Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 109325, 9 pages
http://dx.doi.org/10.1155/2015/109325

*Research Article*

# On Two-Level State-Dependent Routing Polling Systems with Mixed Service

**Guan Zheng,[1] Yang Zhijun,[2] Qian Wenhua,[1] and He Min[1]**

[1] *School of Information Science and Technology, Yunnan University, Kunming 650091, China*
[2] *Educational and Scientific Institute, Educational Department of Yunnan Province, Kunming 650223, China*

Correspondence should be addressed to Guan Zheng; gz_627@sina.com

Received 19 March 2015; Accepted 1 September 2015

Academic Editor: Ruihua Liu



Based on priority differentiation and efficiency of the system, we consider an $N + 1$ queues' single-server two-level polling system which consists of one key queue and $N$ normal queues. The novel contribution of the present paper is that we consider that the server just polls active queues with customers waiting in the queue. Furthermore, key queue is served with exhaustive service and normal queues are served with 1-limited service in a parallel scheduling. For this model, we derive an expression for the probability generating function of the joint queue length distribution at polling epochs. Based on these results, we derive the explicit closed-form expressions for the mean waiting time. Numerical examples demonstrate that theoretical and simulation results are identical and the new system is efficient both at key queue and normal queues.

## 1. Introduction

In this paper, we study a class of $N+1$ queues' polling systems that consists of one key queue, $Q_h$, and $N$ normal queues, $Q_1, Q_2, \dots, Q_N$, which are attended by a single server. Studies on the polling systems have attracted extensive attentions in the last years due to their vast area of applications in communication network, production, and transportation. Excellent surveys on polling systems analysis and their applications may be found in [1–4]. However, many studies in the literatures assume that the server visit the queues in a fixed, cyclic order. This might not be a realistic assumption, as queues might have different priority level; queues with high priority should be visit more frequently than the lower ones; sometime queues might be empty and then there is no need to visit. As such, we study the case where the server just visits active queues with customers. Note that as a consequence, after skipping the empty queues, server could provide more visit opportunity to active queues with customers. Furthermore, parallel process of service period and switch-over period allows a successive service between two active queues without the duration of switch-over time. To provide priority differentiation service, queues are separated as one key queue and $N$ normal queues. Two-level route order and mixed service scheme are used to provide high priority to key queue.

It is observed that in the wide body of literature on polling system hardly can any studies be found that take the consideration of queue state-dependent routing and service priority simultaneously. The reason for this may lie in the fact that the analysis of state-dependent routing polling model is much more complex than that of cyclic polling model, especially in priority differentiated model. In particular, waiting time and queue length analysis of two-level priorities polling systems can be found in [5–7], in which the server visits queues in a two-level route; that is, the server polls key queue with exhaustive scheme after each gated service to normal queue [5]. This work is extended in [6] with assigning 1-limited service discipline to normal queues. More recently, Yang et al. set the exhaustive service for normal queue and gated service for key queue to ensure fairness but just acquire the first moment performance of the system as mean queue length at the polling epoch and the mean cyclic time [7]. The parallel discipline is used to improve the delay performance in [8], in which when the current polling queue has customers in storage the server will process service while switching to the successive queue simultaneously and begins to serve the successor once it finishes the service of the current one. This scheme could improve polling efficiency in high traffic cases. However, the parallel mechanism will be invalid when there is no customer in the queue. In low traffic cases, useless polling to idle queue becomes an obvious liability in cyclic polling model. Routing depends on the event whether a queue is empty or it is not helpful to this problem [9]. In this paper, we consider the special setting to a two-level mixed service polling model, where the key queue is served exhaustively while normal queues are served in 1-limited mechanism. Furthermore, the server no longer checks all the stations in a fixed order; only active stations with transfer requirements could be served and then the switch-over period and service period are processed paralleled. This mechanism increases the system utilization and reduces the mean waiting time.

Although the exhaustive service discipline in principle fits the branching property, the present model involves 1-limited service discipline, which does not satisfy the above-mentioned branching property. The explicit analysis of non-branching service disciplines is mostly in special setting, such as [10, 11] studied on two-queue polling systems and [12] studied on symmetric 1-limited model. In this paper, we follow the special setting in [8] and analyze the mean waiting time of the present model under the assumption on the symmetrical characteristic among normal queues, as will be described in greater detail in Section 2.

Initially, we follow an approach similar to the analysis of [5], which uses a recursive iteration of a functional equation, for the probability generating function (PGF) of the joint queue-length distribution at moments the server starts a visit period.

The main contributions of this paper can be summarized as follows. Firstly, we extend the parallel two-level poling system in [8] by using queue state-dependent routing, in which only active queues with customers could be visited by server. This scheme is helpful to avoid the consumptions induced by idle visit. Secondly, under the assumption of a stable system, we obtain the explicit expressions for the PGF for the joint queue length distribution at polling epochs as a starting point of key queue and normal queue separately. Thirdly, we achieve the exact closed-form expression of the mean waiting time under the assumption on the symmetrical characteristic of normal queue.

The rest of the paper is structured as follows. In Section 2, we give a formal description of the polling model that we study and we introduce the necessary notation. Based on this, in Section 3, we derive the expressions for the mean waiting time of the present model under the assumption of a semisymmetric (symmetrical characteristic of normal queue) stable system, by taking a functional equation for the PGF for the joint queue length distribution at polling epochs as a starting point. In Section 4, numerical results obtained with the proposed analytical models are shown and their very good agreement with realistic simulation results is discussed. Finally, concluding remarks and directions for future research are given in the end.

## 2. Model Description

Consider a discrete time (timeline is divided into time slot) polling system consisting of $N$ ($N \geq 2$) infinite-buffer queues $Q_1, Q_2, \ldots, Q_N$, and $Q_h$. The single server visits active queues in a two-level state-dependent routing order and serves the customers with mixed service discipline.

