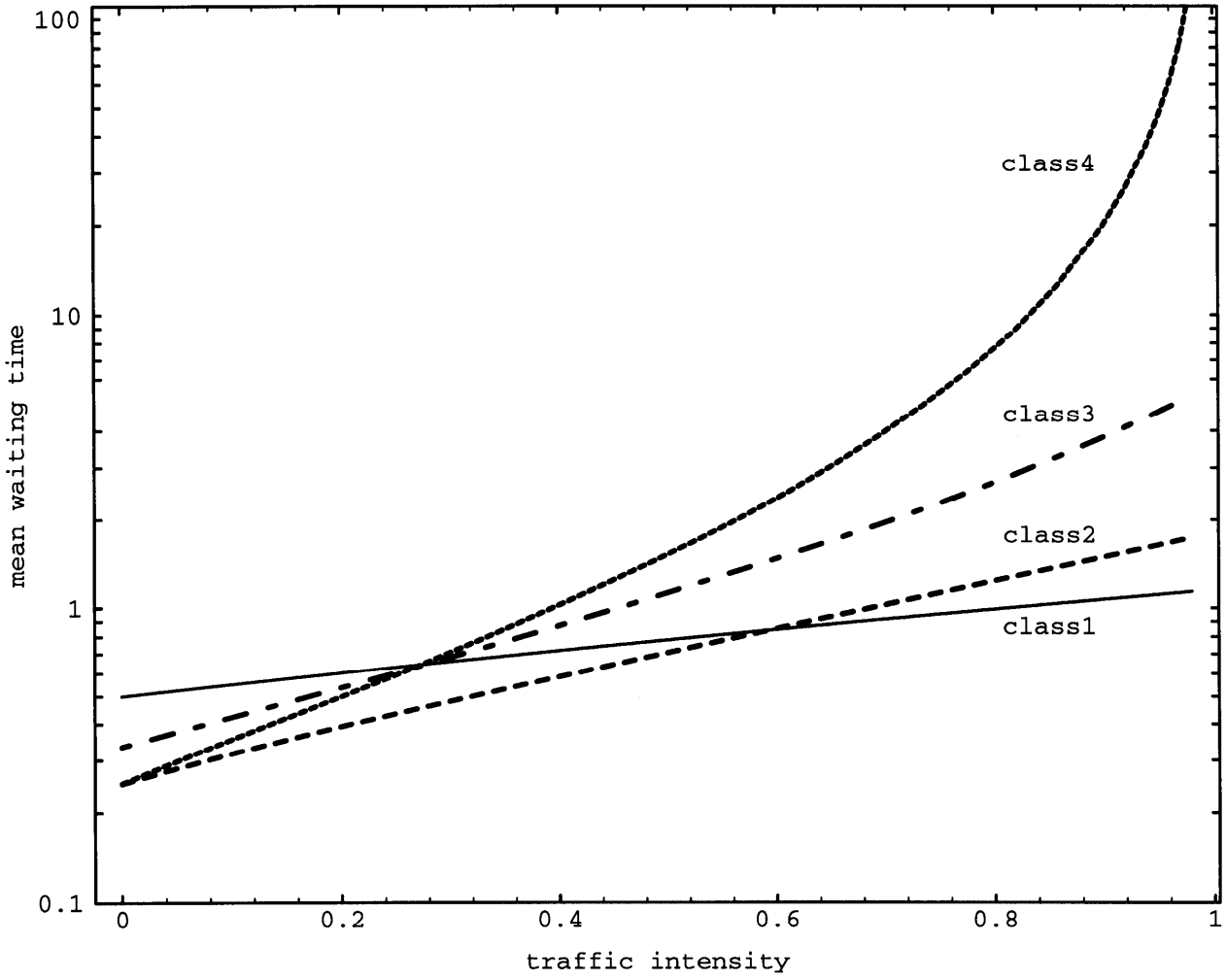

Figure 8: The mean waiting time in the NPNV model.

