Journal of Applied Mathematics and Stochastic Analysis
9, Number 2, 1996, 143-157

M/G/1/K SYSTEM WITH PUSH-OUT SCHEME UNDER VACATION POLICY

SHOJI KASAHARA
Kyoto University
Educational Center for Information Processing
Kyoto 606-01, Japan

HIDEAKI TAKAGI
University of Tsukuba
Institute of Socio-Economic Planning
1-1-1 Tennoudai, Tsukuba-shi
Ibaraki 305, Japan

YUTAKA TAKAHASHI and TOSHIHARU HASEGAWA
Kyoto University
Department of Applied Mathematics and Physics
Faculty of Engineering
Kyoto 606-01, Japan

(Received July, 1995; Revised November, 1995)

ABSTRACT

We consider an M/G/1/K system with push-out scheme and multiple vacations. This model is particularly important in situations where it is essential to provide short waiting times to messages which are selected for service. We analyze the behavior of two types of messages: one that succeeds in transmission and the other that fails. We derive the Laplace-Stieltjes transform of the waiting time distribution for the message which is eventually served. Finally, we show some numerical results including the comparisons between the push-out and the ordinary blocking models.

Key words: M/G/1/K, Push-out Scheme, Vacation, Blocking Model, FCFS, LCFS.

AMS (MOS) subject classifications: 60K25.

1. Introduction

This paper considers a queueing system with a finite buffer and server vacation. Messages are admitted into the system in accordance with an appropriate buffering policy. That is, a finite number of messages can be held in the system at any time since the system contains a finite capacity of a buffer. There are two control policies for processing messages. One is the buffering policy by which messages are selected for admission into the system. The other is the service policy by which messages are selected for admission into the service facility.

Buffering policies specify those messages that are admitted to enter and those to be removed from the buffer when the buffer is full. Rubin and Ouaily [6] classified the buffering policies into

the following types (Fig. 1).

- Non-Preemptive-Buffering (NPB)
  An arriving message that finds the system full is blocked.
- Preemptive-Buffering (PB)
  If an arriving message finds the system full, the message which has waited the longest time is pushed out from the buffer and the arriving message is allocated a buffer space.

The service policy determines the selection of messages waiting for service when the service facility becomes available. This policy includes, for example, First-Come First-Served (FCFS), Last-Come-First-Served (LCFS), and random order of service.

Queueing systems with a finite buffer and server vacation have been extensively studied to model and analyze a number of computer communication systems. In particular, queueing systems with buffering policy have many applications like time-critical message transmission, sensor telemetry, radar communication and processing systems. In those applications, the information content of a message is associated with a timeliness index, so that the most recent message to arrive contains the most valuable information, and thus needs to be given preference for selection for service. On the other hand, the data transmission is the primary job for those systems and, when there are no messages in the buffer, they start secondary jobs like testing and maintenance work. From a queueing theoretical point of view, those periods spent for the service of secondary jobs are considered as vacations.

Recently, with the increase of demands for multi-media communication, many protocols and architectures to accommodate traffics of different characteristics from multiple sources in a common channel have been proposed and implemented so far [1, 8]. In this communication environment, messages are classified from two orthogonal points of view, delay and loss probability [7]. Delay (loss probability) sensitive messages are insensitive to loss probability (delay) in general. These factors can be expressed by assigning a timeliness index to each message, which means after some critical value for its delay, each message becomes useless. For effective transmission of two types of messages, switching systems require the use of finite pre-emptive buffering service system since it is essential to provide short waiting time to those messages which are delay sensitive. If we focus our attention on the behavior of delay sensitive messages, the transmission of loss sensitive messages are considered as a secondary job for those switching systems. Thus, we can apply our model to evaluate the behavior of delay sensitive messages.

There are several literatures concerning buffering policies. A communication system under a pre-emptive buffering was investigated by Rubin and Ouaily in the context of an M/G/1/K with push-put scheme [6]. (A technical error in the analysis of Rubin and Ouaily is pointed out in this paper.) Kröner analyzed loss probabilities for a partial buffer sharing scheme under FCFS [4]. Sumita and Ozawa analyzed loss probabilities and the waiting time of systems with a push-out scheme [7].

Concerning queueing systems with server vacation, there are a number of previous works. An excellent survey of queueing systems with vacations, including some applications, was written by Doshi [2, 3]. An M/G/1/K with multiple vacation has also been analyzed by Lee [5], but not analytical results are available for the model with push-out scheme.

This paper is organized as follows: In section 2, we describe our mathematical model in detail. In section 3, we derive the relation of the mean waiting times for NPB, PB-served and PB-pushed-out messages. We also summarize Lee's results [5] to obtain the joint probability distributions for the number of messages in the system and the remaining service or vacation time. In section 4, the Laplace-Stieltjes transform (LST) of the waiting time distribution for an eventually served message is derived. In section 5, we show the numerical results.

2. Model

We consider an M/G/1/K push-out model with multiple vacations (Fig. 2). Messages arrive at the system according to a Poisson process with rate λ. The service time distribution function and its LST are denoted by $S(x)$ and $S^*(s)$, respectively. The mean service time is $1/\mu$.

When the system becomes idle, the server takes a vacation. The vacation policy of our model is multiple vacations. The server takes vacations repeatedly until it finds at least one waiting message accommodated upon returning from a vacation. The vacation time distribution function and its LST are denoted by $V(x)$ and $V^*(s)$, respectively. The mean vacation time is $1/v$.

