Finite Buffer Polling Models with Routing

Scott E. Grasman [1]

grasmans@umr.edu

219 Engineering Management

University of Missouri - Rolla, Rolla, Missouri 65409-0370

Tava Lennon Olsen

olsen@olin.wustl.edu

John M. Olin School of Business, Campus Box 1133

Washington University in St. Louis, St. Louis, Missouri 63130-4899

John R. Birge

jrbirge@nwu.edu

McCormick School of Engineering and Applied Science

Office of the Dean, 2145 Sheridan Road

Northwestern University, Evanston, Illinois 60208

April 2002, revised April 2003

[1] Corresponding Author

Abstract

This paper analyzes a finite buffer polling system with routing. Finite buffers are used to model the limited capacity of the system, and routing is used to represent the need for additional service. The most significant result of the analysis is the derivation of the generating function for queue length when buffer sizes are limited and a representation of the system workload. The queue lengths at polling instants are determined by solving a system of recursive equations, and an embedded Markov chain analysis and numerical inversion are used to derive the queue length distributions. This system may be used to represent production models with setups and lost sales or expediting.

Keywords: Queueing, polling, routing, setups

1 Introduction

A polling model is a system of multiple queues served by a single server, which requires a setup when switching queues. Polling models have been used extensively to model many computer and communications systems (see, e.g., Levy and Sidi [8] and Takagi [14]), and recently have been used to model other demand-systems such as production and inventory systems (see, e.g., Federgruen and Katalan [3] or Olsen [10]). For a comprehensive review of queueing analysis of polling models see Takagi [12] [13] [16]. Takagi [12] presents an overview of polling model analysis and applications, an extensive list of references, and various analysis and results. Takagi [13] [16] provide updates on research involving polling models.

A number of papers address finite buffers or routing, but to our knowledge, none address both. Sidi et al. [9] utilize the buffer occupancy method to analyze a polling model with routed customers and infinite buffers. They provide results on the expected number of customers in the system at arbitrary instants and the expected delay for a system with Poisson external arrivals, general cyclic service, and general switchover times. Takagi [15] analyzes a finite capacity polling model with Poisson arrivals, general service, and general switchover times. The Laplace-Stieltjes Transform (LST) of the waiting time distribution is derived using the M/G/1/K queue with vacations results of Lee [7]. Jung and Un [6] provide an analysis of a finite buffer polling system with exhaustive service based on virtual buffering. The buffer occupancy method and the M/G/1/K vacation results of Lee [7] are used to derive the mean waiting time and blocking probabilities.

In this paper, we model a finite buffer polling model with routing using the buffer occupancy method and the concept of virtual buffering. A finite buffer queue implies that the buffer at each queue has limited capacity and that when the queue is full, new arrivals are turned away. Since routing is allowed, upon completion of service, a customer may leave the queueing network or be redirected to another queue in the network. Both routing and finite buffers are realistic modeling elements for the types of communications systems described in Levy and Sidi [8].

Under virtual buffering, which was introduced for polling models by Jung and Un [6], an infinite buffer is virtually present at a queue during vacation periods, i.e., service at other queues and switchover periods). At a polling instant, which is the instant when the server arrives at a new queue in order to serve that queue, the virtual buffer is removed and all excess customers are lost. Thus, the analysis calculates the number of arrivals (external and routed) to the virtual infinite buffer queue during vacation periods, and then calculates the buffer occupancy variables by considering the probability of k arrivals during the vacation period. When the server is serving the queue, the queue behavior, e.g., busy period and number served, is that of an M/G/1/K queue.

The buffer occupancy approach is based on computing moments of the number of customers present at a polling instant. Specifically, the buffer occupancy approach computes the first and second moments of the number of customers present at a polling instant, which are required for deriving the expected queue length at arbitrary instants (as well as the mean delay of the system). The main principle of the buffer occupancy method is to follow the evolution of the system in the forward direction and compute the moments using a set of linear equations as introduced by Cooper and Murray [2] and Cooper [1]. The approach will be discussed in more detail in Section 2.

The remainder of the paper is organized as follows. The model considered is described in Section 2. Section 3 provides the queueing analysis, including the derivation of the generating function for the number of customers present at polling instants, calculation of the buffer occupancy variables, and an expression for system workload. Section 4 provides the derivation of the queue length distribution and expected queue length. Practical and numerical applicability of the model is discussed in Section 5, followed by conclusions and future work discussion in Section 6.

2 Model Description

The polling model considered in this paper is depicted in Figure 1 and consists of a single server and N finite buffer queues. When the buffer is not full, customers arrive according to independent Poisson processes with rate λ_i. Customers that find the buffer full are lost. Customers arriving at queue i are called type i customers and have a service time with LST $S_i^*(\theta)$, mean s_i and second moment $s_i^{(2)}$, $1 \leq i \leq N$.

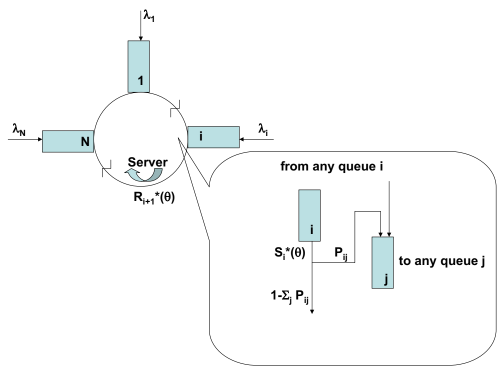

Figure 1: Finite Buffer Polling Model with Routing

The service at queue i follows the exhaustive discipline, in which the server continues to serve until the queue is empty and then proceeds to the next queue. Thus, all customers found in the queue at the beginning of service and all those that arrive and enter the queue during the service period are served in the given service period. After completion of service, customers may be routed to another queue, but not immediately back to the same queue with probability, p_{ij}, $1 \leq i, j \leq N$, or may leave the system. Routed customers that find the queue full are lost. The server then moves to the next queue and incurs a switchover period whose duration is an independent random variable with LST, $R^*_{i+1}(\theta)$, mean r_{i+1} and second moment $r^{(2)}_{i+1}$, $1 \leq i+1 \leq N$. Service order follows a cyclic pattern, meaning that the server serves the queues in order 1 to N, and then returns to queue 1 after serving queue N. The duration of service at a queue is known as a busy period and the time that the server is away from a queue either switching or serving another queue is known as a vacation.