*In the arrival process*, type-$j$ ($j = 1, 2, \ldots, N, h$) customers arrive at $Q_j$ according to an independent Poisson arrival process. The generating function of arrival process in queue $j$ is $A_j(z_j)$, with the variance of $\sigma_{\lambda j}^2 = A_j''(1) + \lambda_j - \lambda_j^2$ and the arrival rate of $\lambda_j = A_j'(1)$. The total arrival rate is $\sum_{i=1}^{N} \lambda_i + \lambda_h$.

In the *service process*, we assume that customers in queue $j$ ($j = 1, 2, \ldots, N, h$) receive individual service. The service time of a customer at each queue is independent of each other. Their generating function is $B_j(z_j)$, with the variance of $\sigma_{\beta j}^2 = B_j''(1) + \beta_j - \beta_j^2$ and the mean value $\beta_j = B_j'(1)$. We propose a two-level server routing make the high priority queue be visited more frequently than others and add mix-service discipline to ensure the high priority of $Q_h$. The load offered to $Q_j$ is $\rho_j = \lambda_j \beta_j$, and the total offered load is equal to $\sum_{i=1}^{N} \rho_i + \rho_h$.

*State-Dependent Routing*. Queues are partitioned as active queue and idle queue by their buffer condition. Only active queues with customers waiting in the buffer could be visited by the server in order. Idle queue with empty buffer would be skipped in the current polling round.

*Two-Level Polling*. The server visits queues governed by a two-level routing. In the first polling level, the server polls between the high priority queue $Q_h$ and an active normal queue; in the second level, for each time after the exhaustive service at $Q_h$, one normal active queue is visited in a cyclic order; that is, the server routing in this model is $1 \rightarrow h \rightarrow \cdots \rightarrow i \rightarrow h \rightarrow i+1 \rightarrow \cdots \rightarrow h \rightarrow N$.

In the *switch-over process*, a parallel mechanism is used. When the server polls an active queue at time with customers in its buffer, the server will provide service and inquire the next active queue simultaneously and then switch to serve the successor immediately without the switch-over time once it has finished the current service. Combined with the state-dependent routing scheme, over the course of a visit period, the server serves the active queues and normal queue in sequence continuously until the entire system is empty; there will be no consumption of switch-over time anymore in the present model. More especially, we assume the server consume one time slot to confirm the system state when the system is entirely empty.

*Mix-Service Discipline.* Exhaustive discipline is specified for the key queue and 1-limited discipline for normal queues, so that the entire customers in the key queue could be served in the present server round, while those who are in normal queues might need several rounds when there are more than one customer in the buffer. Let $F_h$ denote the duration of a service period for the customers arrive during arbitrary time slot in $Q_h$. This service period consists of the services of its ancestral customers arriving during the exact slot and the services of the offspring line of the ancestral customers [13]. The generating function of $F_h$ is denoted by $F_h(z_h) = E[z_h^{F_h}]$. Such a functional equation has already been derived in [14] as $F_h(z_h) = A_h(B_h(z_h F(z_h)))$.

In the remainder of this paper, we are interested in the queue length distributions at the polling epoch of $Q_i$ and $Q_h$. Let $\xi_j(n)$ denote the number of customers present at $Q_j$ at $t_n$ when the server starts a visit period at $Q_i$, and let $\xi_j(n^*)$ denote the number of customers present at $Q_j$ at $t_n^*$ when the server starts a visit period at $Q_h$ successively with the service of $Q_i$. The joint distribution of $\xi_j(n+1)$ and $\xi_j(n^*)$ is represented by the $N$-dimensional PGF $G_{i+1}(z_1, \ldots, z_N, z_h)$ and $G_{ih}(z_1, \ldots, z_N, z_h)$.

We analyze the system under stability conditions ($\sum_{i=1}^{N} \rho_i + \rho_h < 1$) [12]. Normal queues in the present model are served in a 1-limited manner, which does not satisfy the well-known branching property in polling systems. Therefore, more specifically, in the analyses of mean waiting time, we assume the normal queues are symmetric; that is, normal queues have the same customer arrival rate and service rate.

## 3. Analysis for Steady-State Systems

In this section, we derive explicit expression for the joint queue length distribution. In Section 3.1, we first obtain expressions for $G_{i+1}(z_1, \ldots, z_N, z_h)$ and $G_{ih}(z_1, \ldots, z_N, z_h)$, the joint queue length PGF at the polling epoch at $Q_{i+1}$ and $Q_h$. These results ultimately lead in Section 3.2 to the first and second moment of the PGF, and obtain the expressions for $E[W_i]$ and $E[W_h]$, the mean waiting time of type-$i$ and type-$h$ customers that arrive at an arbitrary point in time.

*3.1. Joint Queue Length Distribution at Polling Epoch.* Assuming that the server begin the service of $Q_i$ at $t_n$, define a random variable $\xi_j(n)$ as the number of type-$j$ ($j = 1, 2, \ldots, N, h$) customers at time $t_n$. Then the status of the entire polling model at time $t_n$ can be represented as $\{\xi_1(n), \ldots, \xi_N(n), \xi_h(n)\}$. Denote $\xi_j(n+k)$ as the number of type-$j$ customers at $t_{n+k}$, the polling epoch of $Q_{i+k}$. The status of the entire polling model at time $t_{n+k}$ can be represented as $\{\xi_1(n+k), \ldots, \xi_N(n+k), \xi_h(n+k)\}$ while $\xi_i(n^*)$ is the number of type-$j$ customers in at time $t_n^*$, at which the server begins providing service to $Q_h$ and the status of the entire polling model at time $t_n^*$ can be represented as $\{\xi_1(n^*), \ldots, \xi_N(n^*), \xi_h(n^*)\}$. Under the necessary and sufficient condition for the stability of the system $\sum_{i=1}^{N} \rho_i + \rho_h < 1$, the probability distribution is defined as

$$\begin{aligned} &\lim_{n\to\infty} P\left[\xi_j(n) = x_j;\ j = 1, \ldots, N, h\right] \\ &\quad = \pi_i(x_1, \ldots, x_N, x_h), \\ &\lim_{n\to\infty} p\left[\xi_j(n^*) = y_j;\ j = 1, \ldots, N, h\right] \\ &\quad = \pi_{ih}(y_1, \ldots, y_N, y_h). \end{aligned} \tag{1}$$

