We study the queue length performance in non-exhaustive asymmetric polling systems with Bernoulli feedback. We obtain the PGF (Probability Generating Functions) and the mean values of the queue lengths. For the gated polling model, we define two new service policies: departure-gated policy and service-gated policy, and we demonstrate their difference in queueing performance. In an asymmetric system, the departure-gated policy is better than the service-gated policy. In a symmetric system however, these two non-exhaustive policies are not as good as exhaustive policy. We also observe that the performance of L-limited system degrading faster as the load increases.

1. INTRODUCTION
The primary feature of various polling models is that a single service resource is shared by multiple queues in a cyclic order. These models are suitable for the modeling of a variety of systems in real world. In particular, they are widely used to evaluate various demand-based multiple-access schemes in computer and communications systems such as IEEE802.4 (token bus), IEEE802.5 (token ring), and recently cyclic scheduling in the study of ATM (Asynchronous Transfer Mode) networks [e.g.,16]. However, the basic polling models often fail to capture several important features in some applications. One of the major limitations is that a customer in each queue is assumed to leave system forever upon service completion. By including feedback, the polling models can be applied to analyze the performance of burst-level (or subburst-level) bandwidth allocation schemes [8, 10], in addition to its applications in the packet transmission in an error-prone channel and segmented message transmission.

The basic model has been extended from a variety of application aspects[1-4]. Specifically, the polling model with Bernoulli feedback is a very important general model. Various work includes mean sojourn time for a symmetric system [3,4] and for an asymmetric system [3], mean sojourn time of 3 service disciplines [4]. In [1], Takagi gives the mean customer sojourn time for a symmetric L-limited polling system. Moraes [2] corrected the errors in [1]. Moraes also presented an approximate analysis of mean sojourn time for an asymmetric system [3]. In [4], Takine, Takagi and Hasegawa proposed the expressions of the mean sojourn times in symmetric systems for three service disciplines of exhaustive, gated and 1-limited. They also provided a pseudo-conservation law with respect to the mean sojourn times for an asymmetric polling system with a mixture of the three disciplines. However, so far as we know, the queue length distribution has not been analyzed in symmetric or asymmetric polling system with Bernoulli feedback, though many researchers dealt with the queue length distribution of M/G/1 queue with Bernoulli feedback e.g., 5,6,7].

This paper concentrates on the analysis of queue length performance. We will compare the performance differences among gated policy, L-limited policy and exhaustive policy.

2. MODEL DESCRIPTIONS & ASSUMPTIONS
The system consists of N queues and one server. The server cyclically serves each queue in a fixed order. The arrival process at queue i is a Poisson process (denoted as \( \{(N_i(t), t \geq 0)\} \)) with rate \( \lambda_i \). The customers are served FCFS (First Come First Served). The switchover time from queue i to queue \( i+1 \) (mod N+1) is represented by \( R_i \). Its probability mass function (PMF), LST, mean, variance and second moment are respectively \( R_i(x) \), \( R_i(x) \), \( \bar{t}_i \), \( \sigma_i^2 \), and \( \sigma_i^{(2)} \). Each customer service time is \( B_i \), and let \( B_i(x), B_i^*(s), b_i, \) and \( b_i^{(2)} \) be its PMF, LST, mean and second moment respectively. When a customer completes its service, it is fed back instantaneously to the tail of the queue with probability 1 - \( \sigma_i \), or it leaves with the probability \( \sigma_i \).

In the modeling of a gated-service system, we use a definition that is different from the traditional definition. We refer to the customers already present in a queue when a server arrives as "gated-in" customers. In gated polling with Bernoulli feedback in this paper, we further classify the gated service policy into two types. (1) the departure-gated policy: If there are \( n \) gated-in customers when the queue is polled, the server would keep on serving the customers (those that required repeated services) until any \( n \) customers have completed their repeated service, or it leaves with the probability \( \sigma_i \).