The buffer occupancy method is used to calculate the moments of queue length at polling instants and works as follows. Let X^j_i be the number of customers present at queue j when queue i is polled in steady-state, then the buffer occupancy variables are the set $\{X^j_i; i = 1, 2, ..., N, j = 1, 2, ..., N\}$. The buffer occupancy analysis approach is based on computing moments of these variables, the most important of which are $E[X^j_i]$ and $E[X^j_i X^k_i]$, which are required for deriving the expected queue length at arbitrary instants (as well as the mean delay of the system). The main principle of the buffer occupancy method is to follow the evolution of the system in the forward direction and compute $E[X^j_i]$ and $E[X^j_i X^k_i]$ from $E[X^j_{i-1}]$ and $E[X^j_{i-1} X^k_{i-1}]$ using a set of linear equations. Expected queue lengths can then be calculated from the first moments.

3 Queueing Analysis

3.1 Derivation of $F_{i+1}(z)$

The buffer occupancy method is used to determine the moments of the steady-state queue length, waiting time, and delay. We determine the relationships between the set of buffer occupancy variables by deriving $F_{i+1}(z)$, the generating function for number of customers present at polling instants.

Lemma 1

$$
\begin{aligned}
F_{i+1}(z) \quad = \quad & R^*_{i+1}\left(\sum_{j=1}^{N}\lambda_k(1-z_k)\right)\left\{\sum_{k=0}^{\bar{K}-1} J_i^k\left(P_i(z),\sum_{j\neq i}\lambda_j(1-z_j)\right)\frac{1}{k!}\left(\frac{d^kF_i(z)}{dz_i^k}\right)_{z_i=0}\right. \\
& \left. + J_i^{\bar{K}-1}\left(P_i(z),\sum_{j\neq i}\lambda_j(1-z_j)\right)\left\{F_i(z_1...1...z_N)-\sum_{k=0}^{\bar{K}-1}\frac{1}{k!}\left(\frac{d^kF_i(z)}{dz_i^k}\right)_{z_i=0}\right\}\right\}.
\end{aligned} \tag{1}
$$

Proof

The general balance equations that represent finite buffer queues under virtual buffering are

$$
X_{i+1}^j \quad = \quad \begin{cases} X_i^j + A_i^j + T_i^j + R_{i+1}^j & : \quad i \neq j \\ R_{i+1}^j & : \quad i = j, \end{cases} \tag{2}
$$

where

X_i^j = number of customers at queue j when queue i is polled,

A_i^j = number of arrivals to queue j during service at queue i,

T_i^j = number of customers routed to queue j during service at queue i, and

R_i^j = number of arrivals to queue j during switchover to queue i.

The random variables, A_i^j and T_i^j, represent the number of arrivals to queue j and the number of customers routed to queue j during service at queue i, respectively. Define

$F_i(z) = E[z_1^{X_i^1} z_2^{X_i^2} ....z_N^{X_i^N}]$ as the joint generating function (GF) for the buffer occupancy variables, and $P_i(z)$ as the generating function for the probability of routing a single customer from queue i. Using Equation (2),

$$\begin{aligned} F_{i+1}(z) &= E[z_1^{X_{i+1}^1} ...z_N^{X_{i+1}^N}] \\ &= E[z_1^{X_i^1+A_i^1+T_i^1+R_{i+1}^1} ...z_i^{R_{i+1}^i} ...z_N^{X_i^N+A_i^N+T_i^N+R_{i+1}^N}] \\ &= E[z_1^{X_i^1+A_i^1+T_i^1} ...z_i^0 ...z_N^{X_i^N+A_i^N+T_i^N}]E[z_1^{R_{i+1}^1} ...z_N^{R_{i+1}^N}], \end{aligned}$$

since arrivals during switchover are independent of service periods.

Since the arrivals during switchover are from a Poisson process, by conditioning on the length of the switchover period, R_{i+1}, the joint generating function for arrivals during switchover becomes

$$\begin{aligned} E[z_1^{R_{i+1}^1} ...z_N^{R_{i+1}^N}] &= E[E[z_1^{R_{i+1}^1} ...z_N^{R_{i+1}^N} | R_{i+1}]] \\ &= E[e^{\lambda_1(z_1-1)R_{i+1}} ...e^{\lambda_N(z_N-1)R_{i+1}}] \\ &= E[e^{-\left(\sum_k \lambda_k(1-z_k)R_{i+1}\right)}] \\ &= R_{i+1}^* \left(\sum_k \lambda_k(1-z_k)\right). \end{aligned} \tag{3}$$

Recall that there is a virtual infinite buffer at queue j during the service of queue i; therefore excess arrivals (external or routed) will be removed at the next polling instant at queue j.

We now derive $E[z_1^{X_i^1+A_i^1+T_i^1} ...z_i^0 ...z_N^{X_i^N+A_i^N+T_i^N}]$. By conditioning on X_i^i, we have that

$$\begin{aligned} E[z_1^{X_i^1+A_i^1+T_i^1} ...z_i^0 ...z_N^{X_i^N+A_i^N+T_i^N}] &= E[E[z_1^{X_i^1+A_i^1+T_i^1} ...z_i^0 ...z_N^{X_i^N+A_i^N+T_i^N} | X_i^i]] \\ &= E[E[z_1^{X_i^1} ...z_i^0 ...z_N^{X_i^N} | X_i^i] E[z_1^{A_i^1+T_i^1} ...z_i^0 ...z_N^{A_i^N+T_i^N} | X_i^i]], \end{aligned}$$

where the given independence again follows from the Poisson arrival processes. Now,

$$\begin{aligned} F_i(z) &= E[z_1^{X_i^1} ...z_N^{X_i^N}] \\ &= \sum_{n_1=0}^{\infty} ... \sum_{n_N=0}^{\infty} z_1^{n_1} ...z_N^{n_N} P(X_i^i = n_1, ..., X_i^N = n_N) \\ &= \sum_{n_1=0}^{\infty} ... \sum_{n_N=0}^{\infty} z_1^{n_1} ...z_N^{n_N} P(X_i^i = n_1, ..., X_i^N = n_N | X_i^i = n_i) P(X_i^i = n_i). \end{aligned}$$

Since the only nonzero term is when $n_i = k$, taking the k^{th} derivative with respect to z_i yields,

$$\frac{d^k F_i(z)}{dz_i^k} = \sum_{n_1=0}^{\infty} .. \sum_{n_i=k}^{\infty} .. \sum_{n_N=0}^{\infty} z_1^{n_1}..1..z_N^{n_N} \frac{n_i!}{(n_i-k)!} z_i^{n_i-k} P(X_i^i = n_1, ..., X_i^N = n_N | X_i^i = n_i) P(X_i^i = n_i).$$