The generating functions at $t_n$ and $t_n^*$ are

$$\begin{aligned} &G_i(z_1, \ldots, z_N, z_h) \\ &\quad = \sum_{x_1=0}^{\infty} \cdots \sum_{x_N=0}^{\infty} \sum_{x_h=0}^{\infty} z_1^{x_1} \cdots z_N^{x_n} z_h^{x_h} \pi_i(x_1, \ldots, x_N, x_h) \\ &\qquad\qquad i = 1, 2, \ldots, N, \\ &G_{ih}(z_1, \ldots, z_N, z_h) \\ &\quad = \sum_{y_1=0}^{\infty} \cdots \sum_{y_N=0}^{\infty} \sum_{y_h=0}^{\infty} z_1^{y_1} \cdots z_N^{y_n} z_h^{y_h} \pi_{ih}(y_1, \ldots, y_N, y_h) \\ &\qquad\qquad i = 1, 2, \ldots, N. \end{aligned} \tag{2}$$

According to the proposed mechanism, the system variables have the following equations. When the server begins the service on $Q_{i+1}$ at $t_{n+1}$, we have

$$\xi_j(n+1) = \begin{cases} \xi_j(n^*) + \eta_j(\nu_h) & j \neq h \\ 0 & j = h. \end{cases} \tag{3}$$

$\nu_j(n)$ is the service time in $Q_j$ and $\eta_k(\nu_j)$ is the number of arrivals to $Q_k$ during $\nu_j(n)$.

The server just finishes the service of $Q_h$ in an exhaustive manner and starts the polling on $Q_{i+1}$ at $t_{n+1}$. Such a functional equation of exhaustive service has already been derived in [12]. Applying these results to our case, we obtain

$$\begin{aligned} &G_{i+1}(z_1, z_2, \ldots, z_N, z_h) = \lim_{n\to\infty} E\left[\prod_{j=1}^{N} z_j^{\xi_i(n+1)} z_h^{\xi_h(n+1)}\right] \\ &\quad = G_{ih}\left(z_1, z_2, \ldots, z_N,\right. \\ &\qquad \left. B_h\left(\prod_{j=1}^{N} A_j(z_j) F_h\left(\prod_{j=1}^{N} A_j(z_j)\right)\right)\right). \end{aligned} \tag{4}$$

The expression can be interpreted as follows. At the start of the visit period at $Q_{i+1}$, type-$i$ customers are those at the polling epoch of $Q_h$ plus the new customers arriving at each queue during the service period of the $Q_h$ in exhaustive scheme, and no type-$h$ customer resumes at that moment.

When the server begins the service on $Q_h$ at $t_n^*$, we have

$$\begin{aligned} &\xi_j(n^*) = \begin{cases} \xi_j(n) + \eta_j(\nu_i), & j \neq i \neq h, \\ \xi_i(n) + \eta_i(\nu_i) - 1, & j = i \quad \xi_i(n) \neq 0, \\ \eta_j(\nu_i), & j = h, \end{cases} \\ &\xi_j(n^*) = \begin{cases} \xi_j(n), & j \neq i \neq h, \\ 0, & j = i \quad \xi_i(n) = 0, \\ 0, & j = h \end{cases} \end{aligned} \tag{5}$$

$\nu_j(n)$ is the service time in $Q_j$, and $\eta_k(\nu_j)$ is the number of arrivals to $Q_k$ during $\nu_j(n)$.

In our case, for normal queues, the server just polls the active queues with customers in parallel 1-limited manner. To gain more insight in the state-dependent service discipline, let $P_i$ denote the queue length at the service epoch in an M/G/1 queue with the same arrival process and service-time distribution as $Q_i$. We assume that the $k$ customers have waited in $Q_i$ at the start of the busy period with probability $p_k \in [0,1)$, $\sum_{k=0}^{\infty} p_k = 1$. Then we can acquire the queue length generating function at the service epoch as $P_i(z_i) = A_i(z_i)\sum_{k=0}^{\infty} p_k z_i^k$, where $A_i(z_i)$ is the PGF of the arrival process as defined in Section 2. Specifically, the server does not provide service when the queue length is zero, so we assume that $k^*$ customers resumed after the end of the busy time in 1-limited service with the probability of $p_k^* \in [0,1)$, and $p_k^* = p_k + 1$ for $k = 0, 1, \ldots$. Consequently, the probability space could be rebuilt as

$$\begin{aligned} P_i^*(z_i) &= B_i(A_i(z_i))\,A_i(z_i)\left(p_0 + \sum_{k=0}^{\infty} p_k^* z_i^k\right) \\ &= B_i(A_i(z_i))\left(\frac{\sum_{k=0}^{\infty} p_k z_i^k - p_0 z_i^0}{z_i} + p_0 z_i^0\right). \end{aligned} \tag{6}$$

With the definition of $P_i(z_i)$, we have

$$\begin{aligned} P_i^*(z_i) = B_i(A_i(z_i))\,&\frac{\left(P_i(z_i) - P_i(z_i)\big|_{z_i=0}\right)}{z_i} \\ &+ P_i(z_i)\big|_{z_i=0}. \end{aligned} \tag{7}$$