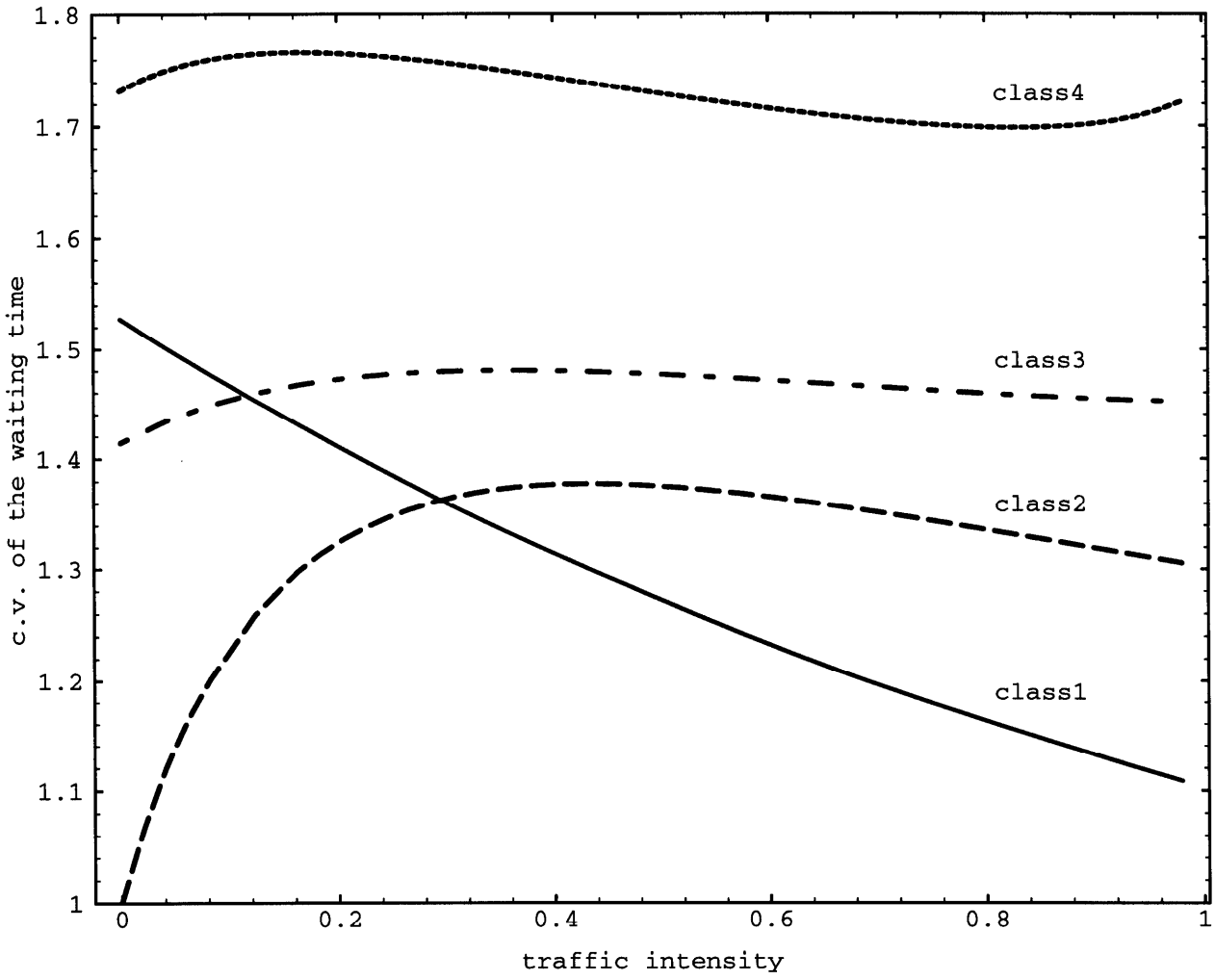

Figure 9: The coefficient of variation of the waiting time in the NPNV model.