The maximum number of messages that can be present in the system is $K(<\infty)$. When a message is in service, the maximum number of messages in the buffer is $K-1$. The buffering policy determines which to discard out of $K-1$ messages (K messages) to accommodate a newly arriving message when the server is busy (taking a vacation) and the system is full.

We consider the following buffering policy: When a new message finds the system full, a message with the longest sojourn time in the buffer is pushed out and lost.

We deal with two service disciplines, FCFS and LCFS.

3. Queue Size Distribution and Mean Waiting Time

3.1 Queue size distribution

Since the message loss happens only when the system is full, the queue length distributions in the NPB and the PB schemes are identical. Thus, we can apply Lee's results for the NPB model to the PB model [5]. This section summarizes his results.

We choose a set of imbedded Markov points at those points in time when either a service is completed or a vacation ends. Let L_n be the number of messages in the system immediately after the nth Markov point, and let

$$\eta_n = \begin{cases} 0 & \text{if a vacation ends,} \\ 1 & \text{if a service is completed,} \end{cases} \tag{1}$$

at the nth Markov point. We consider the limiting probability distributions

$$\omega_k \equiv \lim_{n\to\infty} Prob[\eta_n = 0, L_n = k], \qquad 0 \le k \le K$$

$$\pi_k \equiv \lim_{n\to\infty} Prob[\eta_n = 1, L_n = k], \qquad 0 \le k \le K-1,$$

which satisfy the following equations

$$\begin{aligned} \omega_k &= (\omega_0 + \pi_0) f_k, \quad 0 \le k \le K-1, \\ \omega_K &= (\omega_0 + \pi_0) \textstyle\sum_{m=K}^{\infty} f_m, \\ \pi_k &= \textstyle\sum_{j=1}^{k+1} (\omega_j + \pi_j) a_{k-j+1}, \quad 0 \le k \le K-2, \\ \pi_{K-1} &= \omega_K + \textstyle\sum_{j=1}^{k-1} (\omega_j + \pi_j) \sum_{m=K-j}^{\infty} a_m, \end{aligned} \tag{2}$$

and

$$\sum_{k=0}^{K} \omega_k + \sum_{k=0}^{K-1} \pi_k = 1, \tag{3}$$

where

$$a_k \equiv \int_0^\infty \frac{(\lambda x)^k}{k!} e^{-\lambda x} dS(x), \quad k = 0, 1, 2, \ldots, \tag{4}$$

and

$$f_k \equiv \int_0^\infty \frac{(\lambda x)^k}{k!} e^{-\lambda x} dV(x), \quad k = 0, 1, 2, \ldots. \tag{5}$$

From (2) and (3), we can obtain $\omega_k (k = 0, \ldots, K)$ and $\pi_k (k = 0, \ldots, K-1)$.

Let L, ρ' and P_B denote the queue size, the carried load and the probability that an arriving message is blocked in the NPB model, respectively. The carried load ρ' and the blocking probability P_B are given by

$$\rho' = \frac{(1 - \omega_0 - \pi_0)/\mu}{(\omega_0 + \pi_0)/v + (1 - \omega_0 - \pi_0)/\mu}, \tag{6}$$

$$P_B = 1 - \frac{\rho'}{\rho}, \tag{7}$$

where $\rho = \lambda/\mu$.

The joint distribution of the state ξ of the server, the number L of messages in the system, and the remaining vacation time $\hat{V}$ when the server is on vacation or the remaining service time $\hat{S}$ when the server is busy at an arbitrary instant is given as follows. If we define the state ξ of the server as

$$\xi = \begin{cases} 0 & \text{the server is on vacation,} \\ 1 & \text{the server is busy,} \end{cases} \tag{8}$$

and also define the following equations

$$\Omega_k^*(s) \equiv \int_0^\infty e^{-sx} Prob[\xi = 0, L = k, x < \hat{V} < x + dx], \quad 0 \le k \le K,$$

$$\Pi_k^*(s) \equiv \int_0^\infty e^{-sx} Prob[\xi = 1, L = k, x < \hat{S} < x + dx], \quad 1 \le k \le K,$$

we obtain (see [5] for details)

$$\Pi_k^*(s) = \frac{\sigma}{\lambda} \left[S^*(s) \sum_{j=1}^{k} (\omega_j + \pi_j) \left(\frac{\lambda}{\lambda - s} \right)^{k-j+1} - \sum_{j=0}^{k-1} \pi_j \left(\frac{\lambda}{\lambda - s} \right)^{k-j} \right], \quad 1 \le k \le K-1, \tag{9}$$

$$\Pi_K^*(s) = -\frac{\sigma}{s} \left\{ S^*(s) \left[\sum_{j=1}^{K-1} (\omega_j + \pi_j) \left(\frac{\lambda}{\lambda - s} \right)^{K-j} + \omega_K \right] - \sum_{j=0}^{K-1} \pi_j \left(\frac{\lambda}{\lambda - s} \right)^{K-j-1} \right\}, \tag{10}$$

$$\Omega_k^*(s) = \frac{\sigma}{\lambda}\left[V^*(s)(\omega_0+\pi_0)\left(\frac{\lambda}{\lambda-s}\right)^{k+1} - \sum_{j=0}^{k}\omega_j\left(\frac{\lambda}{\lambda-s}\right)^{k-j+1}\right],\ 0 \le k \le K-1, \tag{11}$$

$$\Omega_K^*(s) = -\frac{\sigma}{s}\left[V^*(s)(\omega_0+\pi_0)\left(\frac{\lambda}{\lambda-s}\right)^{K} - \sum_{j=0}^{K}\omega_j\left(\frac{\lambda}{\lambda-s}\right)^{K-j}\right], \tag{12}$$

where

$$\sigma = \frac{1}{(\omega_0+\pi_0)/v + (1-\omega_0+\pi_0)/\mu}. \tag{13}$$

For later use, let $\Omega_k(x)$ and $\Pi_k(x)$ be the inverse transforms of LST $\Omega_k^*(s)$ and $\Pi_k^*(s)$, respectively.