Applying these results to our case, we obtain

$$\begin{aligned} G_{ih}(z_1,\ldots,z_N,z_h) &= \lim_{n\to\infty} E\left[\prod_{j=1}^{N} z_j^{\xi_i(n^*)} z_h^{\xi_h(n^*)}\right] = \frac{1}{z_i} \\ &\cdot B_i\left(\prod_{j=1}^{N} A_j(z_j)\,A_h(z_h)\right) \\ &\cdot \left[G_i(z_1,\ldots,z_N,z_h) - G_i(z_1,\ldots,z_N,z_h)\big|_{z_i=0}\right] \\ &+ G_i(z_1,\ldots,z_N,z_h)\big|_{z_i=0} \\ &- G_i(z_1,\ldots,z_N,z_h)\big|_{z_1,\ldots,z_N,z_h=0} \\ &+ \prod_{j=1}^{N} A_j(z_j)\,A_h(z_h)\,G_i(z_1,\ldots,z_N,z_h)\big|_{z_1,\ldots,z_N,z_h=0}. \end{aligned} \tag{8}$$

The expression can be interpreted as follows. At the start of the visit period at $Q_h$, in the case that the former $Q_i$ is active, one type-$i$ customer would have been served at $t_n^*$ and new customers arrived at each queue during the service period of the exact type-$i$ customer. The server would skip $Q_i$ to $Q_{i+1}$ when $Q_i$ is empty; in that case, the distribution of the number of customers in the systems is represented by the generating function $G_i(z_1,\ldots,z_i,\ldots,z_{N-1},z_h)|_{z_i=0}$, with the exception as the system is entirely empty, which is represented by the generating function $G_i(z_1,\ldots,z_i,\ldots,z_{N-1},z_h)|_{z_1,\ldots,z_{N-1},z_h=0}$. When the system is entirely empty, the server will stop providing service for one time slot until new customers arrive during this time slot, and this number of customers is represented by the last partition of the addition formula.

### 3.2. Expression for the Mean Waiting Time

*3.2. Expression for the Mean Waiting Time.* Now we have derived expressions for the PGF $G_{i+1}(z_1,\ldots,z_i,\ldots,z_N,z_h)$ and $G_{ih}(z_1,\ldots,z_i,\ldots,z_N,z_h)$ pertaining to the queue length at polling epoch of $Q_{i+1}$ and $Q_h$, we use these results to obtain $E[W_i]$, the mean waiting time of type-$i$ normal customers, and $E[W_h]$, the mean waiting time of type-$h$ high priority customers.

*3.2.1. The First and Second Moment of $G_{i+1}(z_1,\ldots,z_i,\ldots,z_N,z_h)$ and $G_{ih}(z_1,\ldots,z_i,\ldots,z_N,z_h)$.* To start the analysis of mean waiting time of type-$j$ customers, we need to calculate the generating functions and its derivation at the point $\mathbf{z} = \mathbf{1}$, $\mathbf{z}$ is the abbreviation of the $(1\times N+1)$ vector of $(z_1,\ldots,z_i,\ldots,z_N,z_h)$, and $\mathbf{1}$ is the $(1\times N+1)$ vector with 1.

$G_{i+1}(\mathbf{z})$ is the PGF of the joint queue length at the polling epoch of $Q_i$, so we have

$$\begin{aligned} &G_i(z_1,\ldots,z_i,\ldots,z_N,z_h) \\ &= \sum_{x_1=0}^{\infty}\cdots\sum_{x_i=0}^{\infty}\cdots\sum_{x_N=0}^{\infty}\sum_{x_h=0}^{\infty} z_1^{x_1}\cdots z_i^{x_i}\cdots z_N^{x_n} z_h^{x_h} P\left(\xi_1(n)=x_1,\ldots,\xi_i(n)=x_i,\ldots,\xi_N(n)=x_N,\xi_h(n)=x_h\right) \\ &= \sum_{x_1=0}^{\infty}\cdots\sum_{x_i=0}^{\infty}\cdots\sum_{x_N=0}^{\infty}\sum_{x_h=0}^{\infty} z_1^{x_1}\cdots z_i^{x_i}\cdots z_N^{x_n} z_h^{x_h} P\left(\xi_1(n)=x_1,\ldots,\xi_N(n)=x_N,\xi_h(n)=x_h \mid \xi_i(n)=x_i\right) P\left(\xi_i(n)=x_i\right). \end{aligned} \tag{9}$$

Taking the $k$th derivative with respect to $z_i$ yields

$$\begin{aligned} \frac{\partial^k G_i(z_1,z_2,\ldots,z_i,\ldots,z_N,z_h)}{\partial z_i^k} &= \sum_{x_1=0}^{\infty}\cdots\sum_{x_N=0}^{\infty}\sum_{x_h=0}^{\infty} z_1^{x_1}\cdots z_i^{x_i-k}\cdots z_N^{x_N} z_h^{x_h} \\ &\cdot \frac{x_i!}{(x_i-k)!} P\left(\xi_1(n)=x_1,\ldots,\xi_N(n)=x_N,\xi_h(n)=x_h \mid \xi_i(n)=x_i\right) P\left(\xi_i(n)=x_i\right). \end{aligned} \tag{10}$$