Setting $z_i = 0$ yields

$$\begin{aligned}
\left(\frac{d^k F_i(z)}{dz_i^k}\right)_{z_i=0} &= \sum_{n_1=0}^{\infty} ... \sum_{n_N=0}^{\infty} z_1^{n_1}...1...z_N^{n_N} k! P(X_i^i = n_1, ..., X_i^N = n_N | X_i^i = k) P(X_i^i = k) \\
&= k! P(X_i^i = k) \sum_{n_1=0}^{\infty} ... \sum_{n_N=0}^{\infty} z_1^{n_1}...1...z_N^{n_N} P(X_i^i = n_1, ..., X_i^N = n_N | X_i^i = k) \\
&= k! P(X_i^i = k) E[z_1^{X_i^1}...1...z_N^{X_i^N} | X_i^i = k].
\end{aligned}$$

Rearranging terms we have

$$P(X_i^i = k) E[z_1^{X_i^1}...1...z_N^{X_i^N} | X_i^i = k] = \frac{1}{k!} \left(\frac{d^k F_i(z)}{dz_i^k}\right)_{z_i=0}. \tag{4}$$

It is shown in Grasman [4] that

$$E[z_1^{A_i^1+T_i^1}....z_i^0....z_N^{A_i^N+T_i^N} | X_i^i] = J_i^{X_i^i} \left(P_i(z), \sum_{j \neq i} \lambda_j (1 - z_j)\right), \tag{5}$$

where the joint generating function and Laplace-Stieltjes Transform (GF/LST) for the number served in and the length of a busy period in a finite buffer queue starting with k customers is given by

$$J_i^k(z, \theta) = z \sum_{j < \bar{K}_i - k_i} \left(J^{k_i-1+j}(z, \theta) - J^{\bar{K}_i-1}(z, \theta)\right) \frac{(-\lambda)^j}{j!} S_i^{*(j)}(\theta + \lambda) + z J^{\bar{K}_i-1}(z, \theta) S_i^*(\theta), \tag{6}$$

as shown in the Appendix. Equations (4) and (6), explicitly consider the fact that the service at queue i mimics the behavior of an M/G/1/K queue; thus the number of arrivals and the number routed to queue j during the service at queue i are a function of the length of the busy period and the number served during the busy period at the finite buffer queue. Now using Equations (4)and (5), we complete the derivation of the generation function for the number of customers present at a polling instant in an infinite buffer polling model with routing.

$$E[z_1^{X_i^1+A_i^1+T_i^1}...z_i^0...z_N^{X_i^N+A_i^N+T_i^N}] = \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{d^k F_i(z)}{dz_i^k} \right)_{z_i=0} J_i^k \left(P_i(z), \sum_{j \neq i} \lambda_j (1-z_j) \right).$$

Splitting the sum yields

$$\begin{aligned} & E[z_1^{X_i^1+A_i^1+T_i^1}...z_i^0...z_N^{X_i^N+A_i^N+T_i^N}] \\ = & \sum_{k=0}^{\bar{K}-1} J_i^k \left(P_i(z), \sum_{j \neq i} \lambda_j (1-z_j) \right) \frac{1}{k!} \left(\frac{d^k F_i(z)}{dz_i^k} \right)_{z_i=0} \\ & \quad + J_i^{\bar{K}-1} \left(P_i(z), \sum_{j \neq i} \lambda_j (1-z_j) \right) \sum_{k=\bar{K}}^{\infty} \frac{1}{k!} \left(\frac{d^k F_i(z)}{dz_i^k} \right)_{z_i=0}. \end{aligned}$$

In order to eliminate the infinite sum note that

$$\begin{aligned} \sum_{k=\bar{K}}^{\infty} \frac{1}{k!} \left(\frac{d^k F_i(z)}{dz_i^k} \right)_{z_i=0} &= \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{d^k F_i(z)}{dz_i^k} \right)_{z_i=0} - \sum_{k=0}^{\bar{K}-1} \frac{1}{k!} \left(\frac{d^k F_i(z)}{dz_i^k} \right)_{z_i=0} \\ &= \sum_{k=0}^{\infty} P(X_i^i = k) E[z_i^{X_i^1}...1...z_N^{X_n^1} | X_i^i = k] - \sum_{k=0}^{\bar{K}-1} \frac{1}{k!} \left(\frac{d^k F_i(z)}{dz_i^k} \right)_{z_i=0} \\ &= E[z_i^{X_i^1}...1...z_N^{X_n^1}] - \sum_{k=0}^{\bar{K}-1} \frac{1}{k!} \left(\frac{d^k F_i(z)}{dz_i^k} \right)_{z_i=0} \\ &= F_i(z_1...1...z_N) - \sum_{k=0}^{\bar{K}-1} \frac{1}{k!} \left(\frac{d^k F_i(z)}{dz_i^k} \right)_{z_i=0}. \end{aligned}$$

Finally, using this result and replacing the switchover term, we have that

$$\begin{aligned} F_{i+1}(z) = & R_{i+1}^* \left(\sum_{j=1}^{N} \lambda_k (1-z_k) \right) \left\{ \sum_{k=0}^{\bar{K}-1} J_i^k \left(P_i(z), \sum_{j \neq i} \lambda_j (1-z_j) \right) \frac{1}{k!} \left(\frac{d^k F_i(z)}{dz_i^k} \right)_{z_i=0} \right. \\ & \left. + J_i^{\bar{K}-1} \left(P_i(z), \sum_{j \neq i} \lambda_j (1-z_j) \right) \left\{ F_i(z_1...1...z_N) - \sum_{k=0}^{\bar{K}-1} \frac{1}{k!} \left(\frac{d^k F_i(z)}{dz_i^k} \right)_{z_i=0} \right\} \right\}. \end{aligned}$$

□

Calculation of Buffer Occupancy Variables

Since Equation (1) is quite complex, we will introduce some notation, which is easier to work with and will be used for calculation of the buffer occupancy variables. As in Jung and Un [6], we define

$$
\begin{aligned}
G_i(z_1,...,z_N;k_1,...,k_N) &= \frac{1}{k_1!k_2!...k_N!}\frac{\partial^{k_1+...+k_N}F_i(z_1,...,z_N)}{\partial_{z_1}^{k_1}...\partial_{z_N}^{k_N}} \\
&\text{and} \\
G_i(Y_1,...,Y_N;k_1,...,k_N) &= G_i(z_1,...,z_N;k_1,...,k_N)|_{z=Y}.
\end{aligned} \tag{7}
$$