In the modeling of 1-limited service system, recall that the traditional method has been to serve at most one customer when a server visits the queue. In our system with feedback, we would like to extend this notion. That is, upon polling a non-empty queue, the server would keep on serving the customers (those that required repeated services) until one (can be anyone) customer has completed his repeated service requirement and left the system. Note that the server would leave the queue at once upon polling an empty queue.
3. ANALYSIS OF A GATED SERVICE SYSTEM

We want to analyze the departure-gated and service-gated policies described in Section 2. Details of the derivation can be found in [18].

3.1 THE DEPARTURE-GATED POLICY

Under this policy, we want to obtain the PGF and mean of queue length at the polling instants. Let $U_i$ represent the total service time of a customer in queue $i$, and $U_{i1}, U_{i2}, U_{i3}, \ldots$ those of different customers in queue $i$. All $U_j$ and $U_{ij}$ are assumed to be iid variables, and $U_{i0} \equiv 0$. Let $B_{i1}, B_{i2}, B_{i3}, \ldots$ indicate the service times for different services. Therefore,

$$U_i = \sum_{k=1}^{\xi_i} B_{ik}$$

The key to analyzing this model is to establish the following important relationship among the random variables on the system. Let $L_i(t)$ denote the length of queue $i$ at time $t$, $	au_j(m)$ the instant of the $m$th server polling at queue $i$, and $\theta_{ijk}$ the busy period initiated by the $k$th customer in queue $i$, $i = 1, 2, \ldots, N$, $k = 0, 1, 2, \ldots$. Here we define $\theta_{i0} = 0$. For $j = 1, 2, \ldots, N$, let $L_j(t_j) = \lim_{m \to \infty} L_j(\tau_j(m))$. We have

$$L_i(\tau_{i+1}) = N_i(R_i) + N_i \left( \sum_{k=0}^{L_i(\tau_i)} U_{ik} \right)$$

$$L_j(\tau_{i+1}) = L_j(\tau_i) + N_j(R_i) + N_j \left( \sum_{k=0}^{L_j(\tau_i)} U_{jk} \right)$$

$$j \neq i, j = 1, 2, \ldots, N$$

$$L_i(\tau_i + 1) = \sum_{j=1}^{N} L_j(R_i) + \sum_{j=1}^{N} L_j(\tau_{j+1})$$

Let $F_i(z_1, z_2, \ldots, z_N)$ represent the PGF of the joint distribution of $(L_i(\tau_i), L_j(\tau_i), \ldots, L_N(\tau_i))$, and $G_i(z)$, the PGF of the distribution of $L_i(\tau_i)$. From (2) we have

$$F_i(z_1, z_2, \ldots, z_N) = E \left[ \prod_{j=1}^{N} Z_{j}^{L_j(\tau_j)} \right]$$

$$= E \left[ \prod_{j=1}^{N} Z_{j}^{N_j} \right] E \left[ \prod_{j=1}^{N} Z_{j}^{L_j(\tau_j)} \right]$$

$$= R_i \left( \sum_{j=1}^{N} L_j \right) \left( \sum_{j=1}^{N} \prod_{k=1}^{\xi_j} Z_{j+k}^{L_j(\tau_j)} \right)$$

$$= R_i \left( \sum_{j=1}^{N} L_j \right) \left( \sum_{j=1}^{N} \prod_{k=1}^{\xi_j} Z_{j+k}^{L_j(\tau_j)} \right)$$

and

$$G_i(z) = \prod_{j=1}^{N} R_j \left( \sum_{k=0}^{\xi_j} \left( \sum_{k=0}^{\xi_j} \prod_{k=1}^{\xi_j} Z_{j+k}^{L_j(\tau_j)} \right) \right)$$

Using (4), one can normally calculate the moments of the queue length distribution at polling instant. However, we give an alternative approach to directly computing the mean queue length as follows.

