Probability in the Engineering and Informational Sciences, **22**, 2008, 81–105. Printed in the U.S.A.
DOI: 10.1017/S0269964808000065

$M/G/\infty$ POLLING SYSTEMS WITH RANDOM VISIT TIMES

M. VLASIOU
Georgia Institute of Technology
H. Milton Stewart School of Industrial & Systems Engineering
Atlanta, GA 30332-0205
E-mail: vlasiou@gatech.edu

U. YECHIALI
Department of Statistics and Operations Research
School of Mathematical Sciences
Raymond and Beverly Sackler Faculty of Exact Sciences
Tel Aviv University
Tel Aviv 69978, Israel
E-mail: uriy@post.tau.ac.il

We consider a polling system where a group of an infinite number of servers visits sequentially a set of queues. When visited, each queue is attended for a random time. Arrivals at each queue follow a Poisson process, and the service time of each individual customer is drawn from a general probability distribution function. Thus, each of the queues comprising the system is, in isolation, an $M/G/\infty$-type queue. A job that is not completed during a visit will have a new service-time requirement sampled from the service-time distribution of the corresponding queue. To the best of our knowledge, this article is the first in which an $M/G/\infty$-type polling system is analyzed. For this polling model, we derive the probability generating function and expected value of the queue lengths and the Laplace–Stieltjes transform and expected value of the sojourn time of a customer. Moreover, we identify the policy that maximizes the throughput of the system per cycle and conclude that under the Hamiltonian-tour approach, the optimal visiting order is *independent* of the number of customers present at the various queues at the start of the cycle.

1. INTRODUCTION

A typical polling system consists of a number of queues, attended by a single server in a cyclic fashion. There is a huge body of literature on polling systems that has

© 2008 Cambridge University Press 0269-9648/08 $25.00

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

developed since the late 1950s, when the work of Mack [12] and Mack, Murphy, and Webb [13] concerning a patroling repairman model for the British cotton industry were published. Rather than giving a partial overview of the literature, we refer the interested reader to the following books, surveys, and papers on polling systems: Takagi [17–19], Boxma and Groenendijk [5], Levy and Sidi [11], Yechiali [26], Borst [4], Eliazar and Yechiali [10], and Nakdimon and Yechiali [14].

Polling systems have been used as a central model for the analysis of a wide variety of applications in the areas of repair problems [12,13], telecommunication systems [9], road traffic control [15], computer networks [25], multiple access protocols [3], multiplexing schemes in ISDN [21], satellite systems [1], flexible manufacturing systems [24], and the like.

In many of these applications, as well as in most polling models, it is customary to control the amount of service given to each queue during the server's visit. Common service policies are the *exhaustive*, *gated*, *globally gated*, and *limited* regimes. Under the exhaustive regime, at each visit, the server attends the queue until it becomes completely empty, and only then is the server allowed to move on. Under the gated regime, the only customers served during a visit are the ones who are present when the server enters (polls) the queue, and customers arriving when the queue is attended will be served during the next visit. The globally gated regime, introduced by Boxma, Levy and Yechiali [6], is a modification of the gated one: The only customers served during a visit are those who are present at the beginning of a cycle. Finally, under the k-limited service discipline, only a limited number of jobs (at most k) are served at each server's visit to each queue. These service policies imply that the duration of the visit time in a polled queue is a function of the number of customers present there at a given moment (such as the beginning of the cycle or the moment the server enters the queue).

In this article, we analyze a polling system that differs in two ways from the classical polling model. Rather than considering a *single* server providing service to customers at the various queues, we assume that an *infinite number* of servers are moving as a single group between the queues. Moreover, the service policy we study is *independent* of the queue length. We assume that the group of servers visits each queue for an (possibly random) amount of time that is independent of everything else and that has a distribution that might vary per queue. We further assume that the arrival process of customers to each queue is Poisson and that the service-time distribution for customers in each queue is general. To the best of our knowledge, this article is the first in which an $M/G/\infty$-type polling system is analyzed.