3.2 Mean waiting time

Following the approach of [6], we consider the relation between the mean waiting time of NPB model and that of PB model. Let W_P denote the waiting time during which a message stays in the buffer in PB model. We then have

$$E[W_P] = E[W_P \mid \text{served}]Prob[\text{served}] + E[W_P \mid \text{pushed-out}]Prob[\text{pushed-out}]. \tag{14}$$

Let γ denote the system throughput. In both NPB and PB models, the event that a message is lost occurs when the system is full. Note that the stochastic behavior of the number of messages in the system does not depend on our buffering policy. Hence, L, γ and ρ' are invariant in the NPB and PB models with multiple vacations. Let W_B be the waiting time of a message accepted in the NPB model. Applying Little's theorem to those messages present in the queue, we have

$$\gamma E[W_B] = E[L] - \rho' = \lambda E[W_P]. \tag{15}$$

Since $\gamma \le \lambda$, it follows that

$$E[W_P] \le E[W_B]. \tag{16}$$

Considering the throughput γ, we have

$$\gamma = \lambda(1-P_B) = \lambda(1 - Prob[\text{pushed-out}]). \tag{17}$$

Hence, we obtain

$$Prob[\text{pushed-out}] = P_B, \tag{18}$$

and

$$\begin{aligned} Prob[\text{served}] &= 1 - Prob[\text{pushed-out}] \\ &= 1 - P_B. \end{aligned} \tag{19}$$

Substituting (18) and (19) into (14), we have

$$\lambda E[W_P] = \lambda(1-P_B)E[W_P \mid \text{served}] + \lambda P_B E[W_P \mid \text{pushed-out}]. \tag{20}$$

From (15) and (17), we obtain

$$\lambda E[W_P] = \lambda(1 - P_B)E[W_B]. \tag{21}$$

From (20) and (21), $E[W_P \mid \text{pushed-out}]$ is given by

$$E[W_P \mid \text{pushed-out}] = \frac{1 - P_B}{P_B}(E[W_B] - E[W_P \mid \text{served}]). \tag{22}$$

Thus, we can calculate the mean sojourn time of a pushed-out message from (22) if we obtain $E[W_P \mid \text{served}]$.

4. Waiting Time Distribution for Served Messages

4.1 FCFS systems

We first consider the push-out system under FCFS service discipline. Each arriving message joins the queue at the tail and if the system is full upon arrival, the message at the head of the queue is pushed-out.

Let $W_{k:n}$ denote the waiting time of a tagged message that has k other messages ahead and n others behind it at the end of a service or a vacation. We also define the following LST,

$$W^*_{k:n}(s) = E[e^{-sW_{k:n}} \mid \text{served}]Prob[\text{served}], \tag{23}$$

where $0 \le k \le K-1$ and $0 \le n \le K-k-1$ at the end of a vacation, and $0 \le k \le K-2$ and $0 \le n \le K-k-2$ at the end of a service. Note that the LST $W^*_{k:n}(s)$ is the same in both cases of a vacation and a service.

To set $\{W^*_{k:n}(s); 0 \le k \le K-1,\ 0 \le n \le K-k-1\}$ satisfies the following equations.

$$W^*_{0:n}(s) = 1, \quad 0 \le n \le K-2, \tag{24}$$

$$W^*_{k:n}(s) = \sum_{j=0}^{K-k-n-1} S^*_j(s) \cdot W^*_{k-1:n+j}(s) + \sum_{j=K-k-n}^{K-n-2} S^*_j(s) \cdot W^*_{K-n-j-2:n+j}(s),$$

$$1 \le k \le K-1, \quad 0 \le n \le K-k-1, \tag{25}$$

where

$$S^*_j(s) = \int_0^\infty \frac{(\lambda x)^j}{j!} e^{-(s+\lambda)x} dS(x). \tag{26}$$

Using the above LSTs, the LST $W^*(s)$ of the distribution function for the waiting time of a served message in the FCFS system is given by

$$W^*(s) = \frac{1}{1-P_B}\left[\sum_{j=0}^{K-1}\left\{\sum_{k=0}^{K-j-2}\Omega^*_{j:k}(s)\cdot W^*_{j:k}(s) + \sum_{k=K-j-1}^{K-1}\Omega^*_{j:k}(s)\cdot W^*_{K-k-1:k}(s)\right\} + \sum_{k=0}^{K-1}\Omega^*_{K:k}(s)\cdot W^*_{K-k-1:k}(s)\right]$$

$$+\sum_{j=1}^{K-1}\left\{\sum_{k=0}^{K-j-1}\Pi^*_{j:k}(s)\cdot W^*_{j-1:k}(s)+\sum_{k=K-j}^{K-2}\Pi^*_{j:k}(s)\cdot W^*_{K-k-2:k}(s)\right\}$$

$$+\sum_{k=0}^{K-2}\Pi^*_{K:k}(s)\cdot W^*_{K-k-2:k}(s)\Bigg], \tag{27}$$

where

$$\Omega^*_{j:k}(s)=\int_0^\infty \frac{(\lambda x)^k}{k!}e^{-(s+\lambda)x}d\Omega_j(x), \qquad 0\le j\le K, \tag{28}$$

and

$$\Pi^*_{j:k}(s)=\int_0^\infty \frac{(\lambda x)^k}{k!}e^{-(s+\lambda)x}d\Pi_j(x), \quad 1\le j\le K. \tag{29}$$