By averaging both sides of (2), and then letting $m$ approach infinity, we have

$$E[L_i(\tau_i)] = \lambda_i R + \lambda_i \sum_{j=0}^{\infty} E[L_j(\tau_j)],$$

which yields

$$E[L_i(\tau_i)] = \frac{\lambda_i R}{1 - \rho}.$$
4. ANALYSIS OF THE 1-LIMITED SERVICE SYSTEM

We notice that the exact solution to the traditional polling model has not been available for the general case. In the following, we present our analysis of $F_i(z_1, z_2, \cdots, z_N)$ under general conditions. As to the mean queue length, we give the result for the fully symmetric case.

As in the previous analysis, we begin our study by establishing the basic mathematical model. The expression of $F_i(z_1, z_2, \cdots, z_N)$ can be expressed in terms of $L_i(z_1, z_2, \cdots, z_N)$. We use $\alpha_i$ to represent the probability that queue $i$ is not empty when the server polls it. We define $\eta_i$ to be the random variable with binomial distribution, $B_i(1, \sigma_i)$, and $P(X)$ is the unit step function.

We establish the following relationship among the random variables in the system:

$$
E[L_i(\tau_i)] = \frac{\lambda_i R}{\sigma_i(1 - \rho)} .
$$

(10)

5. ANALYSIS OF THE EXHAUSTIVE SERVICE MODEL

The exhaustive service model has been analyzed by the authors [15]. Here we only summarize the result in order to make comparison with those of gated model and 1-limited model. As to the details about deriving the expressions for the mean queue lengths, refer to our another paper on exhaustive service polling model [15]. The PGF and mean of queue length at a polling instant are, respectively.

$$
G_i(z) = \prod_{j=1}^{N} R_j^{(\alpha_j - \lambda_j)} G(j_i^{(\alpha_j - \lambda_j)}, \beta_i^{(\lambda_j - \alpha_j)}, \cdots),
$$

$$
E[L_i(\tau_i)] = \frac{\lambda_i R_i^{(\alpha_i - \lambda_i)}}{\sigma_i(1 - \rho)},
$$

as well as

$$
E \left[ \prod_{j=1}^{N} z_j^{(\alpha_j - \lambda_j)} | L_i(\tau_i) > 0 \right] = \frac{1}{\alpha_i} \left[ F_i(z_1, z_2, \cdots, z_N) - F_i(z_1, z_2, \cdots, z_{i-1}, 0, z_{i+1}, \cdots, z_N) \right]
$$

(13d)

$$
E \left[ \prod_{j=1}^{N} z_j^{(\alpha_j - \lambda_j)} | L_i(\tau_i) = 0 \right] = \frac{1}{1 - \alpha_i}.
$$

(13e)

Substitute (13b)-(13e) into (13a), we obtain

$$
F_i(z_1, z_2, \cdots, z_N) = R_i \left[ \sum_{j=1}^{N} (1 - \lambda_j) z_j \right].
$$

(14)

For a fully symmetric system, we can use the approach in [9, pp120-122], and obtain the mean number of customers found in a queue at the polling instants. Alternatively, substituting $b$ and $\sigma b^{(2)} + 2b^2 - 2b^2 \sigma$ for $b$ and $b^{(2)}$ respectively in (6.18) of [9], the mean customer number is given as

$$
E[L_i(\tau_i)] = \frac{N \lambda_i ^2 \sigma_i ^2 + \delta ^2 + 2N \lambda_i \sigma_i}{2(\sigma - N \lambda_i (\sigma + b))} \left[ \frac{1}{2(\sigma - N \lambda_i (\sigma + b))} - \frac{1}{2(1 - N \rho)} \right].
$$

(15)

In order to take comparison study later, equation (6.18) in [9] is reproduced as follows

$$
E[L] = \frac{\lambda_i R}{2(1 - N \rho)} + \frac{N \lambda_i \sigma}{2(1 - N \rho)} + \frac{N \lambda_i ^2 b^{(2)}}{2(1 - N \rho)}
$$