Setting $z_i = 0$ yields

$$\begin{aligned}
&\left.\frac{\partial^k G_i(z_1, z_2, \dots, z_i, \dots, z_N, z_h)}{\partial z_i^k}\right|_{z_i=0} \\
&= \sum_{x_1=0}^{\infty} \cdots \sum_{x_N=0}^{\infty} \sum_{x_h=0}^{\infty} z_1^{x_1} \cdots 1 \cdots z_N^{x_N} z_h^{x_h} k! P(\xi_1(n) = x_1, \dots, \xi_N(n) = x_N, \xi_h(n) = x_h \mid \xi_i(n) = k) P(\xi_i(n) = k) \\
&= k! P(\xi_i(n) = k) \sum_{x_1=0}^{\infty} \cdots \sum_{x_N=0}^{\infty} \sum_{x_h=0}^{\infty} z_1^{x_1} \cdots 1 \cdots z_N^{x_N} z_h^{x_h} P(\xi_1(n) = x_1, \dots, \xi_N(n) = x_N, \xi_h(n) = x_h \mid \xi_i(n) = k) \\
&= k! P(\xi_i(n) = k) E\left[z_1^{\xi_1(n)} \cdots 1 \cdots z_N^{\xi_N(n)} z_h^{\xi_h(n)} \mid \xi_i(n) = k\right].
\end{aligned} \tag{11}$$

Rearranging terms and setting $k = 0$, we have

$$\begin{aligned}
&G_i(z_1, z_2, \dots, z_i, \dots, z_N, z_h)\big|_{z_i=0} = P(\xi_i(n) = 0) \\
&\quad \cdot E\left[z_1^{\xi_1(n)} \cdots 1 \cdots z_N^{\xi_N(n)} z_h^{\xi_h(n)} \mid \xi_i(n) = 0\right], \\
&G_i(\mathbf{1}_i) = P(\xi_i(n) = 0).
\end{aligned} \tag{12}$$

Extending this result we have

$$\begin{aligned}
G_i(\mathbf{0}) &= P\{\xi_1(n) = 0, \dots, \xi_i(n) = 0, \dots, \xi_N(n) \\
&= 0,\ \xi_h(n) = 0\}.
\end{aligned} \tag{13}$$

$\mathbf{0}$ is the $(1 \times N + 1)$ vector with 0, and $\mathbf{1}_j$ is the $(1 \times N + 1)$ vector with 0 in $j$th position and 1 in all other entries.

Define the first derivative of $G_i(\mathbf{z})$ and $G_{ih}(\mathbf{z})$ at $\mathbf{z} = \mathbf{1}$ as

$$\begin{aligned}
g_i(j) &= \lim_{z_1, \dots, z_i, \dots, z_N, z_h \to 1} \frac{\partial G_i(\mathbf{z})}{\partial z_j}, \\
g_{ih}(j) &= \lim_{z_1, \dots, z_i, \dots, z_N, z_h \to 1} \frac{\partial G_{ih}(\mathbf{z})}{\partial z_j}, \\
&\qquad j, k = 1, 2, \dots, N, h.
\end{aligned} \tag{14}$$

$$g_{ih}(j) = \beta_i \lambda_j [1 - G_i(\mathbf{1}_i)] + g_i(j) + \lambda_j G_i(\mathbf{0}) \tag{15}$$

$$g_{ih}(i) = (\beta_i \lambda_i - 1)[1 - G_i(\mathbf{1}_i)] + g_i(i) + \lambda_i G_i(\mathbf{0}) \tag{16}$$

$$g_{ih}(h) = \beta_i \lambda_h [1 - G_i(\mathbf{1}_i)] + \lambda_h G_i(\mathbf{0}) \tag{17}$$

$$g_{i+1}(i) = g_{ih}(i) + g_{ih}(h) \beta_h \lambda_i \left(1 + F_h'(1)\right) \tag{18}$$

$$g_{i+1}(j) = g_{ih}(j) + g_{ih}(h) \beta_h \lambda_j \left(1 + F_h'(1)\right). \tag{19}$$

Calculate $\sum_{j=1}^{N} g_{j+1}(k)$ yields

$$1 - G_i(\mathbf{1}_i) = \frac{N \lambda_i G_i(\mathbf{0})}{1 - \rho_h - N\rho}. \tag{20}$$

Define the second derivative of $G_i(\mathbf{z})$ and $G_{ih}(\mathbf{z})$ at $\mathbf{z} = \mathbf{1}$ as

$$\begin{aligned}
g_i(j, k) &= \lim_{z_1, \dots, z_i, \dots, z_N, z_h \to 1} \frac{\partial^2 G_i(\mathbf{z})}{\partial z_j \partial z_k} \\
g_{i0}(j, k) &= \lim_{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_N, z_h \to 1} \frac{\partial^2 G_i(\mathbf{z})\big|_{z_i=0}}{\partial z_j \partial z_k} \\
g_{i00}(j, k) &= \lim_{z_1, \dots, z_N, z_h \to 0} \frac{\partial^2 G_i(\mathbf{z})\big|_{z_1, \dots, z_N, z_h=0}}{\partial z_j \partial z_k} \\
g_{ih}(j, k) &= \lim_{z_1, \dots, z_N, z_h \to 1} \frac{\partial^2 G_{ih}(\mathbf{z})}{\partial z_j \partial z_k} \\
&\qquad i = 1, 2, \dots, N \quad j, k = 1, 2, \dots, N, h.
\end{aligned} \tag{21}$$

Substitute (4) and (8) into the above second derivative formulas.