In [6], there is a technical error. The waiting time distribution of a served message $W(t)$ is given by

$$W(t)=\pi_0+\sum_{n=1}^{K}\pi_n[R(t)*B^{(n-1)}(t)],$$

where π_n's are the steady state probabilities that an arriving message finds n messages in the system, $B(t)$ is the service time distribution, $R(t)$ is the remaining service time distribution, $*$ denotes the convolution and $B^{(n-1)}(t)$ is the $n-1$th-convolution. In that equation, the number of messages at an arriving epoch and the remaining service time are treated as being independent, but that is incorrect. The number of messages at an arriving epoch is not independent of the remaining service time. Thus, we have to use the joint distribution of the number of messages and the remaining service time. (We show the corrected LST of the waiting time distribution in the Appendix.)

4.2 LCFS systems

We next consider the LCFS system. Each arriving message joins the queue at the head and, if the system is full, the message at the tail is pushed-out.

As in the case of FCFS, let $\widetilde{W}_k$ denote the waiting time of a tagged message that has k other messages ahead at the end of a service or a vacation. We define the following LST,

$$\widetilde{W}_k(s)=E[e^{-s\widetilde{W}_k}\,|\,\text{served}]Prob[\text{served}], \tag{30}$$

where $0\le k\le K-1$ at the end of a vacation, and $0\le k\le K-2$ at the end of a service. Note that the LST $\widetilde{W}_k(s)$ is the same in the cases of both a vacation and a service.

The set $\{\widetilde{W}_k(s);0\le k\le K-1\}$ satisfies the following equations:

$$\widetilde{W}_0(s)=1, \tag{31}$$

$$\widetilde{W}_k(s)=\sum_{j=0}^{K-k-1}S^*_j(s)\cdot\widetilde{W}_{k+j-1}(s), \quad 1\le k\le K-1. \tag{32}$$

For simplicity, we define the following LSTs:

$$\widehat{S}^*_j(s)=\int_0^\infty \frac{(\lambda x)^j}{j!}e^{-(s+\lambda)x}d\widehat{S}(x), \tag{33}$$

$$\widehat{V}_j^*(s) = \int_0^\infty \frac{(\lambda x)^j}{j!} e^{-(s+\lambda)x} d\widehat{V}(x), \tag{34}$$

where $\widehat{S}(x) = Prob[\widehat{S} \leq x]$ and $\widehat{V}(x) = Prob[\widehat{V} \leq x]$. If k messages arrive at the system during the remaining vacation or service time, the tagged message has k messages ahead at the end of this vacation or service. Thus, we have the LST of the distribution function for the waiting time of a served message in the LCFS, $W^*(s)$ by

$$W^*(s) = \frac{1}{1-P_B}\left[(1-\rho')\sum_{k=0}^{K-1}\widehat{V}_k^*(s)\cdot\widetilde{W}_k(s) + \rho'\sum_{k=0}^{K-2}\widehat{S}_k^*(s)\cdot\widetilde{W}_k(s)\right]. \tag{35}$$

5. Numerical Results

In this section, we show the numerical results for the mean and the coefficient of variation (c.v.) of the waiting time using the analysis presented in section 4.

In our numerical examples, we choose the system size K equal to 5, that is, the buffer size equals 4. As for vacation times, we assume an exponential distribution with mean 1.0. The mean service time is fixed at 1.0, and the performance values are calculated by changing the arrival rate.

First, we compare the mean waiting time under various situations. Using (22), (27) and (35), we calculate the mean waiting times for served and pushed-out messages. From [5], the mean waiting time for NPB model can be also calculated.

Fig. 3 and Fig. 4 illustrate the mean waiting time for three types of messages: NPB, PB-served and PB-pushed-out. Furthermore, mean waiting times under the exponential service distribution are compared with those under deterministic one.

In both figures, the mean waiting times of NPB and PB-served messages tend to the value of 1 as the offered load gets small. This is because each arriving message most likely waits for the remaining vacation time. On the other hand, the mean sojourn time of a pushed-out message is larger than those of others. This phenomenon can be explained as follows. When the arrival rate is very small, there are few messages in the system. Thus, most of arriving messages are eventually served. However, if an arriving message is eventually pushed-out, its sojourn time becomes large due to light traffic.

Next, when the offered load gets large, the mean waiting times for all types of messages converge under exponential and constant service times, in particular, PB-served and PB-pushed-out messages converge to the same value. In both NPB and PB cases, a new arriving message which can be accommodated in the system finds four other messages (including the message in service) ahead when the arrival rate is very large. Hence, the mean waiting time of NPB messages converges to the value 4. In PB case, a new arriving message can enter the system. But there are many other new arriving messages behind that and the probability that the tagged message is eventually served gets small. Thus, the mean waiting times in the buffer of both messages become small.

In FCFS (Fig. 3), the mean waiting time of a PB-pushed-out message is bounded by 5, because each arriving message finds at most five messages ahead. On the other hand, in LCFS (Fig. 4), it may exceed 5. This is because there is no bound on the number of the messages which are served before the service of the tagged one.

In Fig. 4 the mean waiting time of a PB-pushed-out message under the deterministic service time distribution fluctuates remarkably when the arrival rate is small. It can be considered that

under deterministic service distribution, the mean waiting time of a PB-pushed-out message is influenced by the loss probability and the waiting time of a PB-served one.