5. ANALYSIS OF THE EXHAUSTIVE SERVICE MODEL

The exhaustive service model has been analyzed by the authors [15]. Here we only summarize the result in order to make comparison with those of gated model and 1-limited model. As to the details about deriving the expressions for the mean queue lengths, refer to our another paper on exhaustive service polling model [15]. The PGF and mean of queue length at a polling instant are, respectively.

$$
G_i(z) = \prod_{j=1}^{N} R_j^{(\alpha_j - \lambda_j)} G(\theta_i^{(\lambda_j - \alpha_j)}, \beta_i^{(\lambda_j - \alpha_j)}, \cdots)
$$

$$
E[L_i(\tau_i)] = \frac{\lambda_i R_i^{(\alpha_i - \lambda_i)}}{\sigma_i(1 - \rho)},
$$

(17)

As pointed out in [15], we can obtain the mean queue length via equation (17) directly instead of solving the $N^2$ equations.
6. PERFORMANCE EVALUATION
We give some numerical examples to compare the mean queue length performance of different service policies. We assume that the customer service time has an identical and exponential distribution, and the switchover time is constant. The mean queue lengths are computed in the case of fully symmetric stations for all the three kinds of service policies in order for comparison. The system parameters are $b=0.015$, $\sigma=0.50$, $r=0.01$, and $\lambda$ is varied from 1.65 to 7.5. The computed values of queue lengths are normalized with respect to the service time.

Fig. 1 shows the mean queue lengths at polling instant for different service policies. Let $\bar{Q}$ indicate the mean queue length. Using subscripts E, DG, L and SG to represent the exhaustive, departure-gated, 1-limited and service-gated policies, respectively, we have the following observation from our performance evaluation:

1. When system load is light, i.e., $\rho<0.55$, the relative queue performances for different service policies are as follows:

$$\bar{Q}_E \leq \bar{Q}_{DG} \leq \bar{Q}_L \leq \bar{Q}_{SG}$$

2. When traffic load is medium to heavy, the queue performance of 1-limited policy tends to decrease rapidly. The reason is that the server overhead (switchover time) is prominent in this situation.

In Fig.2 to 4, other numerical experiments were carried out for the asymmetric system where the asymmetry is characterized by the arrival rates of different queues. The value of $\lambda$ changes from 0.5 to 3.0. The arrival rates at the queues with heavy traffic are 1.5 times of that at the queues with light traffic. The system was assumed to have five queues including two heavily loaded queues. The mean values of service time, switchover time and feedback probability for every queue are 0.025, 0.02, and 0.5, respectively.

Fig.2 shows the mean queue lengths for the asymmetric exhaustive system. The queues with high traffic intensity have larger queue lengths than those with light traffic intensity in the same total load situation. Fig.3 and Fig.4 are for the departure-gated and service-gated system respectively. It is demonstrated that the difference between the heavily-loaded queues and lightly-loaded queues is not quite large. We have compared the three systems in the light load intensity situation and heavy load intensity situation respectively. The observations are the same as that for Fig.1. It demonstrates that the relative performance for the three policies is same in symmetric as in asymmetric systems.

7. CONCLUSIONS
In this paper, queue length performance for the three common polling models with Bernoulli feedback is considered. We derived the PGFs of queue lengths at the polling instants and/or server's leaving instants. By taking the derivative of the PGFs, we can numerically compute not only the mean but higher moments of queue length. Some examples are used to demonstrate the difference of the queue performance for the different policies. From the numerical results, queue performance for exhaustive policy is the best. The performance of 1-limited system degraded fast with the load increase. For the gate system, the departure-gate policy is better than the service-gate policy.

REFERENCES


Fig. 1 Mean Queue lengths for symmetric systems

Fig. 2 Light traffic vs heavy traffic for exhaustive system

Fig. 3 Light traffic vs heavy traffic for departure-gated system

Fig. 4 Light traffic vs heavy traffic for service-gated system