The specific application that raised our attention and led us to this model is in the field of road traffic control. Polling models for road traffic are typically along the lines of the classical polling system; namely they involve a single server rotating around a number of queues. Other assumptions that are typically being made for such models include deterministic service times (i.e., the amount of time that a car needs to pass a traffic light after possibly standing in the queue) and deterministic visit times (i.e., the time the traffic light remains green); see, for example, van der Heijden, van Harten,

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

and Ebben [22]. Although these models provide fairly good approximations of reality, such assumptions fail to capture the variation both in service times and in visit times. Cars do not need the same amount of time to cross a segment of the road: The ones standing ahead in the queue will inevitably need less time and those that arrive while the queue is empty and the traffic light is still green will not even require the additional time incurred by acceleration. Moreover, recent developments in the technology of traffic lights has led to the design of traffic lights that do not turn green unless a queue is formed, and they turn red either when the queue is empty or after a maximum amount of time, which might also vary within a day. As a result, in this article we provide a framework for studying road traffic control under less restrictive assumptions. We propose an infinite-server polling system, which models the behavior of traffic: While the traffic light is green, all cars present in the queue or approaching the traffic light proceed (receive service) and the time they need to complete service is assumed to be a random variable following a general distribution. Furthermore, we assume that the time the traffic light is green (visit time) is random, although our results are directly applicable in the case of deterministic visit times or, more generally, in the case that the visit times follow a discrete distribution taking positive values.

A common approximation to road traffic is to consider the traffic as *fluid* passing through the road. This approximation is fairly accurate when the traffic is relatively high. Mathematically, high traffic can be modeled by assuming that the arrival rate of customers at each of the queues tends to infinity. The study of such a model provides insights at the queue length (and thus the congestion of a junction) under a heavy load. In this article though we do not study the evolution of the system under a heavy load. We assume that the arrival rate at each queue is fixed. This assumption is usually made for the standard polling systems and provides a reasonable approximation to normal traffic conditions.

The rest of the article is organized as follows. Section 2 introduces the model, gives further notation, and describes formally the evolution of the system. In Section 3 we compute recursively the first moment and the probability generating function of the queue length distributions at a polling instant. Later, in Section 4 we derive the mean and the Laplace–Stieltjes transform of the sojourn time of a customer arriving at queue i, and we show how these expressions simplify in the special case where both the service time and the visit time at queue i are exponentially distributed. Based on the results derived up to that point, in Section 5 we give some numerical results. Specifically, we examine numerically the effect of the first two moments of the visit and service times on the sojourn time of an arbitrary customer. These numerical results indicate that there is an optimal value for the mean visit time to the various queues that minimizes the mean sojourn time of an arbitrary customer. In Section 6 we investigate how we can optimize the visit order of the servers at the various queues so that the expected throughput of the system is maximized. It emerges that even when considering *semidynamic* control policies, in which the group of servers plans a new route for each cycle, the optimal visiting order that maximizes the expected throughput per cycle is *fixed* for all cycles. In other

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

words, because of the infinite number of servers, information regarding the queue lengths of all queues at the beginning of a cycle has no effect on the choice of the optimal strategy.

2. MODEL DESCRIPTION AND NOTATION

We consider a polling system with $N \geqslant 2$ infinite-buffer queues attended by a group of ample number of servers that visits the queues in a fixed cyclic fashion. We index the queues by $i = 1, 2, \ldots, N$ in the order of the servers' movement. We will refer to the polling instant of queue i as the moment when the servers enter that queue. When visiting queue i, the group of servers continues working at this queue for V_i units of time and acts there as an $M/G/\infty$ queue. We assume that the visit times are independent and identically distributed (i.i.d.) random variables.

Customers arrive at all queues according to independent homogeneous Poisson processes with rate λ_i for queue i. After completing their service time, customers leave the system. The service time of each individual customer at queue i is denoted by B_i. It is assumed that all service times in one queue are i.i.d. random variables, which are mutually independent of all service times at any other queue. At the end of a visit to queue i, the group of servers moves to queue $i + 1$, incurring a switch-over time D_i and a realization of V_{i+1} is drawn. We assume that $\{D_i\}$ is a sequence of independent random variables. The total switch-over time during a full cycle is $D = \sum_{i=1}^{N} D_i$, and the length of a full cycle is denoted by the random variable C. We assume that all random variables so far are mutually independent.