One more interesting observation is the relation of the mean waiting time between PB-served and PB-pushed-out messages under two service time distributions. Let $W_{A:B}[C]$ denote the mean waiting time of a 'C' type message under 'A' service discipline and 'B' service distribution.

In Fig. 3, it is observed that $W_{\text{FCFS:Exp}}[\text{Served}] \leq W_{\text{FCFS:Exp}}[\text{Pushed-out}]$, i.e., the mean waiting time of a served message is always smaller than that of a pushed-out one. On the other hand, under the deterministic service time distribution, we see that

$$W_{\text{FCFS:Exp}}[\text{Served}] \leq W_{\text{FCFS:Exp}}[\text{Pushed-out}], \quad 0 \leq \rho \leq 1, \tag{36}$$

$$W_{\text{FCFS:Det}}[\text{Served}] > W_{\text{FCFS:Det}}[\text{Pushed-out}], \quad \rho > 1. \tag{37}$$

In the LCFS case, we can observe the following relations,

$$W_{\text{LCFS:Det}}[\text{Served}] < W_{\text{LCFS:Det}}[\text{Pushed-out}], \quad 0 \leq \rho, \tag{38}$$

$$W_{\text{LCFS:Det}}[\text{Served}] < W_{\text{LCFS:Det}}[\text{Pushed-out}], \quad 0 \leq \rho. \tag{39}$$

Equations (38) and (39) show that the mean waiting time of the served message is always smaller than that of the pushed-out one under both service distributions. Thus, in FCFS, the mean waiting times of the served and pushed-out messages are more influenced by the type of service distribution.

In Fig. 5 and Fig. 6, mean waiting times are compared for two-pushed-out models: the system with vacation and that without vacation. We can calculate the mean waiting time of the system without vacation by [6] (see Appendix). In both figures, we assume $S(x)$ to be exponential (mean service time $= 1.0$). From both figures, we can observe the influence of vacation when the offered load is small. Furthermore, when the offered load becomes large, each mean waiting time converges to the same value. This is because taking vacations hardly affects the performance measures when the offered load is large.

Fig. 7 illustrates the c.v. of the waiting time of the PB-served message under two service time disciplines and two service distributions. In both FCFS and LCFS cases, the values start from 1 because the vacation distribution is exponential and its mean equals 1.0. We also observe that both curves converge rapidly. This means that the fluctuation of the waiting time is small when the offered load becomes large. We note that the variation under LCFS is larger than that under FCFS.

Fig. 8 illustrates the c.v. of the waiting time for NPB and PB-served messages with and without vacation. When the offered load is small, the influence of vacation is recognized. In FCFS cases, c.v.'s converge to the same value when the offered load is large. On the other hand, in LCFS cases, c.v.'s of the PB-served message with and without vacation converge to the same value but that of NPB model diverges to infinity. We observe that the waiting time of the PB-served message with vacation varies least in both FCFS and LCFS cases.

6. Conclusion

In this paper, we have considered a buffer controlling policy, called push-out scheme. We investigated the behaviors of the two types of messages, one is eventually served and the other is pushed-out from the system.

From the numerical results, the following has been found. First, the mean waiting times of NPB and PB-served messages significantly depends on the remaining vacation time. In such a situation, the waiting time of the PB-pushed-out message is larger than others. The mean waiting times of PB-served and PB-pushed-out messages converge as the arrival rate gets large, and those limiting values are smaller than that under NPB case. This is due to the push-out scheme. We found that the mean waiting times under PB case are influenced by the service time distribution. Furthermore, the variation of the waiting time of the PB-served message is small and stable in comparison with that of the NPB one.

Appendix

Waiting time distribution for non-vacation case

In this appendix, we show the results of LSTs of the waiting time distribution for the M/G/1/K with push-out scheme under non-vacation [5, 6].

In the non-vacation case, we choose a set of imbedded Markov points at those epochs when a service is completed. Then, we define the following limiting probability distributions:

$$\pi_k \equiv \lim_{n\to\infty} Prob[L_n = k], \quad k = 0,1,2,\ldots,K-1, \tag{i}$$

where L_n is the number of messages in the system just after the service completion point. The set $\{\pi_k; 0 \le k \le K-1\}$ satisfies the following equations

$$\pi_k = \pi_0 a_k + \sum_{j=1}^{k+1} \pi_j a_{k-j+1}, \qquad 0 \le k \le K-2, \tag{ii}$$

$$\pi_{K-1} = \pi_0 \left(1 - \sum_{j=0}^{K-2} a_k\right) + \sum_{j=1}^{K-1} \pi_j \left(1 - \sum_{k=0}^{K-j-1} a_k\right), \tag{iii}$$

$$\sum_{j=0}^{K-1} \pi_k = 1. \tag{iv}$$

From the above equations, we can determine the values of $\{\pi_k\}$.