Also let $Y_i = 0$ or 1, then define

$$
U_{i,k_i}(Y_i;z_1,...,z_N;k_1,...,k_N) = \frac{1}{k_1!k_2!...k_N!}\frac{\partial^{k_1+...+k_N}}{\partial_{z_1}^{k_1}...\partial_{z_N}^{k_N}}\left\{R^*_{i+1}\left(\sum_{j=1}^{N}\lambda_j(1-z_j)\right)\right.
$$
$$
\left.\left((1-Y_i)J_i^k\left(P_i(z),\sum_{j\neq i}\lambda_j(1-z_j)\right)+(2Y_i-1)J_i^{\bar{K}}\left(P_i(z),\sum_{j\neq i}\lambda_j(1-z_j)\right)\right)\right\}. \tag{8}
$$

From Equation (7), we see that $F_{i+1}(z) = G_{i+1}(z_1,...,z_N;0,...,0)$, and we can now derive a relationship between the joint generating functions at successive polling instants, which is similar to Equation (7) from Jung and Un, [6], but also considers routing. It is shown in Grasman [4] that

$$
F_{i+1}(z) = \sum_{Y_i=0}^{1}\sum_{k_i=0}^{(1-Y_i)(\bar{K}_i-1)} G_i(z_1,...,z_N;k_1,...,k_N)|_{z_i=Y_i}U_{i,k_i}(Y_i;z_1,...,z_N;0,...,0). \tag{9}
$$

Equation (9) is used to recursively solve for the buffer occupancy variables at the individual queues in a similar manner to that found in the appendix of Jung and Un [6].

Special Case: Infinite Buffers

It is shown in Grasman [4] that the right-hand sides of Equations (1) and (6) converge to the results for the infinite buffer case as the buffer size increases to infinity.

Proposition 1 *As* $\bar{K}_i \to \infty$*,*

$$
\begin{aligned}
F_{i+1}(z) &= \sum_{k=0}^{\infty} J_i^k\left(P_i(z),\sum_{j\neq i}\lambda_j(1-z_j)\right)E[z_1^{X_i^1}...1...z_N^{X_i^N}|X_i^i=k]P(X_i^i=k) \\
&= F_i\left(z_1....G_i^*\left(P_i(z),\sum_{j\neq i}\lambda_j(1-z_j)\right)....z_N\right)\forall i,
\end{aligned}
$$

which corresponds to Equation (3.15) of Grasman [4], and $J_i^k(z,\theta) \to G_i^*(z,\theta)$ for all i, where $G_i^*(z,\theta) = zS_i^*(\theta+\lambda_i-\lambda_i G_i^*(z,\theta))$ is the joint GF/LST of the number served in and the busy period of an M/G/1/K queue starting with k customers. The system with infinite buffers may be used to represent production models with backlogging as shown in Grasman [4] and Grasman et al. [5].

3.2 Workload

Define workload in a queue as the time it would require the server working exclusively on that queue to clear all current customers. For service periods, let $M^k =$ the cumulative workload in the queue during a busy period initiated with k customers, (i.e., workload summed over all customers and all service periods during a busy period). Also, let $M_{S,N_S}^k = E[M^k|S,N_S]$, where S is the length of the first service period, and N_S is the number of arrivals during the service period S. Then, for an arbitrary queue with buffer size, $\bar{K}$,

$$M_{S,N_S}^k = \begin{cases} kS + N_S S/2 + E[M^{k+N_S-1}|N_S] & : \quad N_S \le \bar{K}-k \\ kS + E\left[\sum_{i=1}^{\bar{K}-k}(S-U_{(i)})|S,N_S\right] + E[M^{\bar{K}-1}] & : \quad N_S > \bar{K}-k, \end{cases} \tag{10}$$

where $U_{(i)}$ is the time the i^{th} customer arrives during service period S, (i.e., $U_{(i)}$ is an order statistic). All customers spend the entire service time in the queue and thus contribute kS to the cumulative workload during S. Since the N_S arrivals are uniformly distributed as long as $N_S \le \bar{K}-k$ the workload is increased, with the expected time of arrival $S/2$. If $N_S > \bar{K}-k$, then we must only consider the first $\bar{K}-k$ arrivals which are in the system for S minus their actual arrival time.

First, we simplify the expression for M_{S,N_S}^k by noting that

$$E\left[\sum_{i=1}^{\bar{K}-k} U_{(i)}|N_S,S\right] = \sum_{i=1}^{\bar{K}-k} E\left[\int_{u=0}^{S} u f_i^{N_S,S}(u)du|N_S,S\right],$$

where $f_i^{N_S,S}(u)$ is the marginal distribution of $U_{(i)}$ given N_S and S. It can be shown (see, e.g., Ross [11]) that for Poisson arrivals,

$$f^{N_S,S}(u) = \frac{N_S!}{(i-1)!(N_S-i)!}\left(\frac{u}{S}\right)^{i-1}\left(1-\frac{u}{S}\right)^{N_S-i}\frac{1}{S},$$

thus

$$\begin{aligned}
E\left[\sum_{i=1}^{\bar{K}-k} U_{(i)}|N_S, S\right] &= \sum_{i=1}^{\bar{K}-k} E\left[\int_{u=0}^{S} u\frac{N_S!}{(i-1)!(N_S-i)!}\left(\frac{u}{S}\right)^{i-1}\left(1-\frac{u}{S}\right)^{N_S-i}\frac{1}{S}du|N_S, S\right] \\
&= \sum_{i=1}^{\bar{K}-k} \frac{N_S!}{(i-1)!(N_S-i)!}\int_{u=0}^{S}\left(\frac{u}{S}\right)^{i}\left(1-\frac{u}{S}\right)^{N_S-i}du.
\end{aligned}$$

We perform a variable change in order to get the integral into the form of a beta integral. Recall that

$$\begin{aligned}
\beta(z_1, z_2) &= \int_0^1 t^{z_1-1}(1-t)^{z_2-1}dt \\
&= \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+z_2)},
\end{aligned}$$

where $\Gamma(z) = (z-1)!$ if z is an integer. For any given random event ω, let $t = \frac{u}{S(\omega)}$ and $du = S(\omega)dt$. Then,

$$\begin{aligned}
\int_{u=0}^{S(\omega)}\left(\frac{u}{S(\omega)}\right)^{i}\left(1-\frac{u}{S(\omega)}\right)^{N_S(\omega)-1}du &= S(\omega)\int_{t=0}^{1} t^{i}(1-t)^{N_S(\omega)-1}dt \\
&= S(\omega)\beta(1+i, N_S(\omega)-i+1).
\end{aligned}$$