During the visit time of the group of servers to queue i, a customer present at queue i at the polling instant of that queue will successfully complete his service with probability $p_i(V_i) = \mathbb{P}[B_i \leqslant V_i \mid V_i]$. We assume that if the service of a customer of queue i is not completed during a single visit, then at the next visit, a new service time will be drawn from the service time distribution of B_i for that particular customer.

For a generic random variable Y_i, we denote its first two moments by $\mathbb{E}[Y_i]$ and $\mathbb{E}[Y_i^2]$, respectively. Thus, for example, $\mathbb{E}[V_i]$ is the mean visit time of the servers at queue i. By convention, $\sum_{i \neq j} Y_i = \sum_{\substack{i=1 \\ i \neq j}}^{N} Y_i$, and similarly for the product operator. All further notation will be introduced when it is first used.

2.1. Law of Motion

Let X_i^j, $i, j = 1, 2, \ldots, N$, denote the number of customers in queue j at the moment when queue i is polled and let $A_j(t)$ denote the number of Poisson arrivals to queue j during a time interval of length t. The law of motion describing the evolution of the system when the server moves from queue i to queue $i + 1$ connects X_{i+1}^j to X_i^j and is given by

$$X_{i+1}^j = \begin{cases} X_i^j + A_{j,1}(V_i) + A_{j,2}(D_i), & j \neq i \\ \text{Binom}(X_i^i,\, 1 - p_i(V_i)) + \text{Poisson}(\Lambda_i(V_i)) + A_i(D_i), & j = i, \end{cases} \tag{2.1}$$

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

where for all k, $A_{j,k}(t)$ is an i.i.d. copy of $A_j(t)$, Binom(n, p) is a binomial random variable with parameters n and p, and Poisson$(\Lambda_i(t))$ is a Poisson random variable with rate

$$\Lambda_i(t) = \lambda_i \int_0^t \mathbb{P}[B_i > y]\, dy.$$

Note that from (2.1) we see that for all j, the random variables X_i^j are independent of V_i and D_i, which is evident, considering that the number of customers in a queue at the beginning of a visit does not depend on the length of the upcoming visit time or switch-over time.

The relation for $j \neq i$ is straightforward. The number of customers at queue j at the polling instant of queue $i+1$ equals the number of customers that were there at polling instant of queue i plus all customers that arrived during the visit time of queue i and the switch-over time from queue i to queue $i+1$.

For $j = i$, the relation is more involved. When the servers start polling queue i, they encounter X_i^i customers. After V_i time units, only a binomial number of customers out of the initial X_i^i is still present. The probability that a single customer does not complete his service after V_i time units is $1 - p_i(V_i) = \mathbb{P}[B_i > V_i \mid V_i]$. In addition, there is a stream of new arrivals to queue i. The number of customers present at time t in an $M/G/\infty$ queue (starting with zero customers at time $t = 0$) is Poisson distributed with rate $\Lambda_i(t)$, as is given earlier; see Takács [16]. The last term on the right-hand side of (2.1) incorporates the customers that arrived at the queue during the switch-over time from queue i to queue $i+1$.

We will employ this relation to derive the mean queue length and the probability generating function of all queues at a polling instant.

3. QUEUE LENGTHS AT POLLING INSTANTS

One of the main tools used in the analysis of polling systems is the derivation of a set of multidimensional probability generating functions of the number of jobs present in the various queues at a polling instant of queue i. The common method is to derive the probability generating function of a given queue at some polling instant in terms of the probability generating function of the same queue at the previous polling instant. Then, from the set of N (implicit) dependent equations of the unknown probability generating functions, one can obtain expressions that allow for numerical calculation of the mean queue length at each queue. These equations simplify significantly for several cases of the distribution of the visit times. In this section, we use the law of motion (buffer occupancy), which is given by (2.1) and apply this technique to compute recursively the first moment and the probability generating function of the queue length distributions at a polling instant.