Let $\Pi_k(x)$ denote the joint probability distribution that the queue length is k and the remaining service time is less than x at an arbitrary time. Let $\Pi_k^*(s)$ denote the LST of $\Pi_k(x)$. Using $\{\pi_k\}$ $(0 \le k \le K-1)$, we obtain LSTs as

$$\Pi_k^*(s) = \frac{1}{\pi_0 + \rho}\left[S^*(s)\left\{\pi_0\left(\frac{\lambda}{\lambda - s}\right)^k + \sum_{j=1}^{k} \pi_j \left(\frac{\lambda}{\lambda - s}\right)^{k-j+1}\right\}\right.$$

$$\left. - \sum_{j=0}^{k-1} \pi_j \left(\frac{\lambda}{\lambda - s}\right)^{k-j}\right], \quad 1 \le k \le K-1, \tag{v}$$

$$\Pi_K^*(s) = -\frac{1}{(\pi_0 + \rho)s}\left[S^*(s)\left\{\pi_0\left(\frac{\lambda}{\lambda - s}\right)^{K-1} + \sum_{j=1}^{K-1} \pi_j \left(\frac{\lambda}{\lambda - s}\right)^{K-1}\right\}\right.$$

$$-\sum_{j=0}^{K-1}\pi_j\left(\frac{\lambda}{\lambda-s}\right)^{K-j-1}\Bigg]. \tag{vi}$$

where $\rho=\lambda/\mu$. Using (23), we obtain the LST of the waiting time of a served message under FCFS as

$$W^*(s)=\pi_0+(\pi_0+\rho)\left[\sum_{j=1}^{K-1}\left\{\sum_{k=0}^{K-j-1}\Pi^*_{j:k}(s)\cdot W^*_{j-1:k}\right.\right.$$

$$\left.\left.+\sum_{k=K-j}^{K-2}\Pi^*_{j:k}(s)\cdot W^*_{K-k-2:k}(s)\right\}+\sum_{k=0}^{K-2}\Pi^*_{K:k}(s)\cdot W^*_{K-k-2:k}(s)\right], \tag{vii}$$

where

$$\Pi^*_{j:k}(s)=\int_0^\infty\frac{(\lambda x)^k}{k!}e^{-(s+\lambda)x}d\Pi_j(x),\quad 1\le j\le K. \tag{viii}$$

Using (30) and (33), we also obtain the LST of the waiting time under LCFS as

$$W^*(s)=\pi_0+\rho\sum_{j=0}^{K-2}\widehat{S}^*_j(s)\cdot\widetilde{W}_j(s). \tag{ix}$$

References

[1] Armbruster, H. and Arndt, G., Broadband communication and its realization with broadband ISDN, *IEEE Commun. Mag.* **25**:11 (1987), 8-19.

[2] Doshi, B.T., Queueing systems with vacations - A survey, *Queueing Sys.* **1** (1986), 29-66.

[3] Doshi, B.T., Single server queues with vacations, *Stochastic Analysis of Comp. and Commun. System*, Elsevier Science Publishers B.V., North-Holland (1990), 217-265.

[4] Kröner, H., Comparative performance study of space priority mechanisms for ATM networks, *IEEE INFOCOM '90* (1990), 1136-1143.

[5] Lee, T.T., M/G/1/N queue with vacation time and exhaustive service discipline, *Oper. Res.* **32**:4 (1984), 774-784.

[6] Rubin, I. and Ouaily, M., Performance of finite capacity communication and queueing systems under various service and buffer preemptive policies, *IEEE INFOCOM '88* (1988), 505-514.

[7] Sumita, S. and Ozawa, T., Achievability of performance objectives in ATM switching nodes, *International Seminar on Performance of Distributed and Parallel Systems*, North-Holland, Amsterdam (1988), 45-56.

[8] Turner, J.S., New directions in communications (or Which way to the information age?), *IEEE Commun. Mag.* **24**:10 (1986), 8-15.