Therefore,

$$\begin{aligned}
E\left[\sum_{i=1}^{\bar{K}-k} U_{(i)}|N_S, S\right] &= \sum_{i=1}^{\bar{K}-k}\frac{N_S!}{(i-1)!(N_S-i)!}S\beta(i+1, N_S-i+1) \\
&= \sum_{i=1}^{\bar{K}-k}\frac{N_S!}{(i-1)!(N_S-i)!}S\frac{i!(N_S-i)!}{(N_S+1)!} \\
&= \sum_{i=1}^{\bar{K}-k}\frac{Si}{N_S+1} \\
&= S\frac{(\bar{K}-k)(\bar{K}-k+1)}{2(N_S+1)}.
\end{aligned}$$

Substituting into the expression for M^k_{S,N_S}, summing over values of N_S, and taking the expectation with respect to S, yields

$$\begin{aligned}
E[M^k] &= E\left\{\sum_{n=0}^{\bar{K}-k} P(N_S=n|S)\left[S\left((k+\frac{n}{2}\right)+M^{k+n-1}\right]\right. \\
&+ \left.\sum_{n>\bar{K}-k} P(N_S=n|S)\left[kS+\sum_{i=1}^{\bar{K}-k} S - S\left(\frac{(\bar{K}-k)(\bar{K}-k+1)}{2(n+1)}\right)+M^{\bar{K}-1}\right]\right\}
\end{aligned}$$

Rewriting the infinite sum,

$$\begin{aligned} E[M^k] &= E\Bigg\{\sum_{n=0}^{\bar{K}-k} \frac{(\lambda S)^n e^{-\lambda S}}{n!} \left[S(k+\frac{n}{2}) + M^{k+n-1}\right] \\ &+ \sum_{n=0}^{\infty} \frac{(\lambda S)^n e^{-\lambda S}}{n!} \left[kS + \sum_{i=1}^{\bar{K}-k} S - S\left(\frac{(\bar{K}-k)(\bar{K}-k+1)}{2(n+1)}\right) + M^{\bar{K}-1}\right] \\ &- \sum_{n=0}^{\bar{K}-k} \frac{(\lambda S)^n e^{-\lambda S}}{n!} \left[kS + \sum_{i=1}^{\bar{K}-k} S - S\left(\frac{(\bar{K}-k)(\bar{K}-k+1)}{2(n+1)}\right) + M^{\bar{K}-1}\right]\Bigg\}. \end{aligned}$$

The infinite sum reduces to

$$E\left[(1-e^{-\lambda S})\left(\frac{(\bar{K}-k)(\bar{K}-k+1)}{2\lambda} + M^{\bar{K}-1}\right)\right];$$

thus,

$$\begin{aligned} E[M^k] &= E\Bigg\{\sum_{n=0}^{\bar{K}-k} \frac{(\lambda S)^n e^{-\lambda S}}{n!} \Bigg(\left[S(k+\frac{n}{2}) + M^{k+n-1}\right] \\ &- \left[kS + \sum_{i=1}^{\bar{K}-k} S - S\left(\frac{(\bar{K}-k)(\bar{K}-k+1)}{2(n+1)}\right) + M^{\bar{K}-1}\right]\Bigg) \\ &+ (1-e^{-\lambda S})\left(kS + \sum_{i=1}^{\bar{K}-k} S - \frac{(\bar{K}-k)(\bar{K}-k+1)}{2\lambda} + M^{\bar{K}-1}\right)\Bigg\}. \end{aligned}$$

Finally, the expected total workload during a busy period is given by

$$E[M_{serve}] = \sum_{k=0}^{\bar{K}} E[M^k] P(X^p = k),$$

where X^p is the random number of customers in the queue at the polling instant and is independent of future service times and arrivals.

Similarly, for vacation periods, let $\mathcal{M}^k$ = the cumulative workload in the queue during a vacation period initiated with k customers. Also, let $\mathcal{M}^k_{V,N_V} = E[\mathcal{M}^k|V, N_V]$, where V is the length of a vacation period, and N_V is the number of arrivals during the vacation period, V. Then, for an arbitrary queue with buffer size, $\bar{K}$,

$$\mathcal{M}^k_{V,N_V} = \begin{cases} kV + N_V V/2 & : \quad N_V \le \bar{K}-k \\ kV + E\left[\sum_{i=1}^{\bar{K}-k}(V - U_{(i)})|V, N_V\right] & : \quad N_V > \bar{K}-k, \end{cases} \tag{11}$$

where $U_{(i)}$ is the i^{th} customer to arrive during vacation (order statistic). Repeating the previous analysis yields

$$\begin{aligned} E[\mathcal{M}^k] &= E\left\{\sum_{n=0}^{\bar{K}-k} \frac{(\lambda V)^n e^{-\lambda V}}{n!} \left(\left[V(k+\frac{n}{2})\right]\right.\right. \\ &- \left.\left.\left[kV + \sum_{i=1}^{\bar{K}-k} V - V\left(\frac{(\bar{K}-k)(\bar{K}-k+1)}{2(n+1)}\right)\right]\right)\right. \\ &+ \left.(1-e^{-\lambda V})\left(kV + \sum_{i=1}^{\bar{K}-k} V - \frac{(\bar{K}-k)(\bar{K}-k+1)}{2\lambda}\right)\right\}. \end{aligned}$$

Finally, the expected total workload during a vacation period is given by

$$E[M_{vac}] = \sum_{k=0}^{\bar{K}} E[\mathcal{M}^k] P(X^v = k),$$

where X^v is the random number of customers in the queue at the start of a vacation, which is zero in this model due to the exhaustive service discipline. The total workload during a vacation period and service period are then added together to obtain the total workload for a cycle.