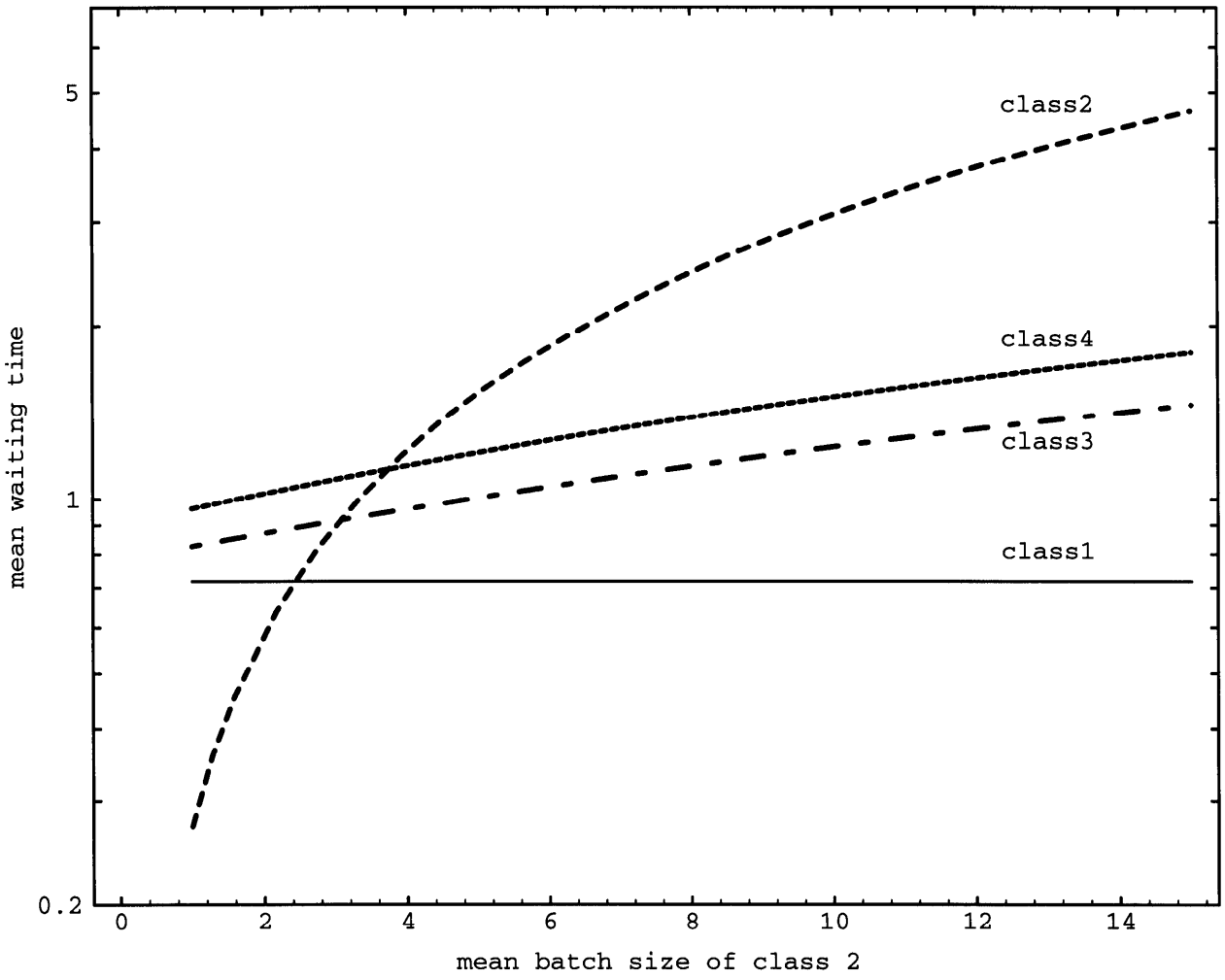

Figure 10: The mean waiting time vs. g_2 in the NPNV model.