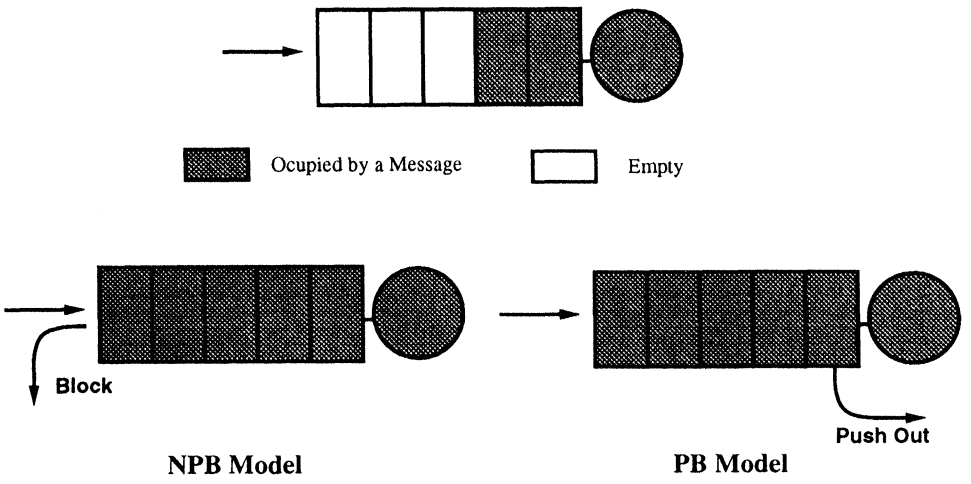

Figure 1: NPB and PB Models

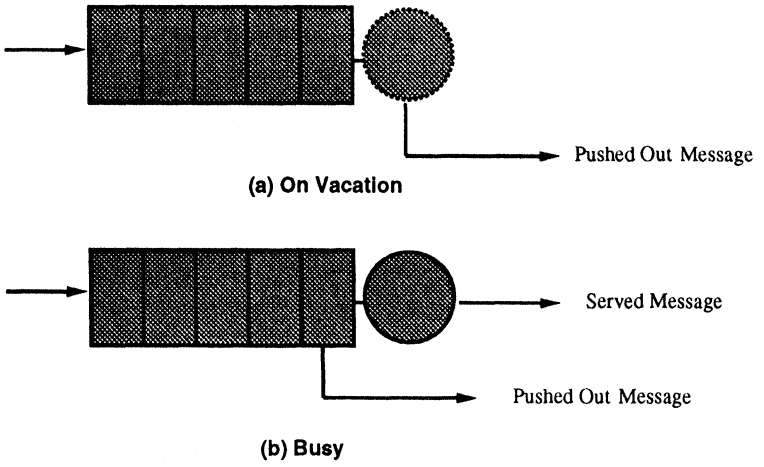

Figure 2: Push-out Model with Vacation

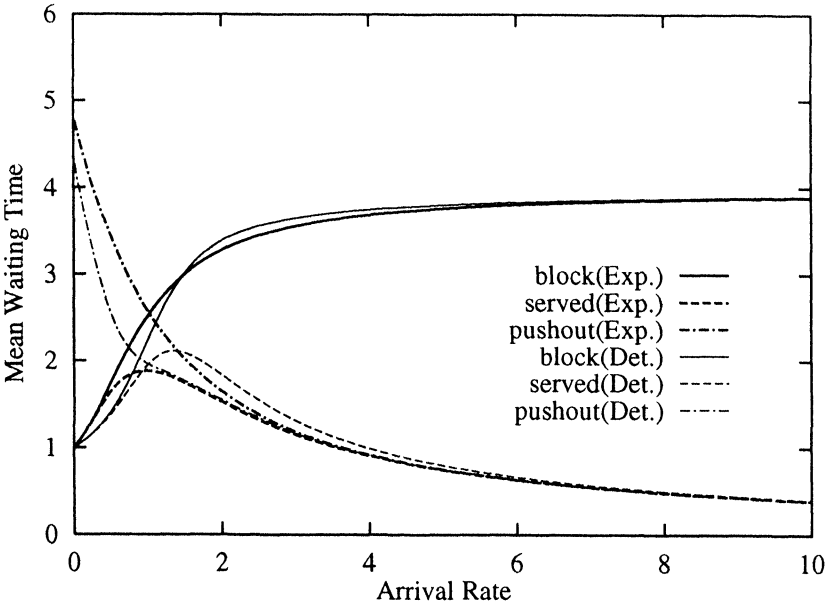

Figure 3: Mean Waiting Time under FCFS

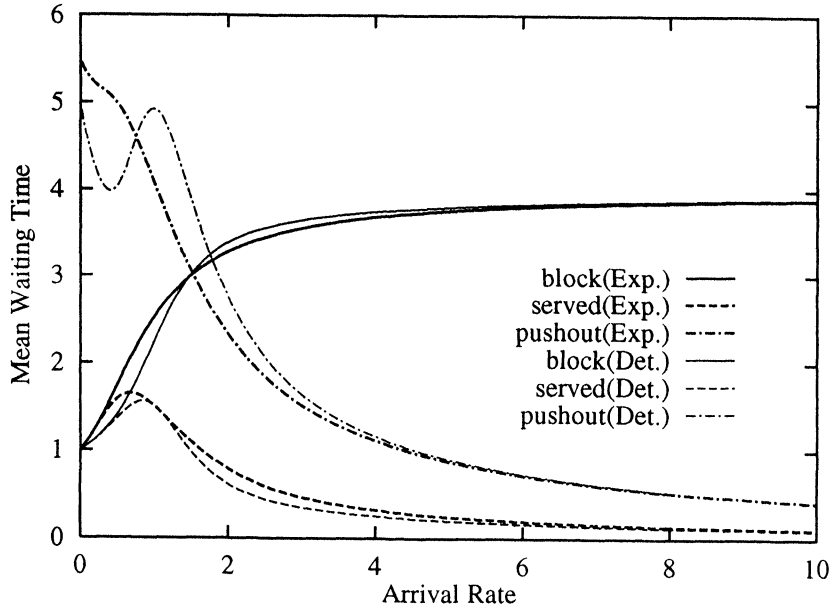

Figure 4: Mean Waiting Time under LCFS

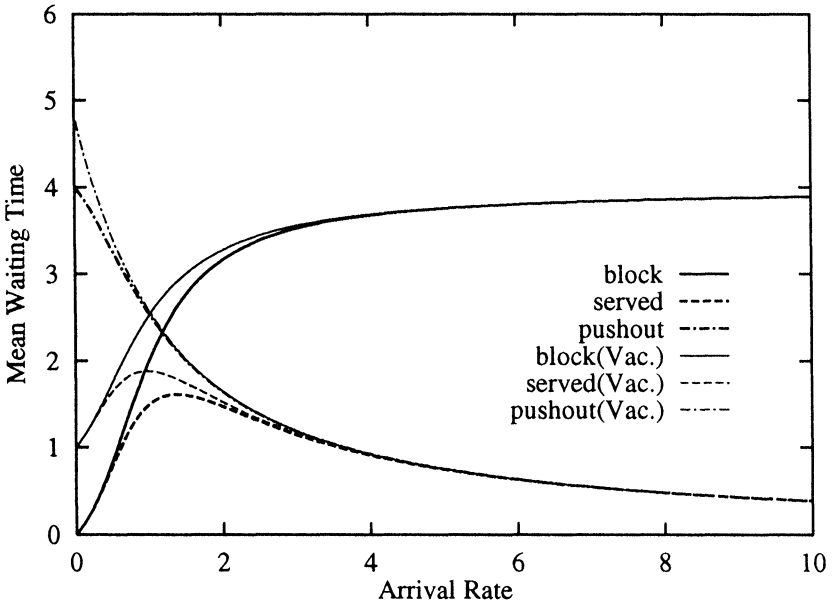

Figure 5: Mean Waiting Time under FCFS (non-Vac. vs. Vac.)