4 Derivation of Queue Length

4.1 Queue Length Distribution

Using an embedded Markov chain analysis similar to that of Lee [7], we examine the queue length, Q, at the polling instant and at the service completion of each individual customer. Without loss of generality, we may choose any queue, and thus we drop the queue subscripts in this section. Let $\gamma = 1$ for a service completion instant and $\gamma = 0$ for a polling instant. Then given either a polling instant or service completion, define $p_n = P(Q = n, \gamma = 1)$ and $q_n = P(Q = n, \gamma = 0)$. Since p_n is the probability of having n customers in the system at a service completion, $p_{\bar{K}} = 0$ because the customer being served has left the system. Then, as in Lee [7], the limiting probability distributions satisfy

$$p_n = \sum_{k=1}^{n+1}(p_k+q_k)g_{n-k+1}, \qquad n=0,1,...,\bar{K}-2; \tag{12}$$

$$p_{\bar{K}-1} = \sum_{k=1}^{\bar{K}-1}(p_k+q_k)g^c_{\bar{K}-k}+q_{\bar{K}}; \tag{13}$$

$$q_n = (p_0+q_0)h_n, \qquad n=0,1,...,\bar{K}-1; \tag{14}$$

$$q_{\bar{K}} = (p_0+q_0)h^c_{\bar{K}}; \tag{15}$$

where g_j and h_j are the probabilities that j customers arrive during a service time with LST, $S^*(\theta)$, and a vacation time with LST, $V^*(\theta)$, respectively, $g^c_j = \sum_{k=j}^{\infty} g_k$, and $h^c_j = \sum_{k=j}^{\infty} h_k$. These relationships hold for the system with routing since the queue length distribution at the polling instant and service completion instants only depends on the length of the current vacation and individual service periods, respectively, and not on previous vacation or service periods; however, the distribution of queue length differs from that of Lee [7], thus his result may not be directly utilized.

The generating function for the number of arrivals during service can easily be shown to be $S^*(\lambda(1-z))$ since the arrivals during service are distributed according to the Poisson distribution; however, contrary to Lee [7], the arrivals during a vacation period are not Poisson due to routing. Therefore, we must define $h^{ext}_j = P(j$ Poisson (external) arrivals during a vacation) and $h^{int}_j = P(j$ routed (internal) arrivals during a vacation). It can be easily shown that the generating function for the number of external arrivals during a vacation period is $V^*(\lambda(1-z))$ since the external arrivals during service are distributed according to a Poisson distribution. The distribution of h^{int}_j is found using the buffer occupancy variables, which were derived in Section 3.1 since the number of routed customers may be determined using the number served at other queues during the vacation period. The number of internal arrivals depends on the number of arrivals during the previous cycle, but is independent of the number of arrivals in this cycle, thus, $h_j = \sum_{k=0}^{j} h^{ext}_k h^{int}_{j-k}$.

We recursively solve for the limiting probabilities, by defining $\beta_n = \frac{p_n+q_n}{p_0+q_0}$ and noting that

$\beta_0 = 1$. From Equations (13) and (14), for $0 \leq n \leq N-2$,

$$\beta_n = \frac{\sum_{k=1}^{n+1}(p_k + q_k)g_{n-k+1} + (p_0+q_0)h_n}{p_0+q_0} = \sum_{k=1}^{n+1} g_{n-k+1}\beta_k + h_n\beta_0. \tag{16}$$

Since the limiting probabilities sum to 1, i.e., $1 = \sum_{j=0}^{\bar{K}-1}(p_j+q_j) + q_{\bar{K}}$,

$$1 = \sum_{j=0}^{\bar{K}-1}(p_j+q_j) + (p_0+q_0)h^c_{\bar{K}}$$

$$\Rightarrow \frac{1}{p_0+q_0} = \frac{\sum_{j=0}^{\bar{K}-1}(p_j+q_j) + (p_0+q_0)h^c_{\bar{K}}}{p_0+q_0}$$

$$\Rightarrow \frac{1}{p_0+q_0} = \sum_{j=0}^{\bar{K}-1}\beta_j + h^c_{\bar{K}}.$$

Finally,

$$p_0 + q_0 = \frac{1}{\sum_{j=0}^{\bar{K}-1}\beta_j + h^c_{\bar{K}}}. \tag{17}$$

The values for q_n may then be found from Equations (14) and p_n may be found from the definition of β_n.

4.2 Expected Queue Length

The LST for waiting time distribution, $W^*(\theta)$, given by Lee [7], holds as long as the appropriate distributions are used for service time and vacation time. Thus, for any queue,

$$W^*(\theta) = \frac{\rho(1-\rho')S^*(\theta)(1-(\frac{\lambda}{\lambda-\theta})^{\bar{K}}S^*(\theta)^{\bar{K}})(V^*(\theta)-1)}{E[V]\rho'(\lambda-\theta-\lambda S^*(\theta))} + \frac{E[S](1-\rho')+E[V]\rho'}{E[V]\rho'}S^*(\theta)^{\bar{K}}\sum_{j=0}^{\bar{K}-1}p_j\left(\frac{\lambda}{\lambda-\theta}\right)^{\bar{K}-j}, \tag{18}$$

where $\rho' = P(\gamma = 1)$. Little's Law and Equation (18) are used to derive the expected value of queue length as

$$E[Q] = \frac{\rho}{\rho'}\left(\frac{\sigma}{\lambda^2}\sum_{j=1}^{\bar{K}-1}jp_j + \frac{\bar{K}}{\lambda}\left(1-\frac{\rho'}{\rho}\right)\right), \tag{19}$$

where $\sigma = (E[S](1-\rho') + E[V]\rho')(E[V]E[S])$.

5 Applicability of Model

5.1 Practical Applicability

The system we consider may be applied to production systems with multiple processing steps or rework. After completion of service at one queue, an item may be routed to another queue for the next processing step or rework. An applicable model for the finite buffer is to represent a basestock inventory system with lost sales or expediting. When the queue is empty, the inventory level is equal to the basestock level, thus the server will continue to serve the queue until it is empty. If the capacity of the queue is equal to the inventory allocated to the item, then when the queue is full, inventory is fully depleted and new demand is lost. Figure 2 provides an illustration.

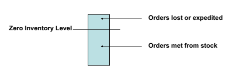

Figure 2: Queueing Representation of Lost Sales or Expediting

Consider a multiproduct system with lost sales, setups, and random yield described in the introduction. The system may be represented, as in Figure 3, as consisting of a single server and N queues representing $N/2$ items. The odd numbered queues correspond to each of the items, while the even numbered queues are used for temporary storage of defects. Since the yield of each item is assumed to be independent of all others, each item either exits the system or is routed to a temporary storage queue according to the Bernoulli distribution. Items are routed from queue i to queue j $(j = i + 1)$ with probability p_{ij} due to defective production; items that find the queue full are lost.

After the inventory position of an item reaches its basestock level (minus any defective production), the server then moves to the storage queue and routes the customers (defects) back to their original queue. The service at any queue follows a general distribution, and after completion of service at both the real and storage queues, the server moves in a cyclic fashion, incurs a switchover period whose duration corresponds to the setup of the next item,

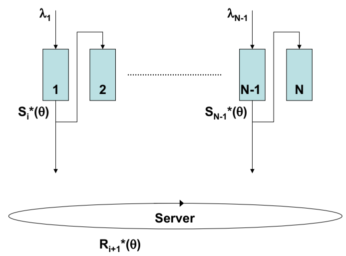

Figure 3: Polling Model Representation of Random Yield Production System

and follows a general distribution. Backorders are not allowed so the queue is given a finite buffer equal to the basestock level; when the queue is full, new demand is lost or expedited. Grasman [4] and Grasman et al. [5] provide numerical results for the above model. Other applications include shop floor scheduling and cellular manufacturing.