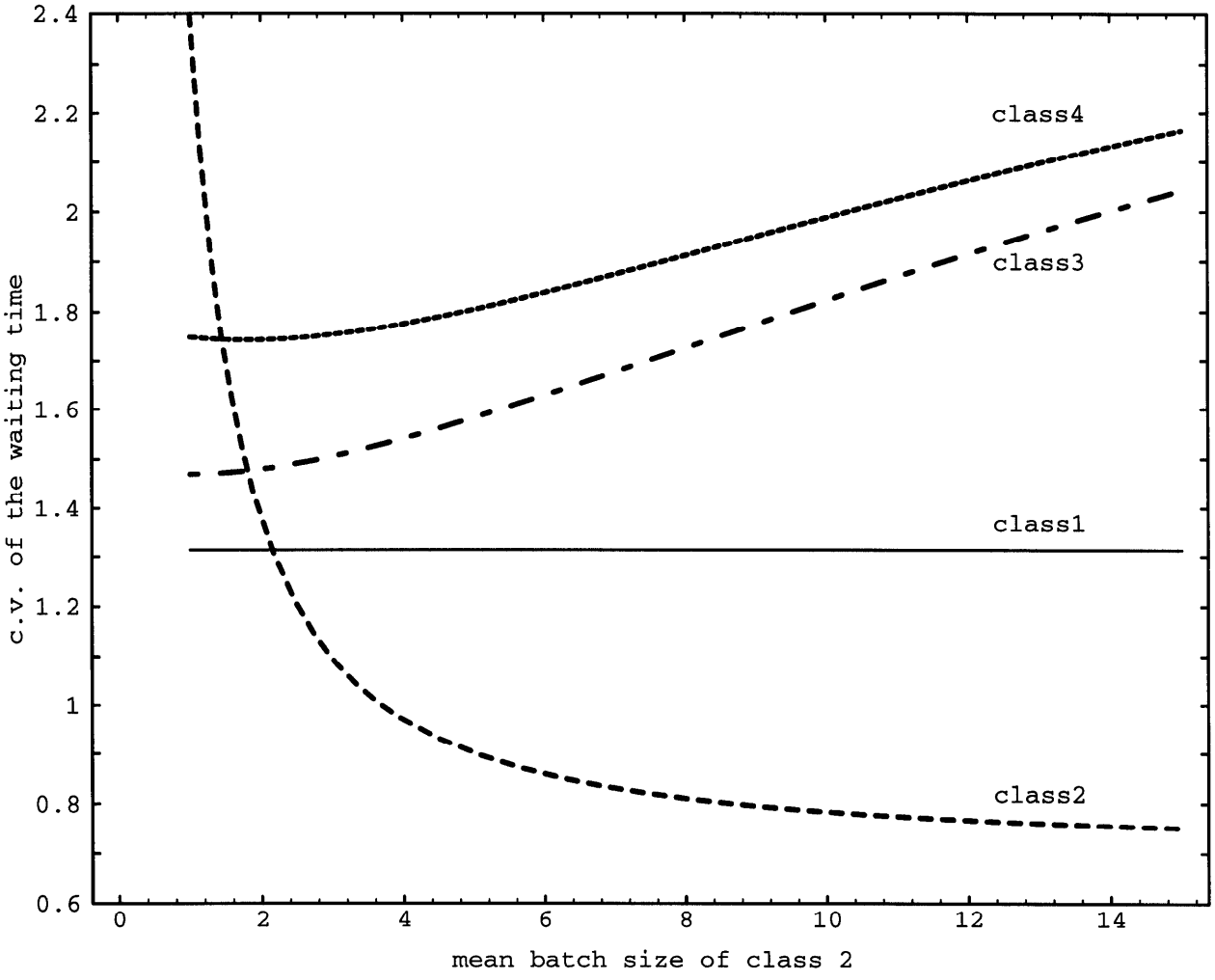

Figure 11: The coefficient of variation of the waiting time vs. g_2 in the NPNV model.