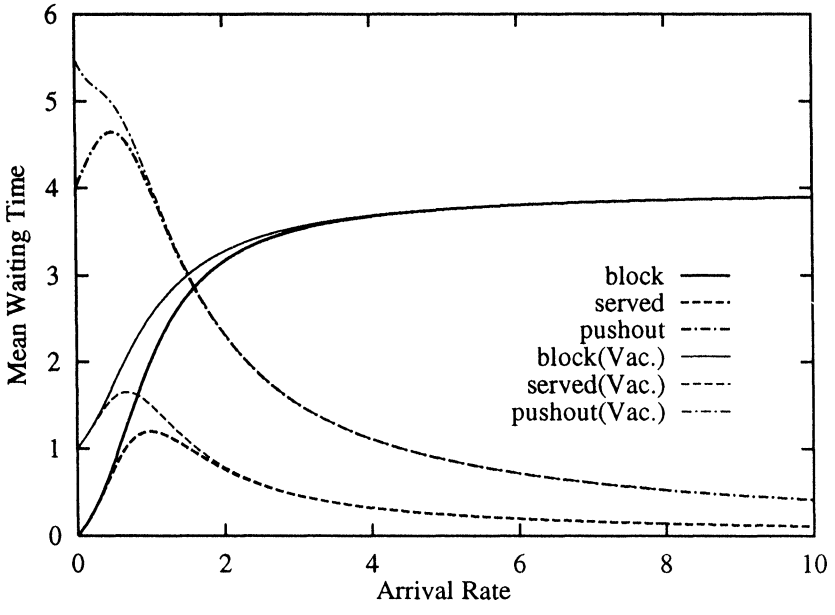

Figure 6: Mean Waiting Time under LCFS (non-Vac. vs. Vac.)

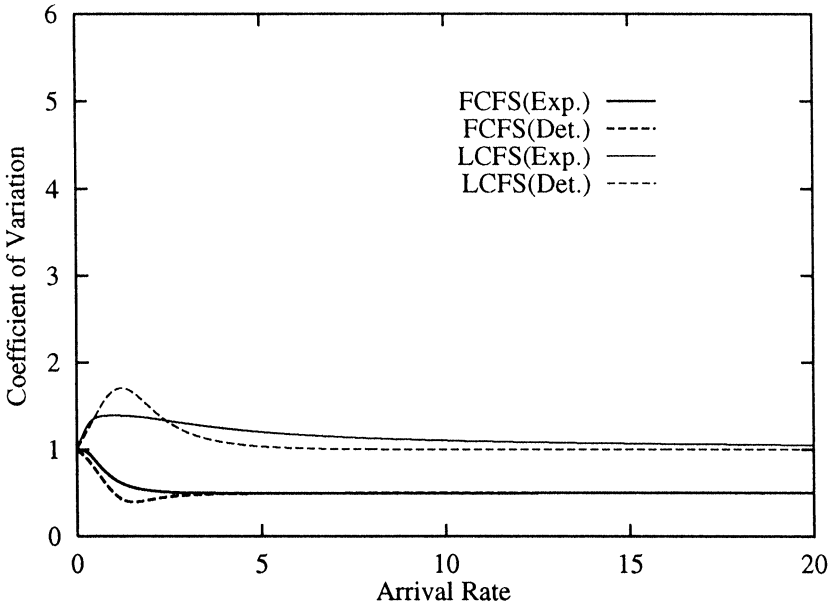

Figure 7: C.V. of Waiting Time

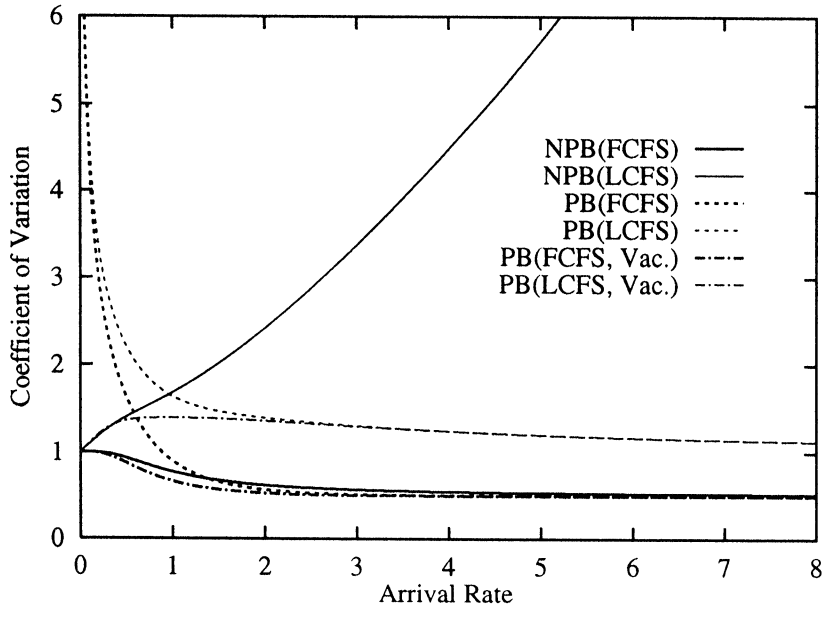

Figure 8: C.V. of Waiting Time