We assume the $N$ normal queues are symmetrical; that is, $\lambda_i = \lambda$, $\beta_i = \beta$, $i = 1, 2, \dots, N$. Then simplifying these we get the second derivative of $G_i(\mathbf{z})$ and $G_{ih}(\mathbf{z})$ at $\mathbf{z} = \mathbf{1}$ as follows:

$$\begin{aligned}
g_{ih}(h, h) &= B''(1) \lambda_h^2 (1 - G_i(\mathbf{1}_i)) + \beta A_h''(1) (1 \\
&\quad - G_i(\mathbf{1}_i)) + A_h''(1) G_i(\mathbf{0}).
\end{aligned} \tag{22}$$

$$\begin{aligned}
g_i(i) &= \frac{1 - G_i(\mathbf{1}_i)}{2} \left\{ \frac{1}{(1 - \rho_h - N\rho)(1 - \rho_h)} \left[ \rho_h \right.\right. \\
&\quad \cdot \frac{A''(1)}{\lambda} \left(1 - \rho_h^2 + A_h'' \beta_h^2 + \lambda_h B_h''\right) + N B''(1) \lambda^2 \\
&\quad \left.\left. + N \beta A''(1) \right] + \frac{\lambda}{(1 - \rho_h)} + \frac{\lambda_h \lambda B_h''(1)}{1 - \rho_h - N\rho} \right\} + 1 \\
&\quad - G_i(\mathbf{1}_i).
\end{aligned} \tag{23}$$

*Remark 1.* Though $g_i(i)$ is the first derivative at $\mathbf{z} = \mathbf{1}$ $G_i(\mathbf{z})$ in definition, it is clear that it contains the second moment parameter as $A_j''(1)$ and $B_j''(1)$. So, $g_i(i)$ is a second moment parameter for the system performance.

*3.2.2. Analysis of* $E[W_h]$ *and* $E[W_i]$. Define $W_h$ and $W_i$ as the waiting time of type-$h$ and type-$i$ customers, which denotes the time from the epoch when a customer arrives at the queue to the time it is served. In the present model, high priority type-$h$ customers are served in the exhaustive service and normal type-$i$ customers are served in 1-limited service. Based on the related research works in [14], the mean waiting time of type-$h$ customers $E[W_h]$ and the type-$i$ customers $E[W_i]$ can be calculated as follows:

$$E\left[W_h\right] = \frac{g_{ih}(h,h)}{2\lambda_h g_{ih}(h)} - \frac{A_h''(1)}{2\lambda_h^2(1+\rho_h)} + \frac{\lambda_h B_h''(1)}{2(1-\rho_h)}, \quad (24)$$

$$E\left[W_i\right] = \frac{1}{\lambda(1-G_i(\mathbf{1}_i))} g_i(i) - \frac{1}{\lambda} - \frac{A''(1)}{2\lambda^2}. \quad (25)$$

Taking (17), (22) in (24) in the above expressions, we have

$$\begin{aligned} E\left[W_h\right] = {} & \frac{1}{2(1-\rho_h)}\left(1-\rho_h + N\lambda B''(1) + \lambda_h B_h''(1)\right) \\ & - \frac{1}{2{\lambda_h}^2(1+\rho_h)} A_h''(1). \end{aligned} \quad (26)$$

Taking (17), (22), and (23) in (25) in the above expressions, we have

$$\begin{aligned} E\left[W_i\right] = {} & \frac{1}{2\lambda}\left\{\frac{1}{(1-\rho_h-N\rho)(1-\rho_h)}\left[\rho_h\right.\right. \\ & \cdot \frac{A''(1)}{\lambda_i}\left(1-\rho_h^2 + A_h''\beta_h^2 + \lambda_h B_h''\right) + NB''(1)\lambda^2 \\ & \left.\left. + N\beta A''(1)\right] + \frac{\lambda}{(1-\rho_h)} + \frac{\lambda_h \lambda B_h''(1)}{1-\rho_h-N\rho}\right\} \\ & - \frac{A''(1)}{2\lambda^2}. \end{aligned} \quad (27)$$

## 4. Numerical Study

In this section we study the accuracy of the theoretical analysis and compare the mean waiting time of the present model with two existing two-level polling models. Consider an $N+1$ queues' model with one high priority queue $Q_h$ and $N$ normal queues $Q_i$ $(i = 1, \ldots, N)$ defined as follows: the service times of all customers are exponentially distributed with mean $\beta$ in $Q_i$ and $\beta_h$ in $Q_h$. The arrival processes are Poisson process with rate $\lambda$ in $Q_i$ and $\lambda_h$ in $Q_h$. The relative parameter values are listed in Table 1, in which $\{a : k : b\}$ means the parameter is varied between $a$ and $b$ in steps of $k$.

From Figure 1, we can clearly see that, firstly, the theoretical value and the simulation result coincided with each other. Secondly, when the total offered load grew with the arrival rate, service time, and the number of queues, with the mean waiting time increasing distinctly in $Q_i$, while the performances in $Q_h$ are much better, both queue and mean waiting time are much lower than normal queues, and the growth in $Q_h$ with the total offered load presents much more smoothly.

It is worth considering whether the state-dependent mechanism improves the performance of the system comparing with the existing two-level polling systems. In order to answer this question, we compare a classical two-level system with switch-over time [6], abbreviated as classical system and a parallel two-level system [8], abbreviated as parallel system in Figure 2. The service discipline in the comparisons is 1-limited service for normal queues and exhaustive service for the key queue. Overall models have the same test bed as shown in Table 1. We just vary the working mechanism.

Figure 2 shows the mean waiting time of normal queues in (a) and mean waiting time of key queue in (b). Comparing with the forgoing, the state-dependent system achieves a better performance in delay guarantee and stability. It is clear in Figure 2(a), for lower load, in most of the cases, that there is no customer in the buffers; thus a switch-over time is necessary when the server switches between $Q_i$ and $Q_h$ in the classical and parallel system, while the empty queues would be skipped in the present model. Therefore, customers in the state-dependent system achieve a lower mean waiting time, which is under 20% of the forgoing. In the heavy traffic, the server could not provide service in the necessary switch-over time for the classical system; consequently, it becomes unstable when the arrival rate of $Q_i$ grows over 0.06 in this case. The parallel system and the state-dependent system have better performance in system stability; especially in state-dependent system, the mean waiting time of the normal customers has less than 50% of which in the parallel system. A conclusion can be drawn from a comparison between Figures 2(a) and 2(b), which is that for all three two-level models the mean waiting time of the customers in key queue is significantly lower than that in normal queues, and as illustrated in Figure 2(b), the mean waiting time for $h$-type customers in state-dependent system is lower than that of the others.