References

[1] Cobham, A., Priority assignment in waiting line problems, *Operations Research* 2 (1954) 70–76. Also, Cobham, A., Priority assignment - a correction, *Operations Research* 3 (1955) 547.

[2] Conolly, B., *Lecture Notes on Queueing Systems*, Ellis Horwood Limited, Sussex, England (1975).

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

[4] Cooper, R. B., *Introduction to Queueing Theory*, Second Edition, North-Holland Publishing Company, New York, 1981, (First Edition: Macmillan, New York (1972)), republished by the Continuing Engineering Education Program, The George Washington University, Washington D.C. (1990).

[5] Doshi, B. T., Queueing systems with vacations - a survey, *Queueing Systems* 1 (1986) 29–66.

[6] Durr, L., A single-server priority queueing system with general holding times, Poisson input, and reverse-order-of-arrival queueing discipline, *Operations Research* 17 (1969) 351–358.

[7] Durr, L., Priority queues with random order of service, *Operations Research* 19 (1971) 453–460.

[8] Fuhrmann, S. W., Second moment relationships for waiting times in queueing systems with Poisson input, *Queueing Systems* 8 (1991) 397–406.

[9] Fujiki, M. and Gambe, E., *Teletraffic Theory*, Maruzen, Tokyo (1980) (in Japanese).

[10] Gaver, D. P. Jr., A waiting line with interrupted service, including priorities, *Journal of the Royal Statistical Society, Series B* 24 (1962) 73–90.

[11] Heyman, D. P. and Sobel, M. J., *Stochastic Models in Operations Research, Volume I: Stochastic Processes and Operating Characteristics*, McGraw-Hill Publishing Company, New York (1982).

[12] Holley, J. L., Waiting line subject to priorities, *Operations Research* 2 (1954) 341–343.

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

[14] Kawasaki, N., Takagi, H., Takahashi, Y., and Hasegawa, T., Waiting time analysis of $M^X/G/1$ queues with/without vacations under random order of service discipline, Technical Report #95003, Department of Applied Mathematics and Physics, Faculty of Engineering, Kyoto University, Kyoto (1995).

[15] Kella, O. and Yechiali, U., Priorities in M/G/1 queue with server vacations, *Naval Research Logistics* 35 (1988) 23–24.

[16] Kesten, H. and Runnenburg, J. Th., Priority in waiting line problems, I and II, *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Series A* 60 (1957) 312–324 and 325–336.

[17] Kingman, J. F. C., On queues in which customers are served in random order, *Proceedings of the Cambridge Philosophical Society* 58 (1962) 79–91.

[18] Kleinrock, L., *Queueing Systems, Volume 1: Theory*, John Wiley and Sons, New York (1975).

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

[20] Little, J. D. C., A proof for the queuing formula: $L = \lambda W$, *Operations Research* 9 (1961) 383–387.

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

[22] Scholl, M. and Kleinrock, L., On the M/G/1 queue with rest periods and certain service-independent queueing disciplines, *Operations Research* 31 (1983) 705–719.

[23] Shanthikumar, J. G., Analysis of priority queues with server control, *Opsearch* 21 (1984) 183–192.

[24] Takács, L., Delay distributions for one line with Poisson input, general holding times, and various orders of service, *The Bell System Technical Journal* 42 (1963) 487–503.

[25] Takács, L., Priority queues, *Operations Research* 12 (1964) 63–74.

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

[27] Takagi, H. and Kudoh, S., Symbolic higher-order moments of the waiting time in an M/G/1 queue with random order of service, *Communications in Statistics-Stochastic Models* 13 (1997) 167–179.

[28] Takagi, H. and Takahashi, Y., Priority queues with batch Poisson arrivals, *Operations Research Letters* 10 (1991) 225–232.

[29] Welch, P. D., Some contributions to the theory of priority queues. Ph.D. Thesis, Department of Mathematical Statistics, Columbia University; IBM Research Report RC-922, IBM Research Center, Yorktown Heights, New York (1963).