5.2 Numerical Applicability

The numerical procedure presented in this paper for analysis of a finite buffer polling model with routing requires the calculation of the entire set of state probabilities for each queue at a polling instance, and requires the solution of a system of linear equations representing the unknown state probabilities. The solution of this system of linear equations is the limiting factor in the numerical applicability, thus we present a brief numerical analysis in this section.

Specifically, for a system with N queues with capacity, K_i for $i = 1, 2, ..., N$, the number of linear equations/unknowns is:

$$\prod_{i=1}^{N}(K_i + 1) - 1. \tag{20}$$

As can be seen from Equation (20), the system size grows quite rapidly as N increases. Symmetric queues, i.e., $K_i = K \ \forall \ i$, represent the maximum size of the system of linear equations for a given number of queues and total buffer capacity, thus we use a symmetric example to present numerical tractability. For symmetric queues, Equation (20) reduces to:

$$(K+1)^N - 1. \tag{21}$$

Figures 4 and 5, show the tractable buffer size as a function of the number of queues and the tractable number of queues as a function of buffer size based on the size of the system of linear equations/unknowns. From Figure 4, it can be seen that the numerical procedure is tractable for a small number of queues with large buffer sizes, a moderate number of queues with moderate buffer sizes, and for large number of queues with buffer size of 1.

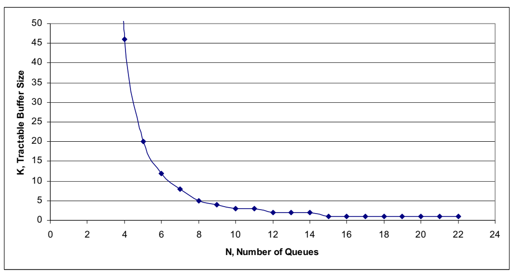

Figure 4: Tractable (Symmetric) Buffer Size as a Function of Number of Queues

Similarly, from Figure 5, it can be seen that the numerical procedure is tractable for small buffer sizes and a large number of queues, moderate buffer sizes and a moderate number of queues, and for large buffer sizes and a small number of queues.

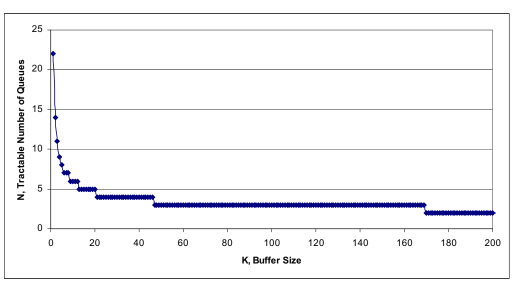

Figure 5: Tractable Number of Queues as a Function of (Symmetric) Buffer Size

In order to calculate the unknown coefficients, the generating function for number of customers present at polling instants, Equation (1), could be differentiated $\sum_{i=1}^{N}(K_i - 1)$ times (or $N(K-1)$ times for symmetric queues). For systems with a tractable number of linear equations/unknowns as presented above, the required number of derivatives is not a limiting factor of the numerical procedure, however, the expressions become quite complex. These derivatives are avoided by using the procedure presented in Section 3.1. In addition, the numerical procedure requires calculation of the queue length distribution (and expected queue length) from the waiting time distribution as shown is Section 4. Again, these calculations are not limiting factors. Finally, the stability of the numerical procedure is limited only by the numerical accuracy (rounding) of the solutions.

6 Summary and Future Work

This paper presented the analysis of a finite buffer polling system with routing. The most significant results of the analysis are the derivation of the generating function for queue length when buffer sizes are limited and a representation of the system workload. This work may be applied to a multiproduct production system with lost sales, setups, and random yield by representing the system as a standard polling model and controlling the inventory of each product using a basestock policy. This work may be further extended by developing cost functions using the system workload, which may be used to determine optimal basestock levels for the system (see Grasman et al. [5]).

Acknowledgments

This research was supported in part by National Science Foundation Grant DMI-9713727.

Appendix

Here we derive $J^k(z,\theta)$, the joint GF/LST of the number served in and the busy period of an M/G/1/K queue starting with k customers and show the derivation for an arbitrary queue (thus leaving off the queue subscript). Let N^k and B^k be the number served in and length of an M/G/1/K queue starting with k customers, then

$$J^k(z,\theta) = \begin{cases} E[z^{N^k} e^{-\theta B^k}] & : \quad k \leq \bar{K} \\ E[z^{N^{\bar{K}}} e^{-\theta B^{\bar{K}}}] & : \quad k \geq \bar{K}, \end{cases} \tag{22}$$

where $\bar{K}$ is the buffer size. Assume, without loss of generality, that $k \leq \bar{K}$ for the remaining analysis. First, conditioning on Y, the number of arrivals during service S, of an arbitrary customer,

$$J^k(z,\theta) = E[E[z^{N^k} e^{-\theta B^k} | Y]].$$

Given $Y = y \leq \bar{K} - k$, the number of customers in the queue after the k^{th} customer's service is $k-1$ plus the number of arrivals that enter during service. From this, we can develop the recursive relationships $N^k \overset{D}{=} 1 + N^{k-1+y}$ and $B^k \overset{D}{=} S + B^{k-1+y}$, where $\overset{D}{=}$ represents equal in distribution. Using these relationships, we write the expectation as

$$\begin{aligned} J^k(z,\theta) = & \sum_{j<\bar{K}-k} E[z^{1+N^{k-1+j}} e^{-\theta(S+B^{k-1+j})} | Y = j] P(Y = j) \\ & + \sum_{j \geq \bar{K}-k} E[z^{1+N^{\bar{K}-1}} e^{-\theta(S+B^{\bar{K}-1})} | Y = j] P(Y = j). \end{aligned}$$

Collecting like terms, and using the conditional independence of (N^{k-1+y}, B^{k-1+y}) and S,

$$\begin{aligned} J^k(z,\theta) = & \ z \sum_{j<\bar{K}-k} J^{k-1+j}(z,\theta) E[e^{-\theta S} | Y = j] P(Y = j) \\ & + z J^{\bar{K}-1}(z,\theta) \sum_{j \geq \bar{K}-k} E[e^{-\theta S} | Y = j] P(Y = j). \end{aligned}$$