## 5. Conclusion

When comparing the model of the present paper with the existing literature, the contribution of the present paper is twofold. One of the most striking differences is the queues which are partitioned as active queue and idle queue by their buffer condition, and only active queues with customers waiting in the buffer could be visited by the server in a two-level order. As illustrated in the numerical example, both $i$-type customers in normal queues and $h$-type customers in key queue acquire better delay performance than those in systems without queue-stated differentiation. Another notable contribution of the paper is that we achieve the closed-form exact expressions of the mean waiting time for customers in normal queues and key queue, under the assumption of the symmetric of normal queues. The total unknowns in these equations are all first moments of random variables and, thus, no correlation terms are required.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

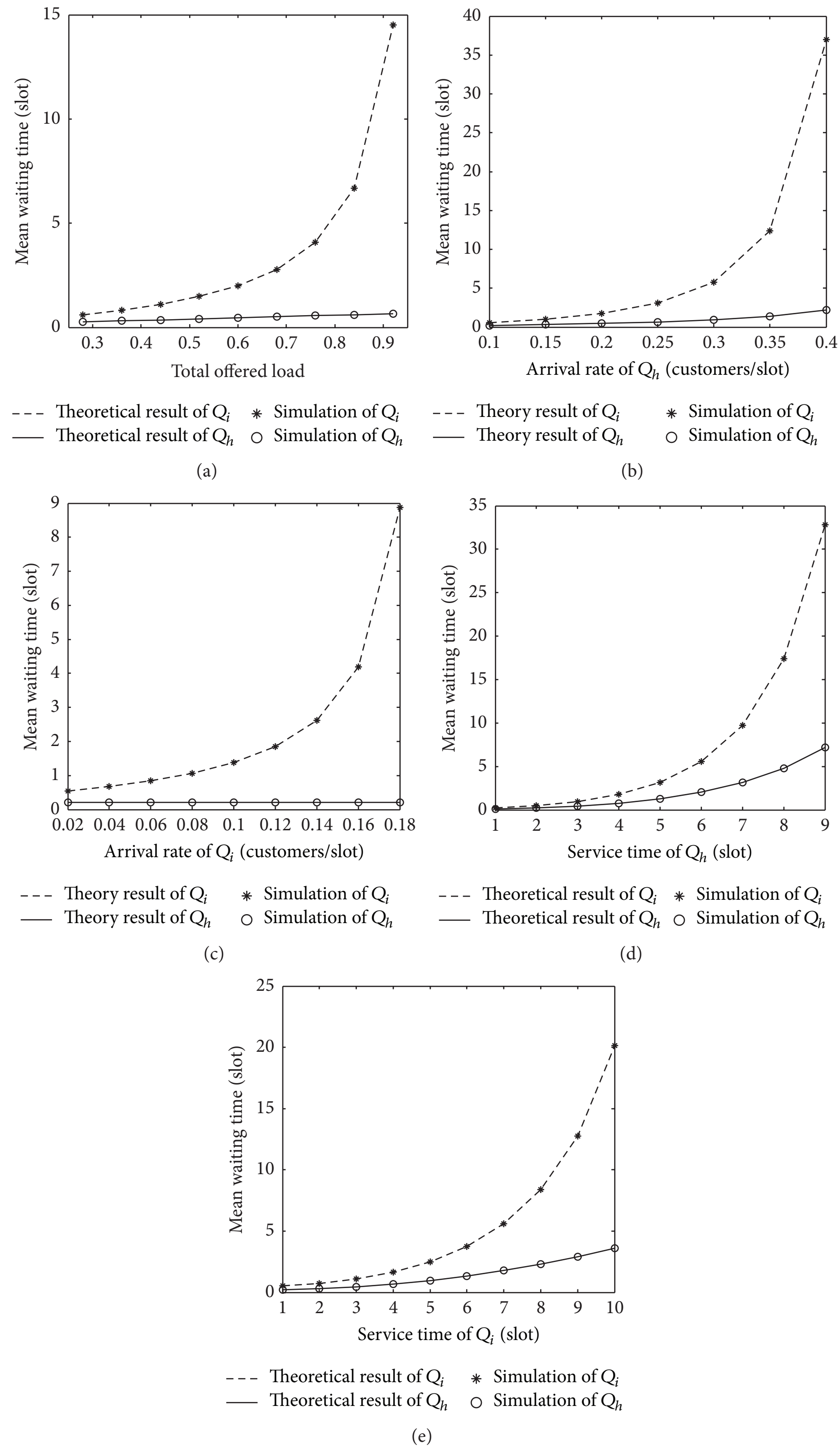


Figure 1: Theoretical and simulation values of $E[W_h]$ and $E[W_i]$ from different values of the load increasing with the increasing of the number of normal queues. (a) is the total offered load increasing with the growth of the number of normal queues. (b) is the total offered load increasing with the growth of the arrival rate of $Q_h$. (c) is the total offered load increasing with the growth of the arrival rate of $Q_i$. (d) is the total offered load increasing with the growth of the service time of $Q_h$. (e) is the total offered load increasing with the growth of the service time of $Q_i$.

Table 1: Test bed used to compare the mean waiting time.

| Parameter | Number of normal queues | Arrival rate | | Service time | | Switch over time |
|---|---|---|---|---|---|---|
| Notation value | $N$ | $\lambda$ | $\lambda_h$ | $\beta$ | $\beta_h$ | $\gamma$ |
| Figure 1(a) | {1 : 1 : 9} | 0.04 | 0.1 | 2 | 2 | |
| Figure 1(b) | 4 | 0.02 | {0.1 : 0.05 : 0.4} | 1 | 2 | — |
| Figure 1(c) | 4 | {0.02 : 0.02 : 0.18} | 0.1 | 1 | 2 | — |
| Figure 1(d) | 4 | 0.02 | 0.1 | 2 | {1 : 1 : 9} | — |
| Figure 1(e) | 4 | 0.02 | 0.1 | {1 : 1 : 10} | 2 | — |
| Figure 2 | 4 | {0.01 : 0.01 : 0.09} | 0.1 | 2 | 2 | 1 |

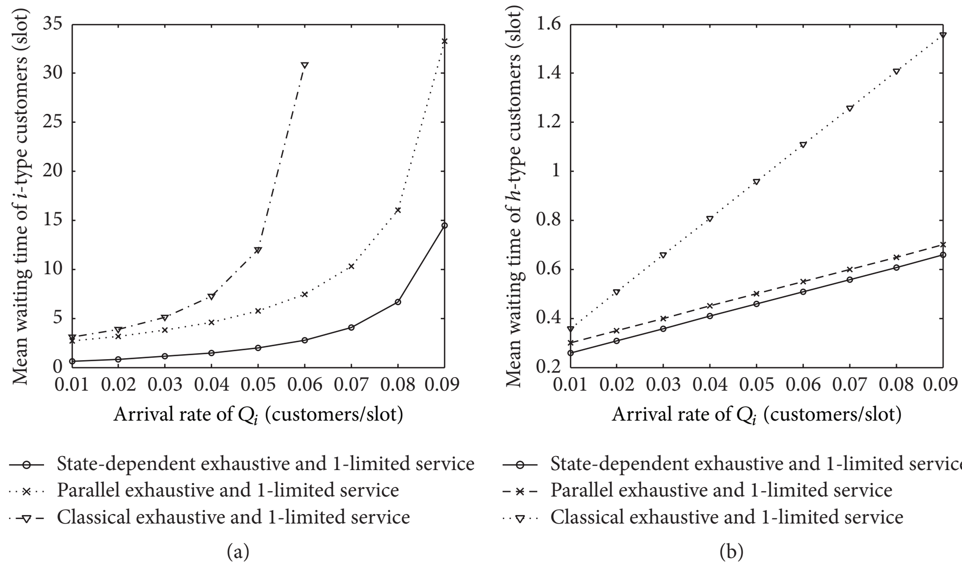


Figure 2: Comparing of mean waiting time among the classical two-level system [6], the parallel two-level system [8], and the state-dependent two-level system. (a) is the theoretical value comparison of $E[W_i]$ with the growth of the arrival rate in $Q_i$. (b) is the theoretical value comparison of $E[W_h]$ with the growth of the arrival rate in $Q_i$.

## Acknowledgments

This work was supported by a Grant from the National Science Foundation of China (nos. 61463051 and 61463054), the National Science Foundation of Yunnan Province (no. 2012FD002), and the Science Foundation of Yunnan Provincial Department (no. 2014Z010).

## References

[1] H. Takagi, "Analysis of polling systems," *Performance Evaluation*, vol. 5, no. 3, p. 206, 1985.

[2] H. Levy and M. Sidi, "Polling systems: applications, modeling, and optimization," *IEEE Transactions on Communications*, vol. 38, no. 10, pp. 1750–1760, 1990.

[3] M. A. A. Boon, R. D. van der Mei, and E. M. M. Winands, "Applications of polling systems," *Surveys in Operations Research and Management Science*, vol. 16, no. 2, pp. 67–82, 2011.

[4] V. M. Vishnevskii and O. M. Semenova, "Mathematical models to study the polling systems," *Automation and Remote Control*, vol. 67, no. 2, pp. 173–220, 2006.

[5] Z.-J. Yang, D.-F. Zhao, H.-W. Ding, and Y.-F. Zhao, "Research on two-class priority based polling system," *Acta Electronica Sinica*, vol. 37, no. 7, pp. 1452–1456, 2009.

[6] Q. Liu, D. Zhao, and D. Zhou, "An analytic model for enhancing IEEE 802.11 point coordination function media access control protocol," *European Transactions on Telecommunications*, vol. 22, no. 6, pp. 332–338, 2011.

[7] Z.-J. Yang, H.-W. Ding, and C.-L. Chen, "Research on E(x) characteristics of two-class polling system of exhaustive-gated service," *Tien Tzu Hsueh Pao/Acta Electronica Sinica*, vol. 42, no. 4, pp. 774–778, 2014 (Chinese).

[8] Z. Guan, D. Zhao, and Y. Zhao, "A discrete time two-level mixed service parallel polling model," *Journal of Electronics*, vol. 29, no. 1-2, pp. 103–110, 2012.

[9] J. P. Dorsman, O. J. Boxma, and R. D. van der Mei, "On two-queue Markovian polling systems with exhaustive service," *Queueing Systems*, vol. 78, no. 4, pp. 287–311, 2014.

[10] M. A. Boon, I. J. Adan, and O. J. Boxma, "A two-queue polling model with two priority levels in the first queue," *Discrete Event Dynamic Systems*, vol. 20, no. 4, pp. 511–536, 2010.

[11] Z. Dongfeng, L. Bihai, and Z. sumin, "Study of a polling systems with limited service," *Journal of Electronics*, vol. 19, no. 1, pp. 44–49, 1997 (Chinese).

[12] O. C. Ibe and X. Cheng, "Stability conditions for multi-queue systems with cyclic service," *IEEE Transactions on Automatic Control*, vol. 33, no. 1, pp. 102–103, 1988.

[13] O. J. Boxma and W. P. Groenendijk, “Pseudoconservation laws in cyclic-service systems,” *Journal of Applied Probability*, vol. 24, no. 4, pp. 949–964, 1987.

[14] D. Zhao and S. Zheng, “Analysis on a polling model with exhaustive service,” *Acta Electronica Sinica*, vol. 22, no. 5, pp. 102–107, 1994 (Chinese).