We now derive $E[e^{-\theta S}|Y=j]P(Y=j)$ using an indicator variable.

$$
\begin{aligned}
E[e^{-\theta S}|Y=j]P(Y=j) &= E[e^{-\theta S}I[Y=j]] \\
&= E[E[e^{-\theta S}I[Y=j]]|S]] \\
&= E[e^{-\theta S}E[I[Y=j]]|S]] \\
&= E[e^{-\theta S}P(Y=j|S)] \\
&= E[e^{-\theta S}e^{-\lambda S}\frac{(\lambda S)^j}{j!}] \qquad (23) \\
&= E[e^{-S(\theta+\lambda)}(-S)^j\frac{(-\lambda)^j}{j!}] \\
&= \frac{(-\lambda)^j}{j!}S^{*(j)}(\theta+\lambda), \qquad (24) \\
\text{where } S^{*(j)}(\theta+\lambda) &= E[e^{-S(\theta+\lambda)}(-S)^j] = \frac{\partial^j}{\partial\theta^j}E[e^{-S(\theta+\lambda)}].
\end{aligned}
$$

Continuing using Equation (23),

$$
\begin{aligned}
J^k(z,\theta) &= z\sum_{j<\bar{K}-k} J^{k-1+j}(z,\theta)E[e^{-(\theta+\lambda)S}\frac{(\lambda S)^j}{j!}] \\
&\quad +zJ^{\bar{K}-1}(z,\theta)\sum_{j\geq\bar{K}-k} E[e^{-(\theta+\lambda)S}\frac{(\lambda S)^j}{j!}] \\
&= z\sum_{j<\bar{K}-k} J^{k-1+j}(z,\theta)E[e^{-(\theta+\lambda)S}\frac{(\lambda S)^j}{j!}] \\
&\quad +zJ^{\bar{K}-1}(z,\theta)E[\sum_{j\geq\bar{K}-k} e^{-(\theta+\lambda)S}\frac{(\lambda S)^j}{j!}].
\end{aligned}
$$

where the sum may be moved inside the expectation as all terms are nonnegative. Since $\sum_{j=0}^{\infty} e^{-\lambda S}\frac{(\lambda S)^j}{j!} = 1$, we can rewrite the expression as

$$
\begin{aligned}
J^k(z,\theta) &= z\sum_{j<\bar{K}-k} J^{k-1+j}(z,\theta)E\left[e^{-(\theta+\lambda)S}\frac{(\lambda S)^j}{j!}\right] \\
&\quad +zJ^{\bar{K}-1}(z,\theta)E\left[\left(1-\sum_{j<\bar{K}-k} e^{-\lambda S}\frac{(\lambda S)^j}{j!}\right)e^{-\theta S}\right].
\end{aligned}
$$

Multiplying out the terms, distributing the expected value, and pulling the sum back out of the expectation, we have

$$\begin{aligned}J^k(z,\theta) &= z\sum_{j<\bar{K}-k} J^{k-1+j}(z,\theta)E\left[e^{-(\theta+\lambda)S}\frac{(\lambda S)^j}{j!}\right] \\ &\qquad +zJ^{\bar{K}-1}(z,\theta)E[e^{-\theta S}] - zJ^{\bar{K}-1}(z,\theta)\sum_{j<\bar{K}-k} E\left[e^{-(\theta+\lambda)S}\frac{(\lambda S)^j}{j!}\right].\end{aligned}$$

Finally, collecting terms and writing in terms of Equation (24),

$$J^k(z,\theta) = z\sum_{j<\bar{K}-k}\left(J^{k-1+j}(z,\theta) - J^{\bar{K}-1}(z,\theta)\right)\frac{(-\lambda)^j}{j!}S^{*(j)}(\theta+\lambda) + zJ^{\bar{K}-1}(z,\theta)S^*(\theta). \tag{25}$$

References

[1] Cooper, R.B., "Queues Served in Cyclic Order: Waiting Times," *The Bell System Technical Journal* (49)3, 399-413, 1970.

[2] Cooper, R.B., and G. Murray, "Queues Served in Cyclic Order," *The Bell System Technical Journal* (48)3, 675-689, 1969.

[3] Federgruen, A., and Z. Katalan, "Determining Production Schedules Under Base-stock Policies in Single Facility Multi-item Production Systems," *Operations Research*(46)6, 883-898, 1998.

[4] Grasman, S.E., "Production Strategies for Random Yield Processes," *Ph.D. Dissertation*. University of Michigan, 2000.

[5] Grasman, S.E., T.L. Olsen and J.R. Birge "Setting Basestock Levels in Multi-product Systems with Setups and Random Yield," *Working Paper*.

[6] Jung, W.Y., and C.K. Un, "Analysis of Finite-Buffer Polling System with Exhaustive Service Based on Virtual Buffering," *IEEE Transactions on Communications*(42)12, 3144-3149, 1994.

[7] Lee, T.T., "M/G/1/N Queue with Vacation Time and Exhaustive Service Discipline," *Operations Research*(32)4, 774-784, 1984.

[8] Levy, H. and Sidi, M., "Polling Models: Applications, Modeling and Optimization, " *IEEE Transactions on Communications.*(38), 1750-1760, 1991.

[9] Sidi, M., H. Levy and S.W. Fuhrmann, "A Queueing Network with a Single Cyclically Roving Server," *Queueing Systems*(11), 121-144, 1992.

[10] Olsen, T.L., "A Practical Scheduling Method for Multi-Class Production Systems with Setups," *Management Science*(45)1, 116-130, 1999.

[11] Ross, S.M., *Stochastic Processes*, second edition. Wiley Press, New York, New York, 1996.

[12] Takagi, H., *Analysis of Polling Systems*. MIT Press, Cambridge, Massachusetts, 1986.

[13] Takagi, H., "Queueing Analysis of Polling Models: An Update," in *Stochastic Analysis of Computer and Communication Systems*, 267-318, H. Takagi, editor. Elsevier Science Publishers, Amsterdam, Netherlands, 1990.

[14] Takagi, H., "Applications of Polling Models to Computer Networks," *Computer Networks and ISDN Systems*(22), 193-211, 1991.

[15] Takagi, H., "Analysis of Finite Capacity Polling Systems," *Advanced Applied Probability*(23), 373-387, 1991.

[16] Takagi, H., "Queueing Analysis of Polling Models: Progress in 1990-1994", in *Frontiers In Queueing: Models and applications in science and engineering*, Chapter 5. CRC Press, New York, 1997.