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

3.1. Mean Queue length

From (2.1) we have the following relation for the mean queue length of queue j at two consecutive polling instants:

$$\mathbb{E}[X_{i+1}^j] = \begin{cases} \mathbb{E}[X_i^j] + \lambda_j \mathbb{E}[V_i] + \lambda_j \mathbb{E}[D_i], & j \neq i \\ (1-p_i)\mathbb{E}[X_i^i] + \mathbb{E}[\Lambda_i(V_i)] + \lambda_i \mathbb{E}[D_i], & j = i, \end{cases} \tag{3.1}$$

where $p_i = \mathbb{P}[B_i \leqslant V_i] = \mathbb{E}[\,p_i(V_i)]$. Summing (3.1) over i, we obtain

$$p_j \mathbb{E}[X_j^j] = \lambda_j \sum_{i \neq j} \mathbb{E}[V_i] + \mathbb{E}[\Lambda_j(V_j)] + \lambda_j \mathbb{E}[D]. \tag{3.2}$$

Indeed, in steady state, the mean number of jobs in queue j at a polling instant equals the fraction of jobs $(1 - p_j)\mathbb{E}[X_j^j]$ left behind at the end of the previous visit, plus the mean number of arrivals during the cycle time out of queue j, which is $\lambda_j\big(\sum_{i \neq j} \mathbb{E}[V_i] + \mathbb{E}[D]\big)$, plus the mean number of customers in a $M/G/\infty$ queue at time V_j. The mean queue length of queue j at the polling instant of queue i is easily derived from (3.1), yielding

$$\mathbb{E}[X_i^j] = \mathbb{E}[X_j^j](1-p_j) + \mathbb{E}[\Lambda_j(V_j)] + \lambda_j \sum_{k=j+1}^{i-1} \mathbb{E}[V_k] + \lambda_j \sum_{k=j}^{i-1} \mathbb{E}[D_k]. \tag{3.3}$$

For example, suppose that B_j is exponentially distributed with parameter μ_j. Then

$$\Lambda_j(V_j) = \lambda_j \int_0^{V_j} e^{-\mu_j y}\, dy = \frac{\lambda_j}{\mu_j}(1 - e^{-\mu_j V_j}).$$

Thus, $\mathbb{E}[\Lambda_j(V_j)] = \lambda_j(1 - \mathbb{E}[e^{-\mu_j V_j}])/\mu_j$. So, in particular, if V_j is also exponentially distributed with parameter γ_j, then we have that $\mathbb{E}[\Lambda_j(V_j)] = \lambda_j/(\gamma_j + \mu_j)$, and the mean queue length of each queue can now easily be computed recursively from (3.3).

3.2. Recursive Relation for the Generating Function

Define the generating function of the queue length of all queues at polling instants of queue i as $G_i(\mathbf{z}) = \mathbb{E}\big[\prod_{j=1}^N z_j^{X_i^j}\big]$. Then, from (2.1), we have that

$$G_{i+1}(\mathbf{z}) = \mathbb{E}\left[\prod_{j \neq i} z_j^{X_i^j + A_{j,1}(V_i) + A_{j,2}(D_i)}\, z_i^{\text{Binom}(X_i^i,\, 1-p_i(V_i)) + \text{Poisson}(\Lambda_i(V_i)) + A_i(D_i)}\right]. \tag{3.4}$$

By conditioning on the vector $(X_i^1, \ldots, X_i^N)$, on V_i, and on D_i, (4.1) becomes

$$\begin{aligned} G_{i+1}(\mathbf{z}) = \mathbb{E}\Bigg[\prod_{j \neq i} z_j^{X_i^j} \mathbb{E}\left[\prod_{j \neq i} z_j^{A_{j,1}(V_i)} \mid V_i\right] \mathbb{E}\left[\prod_{j=1}^N z_j^{A_{j,2}(D_i)} \mid D_i\right] \\ \times\, \mathbb{E}\left[z_i^{\text{Binom}(X_i^i,\, 1-p_i(V_i))} \mid X_i^i, V_i\right] \mathbb{E}\left[z_i^{\text{Poisson}(\Lambda_i(V_i))} \mid V_i\right]\Bigg]. \end{aligned} \tag{3.5}$$

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

Since the number of arrivals at any queue during a fixed amount of time is independent of the number of arrivals at any other queue during the same given period, we have that

$$\begin{aligned}\mathbb{E}\left[\prod_{j=1}^{N} z_j^{A_j(D_i)} \mid D_i = x\right] &= \mathbb{E}\left[\prod_{j=1}^{N} z_j^{A_j(x)}\right] \\ &= \prod_{j=1}^{N} \mathbb{E}\left[z_j^{A_j(x)}\right] = \prod_{j=1}^{N} \sum_{n=0}^{\infty} z_j^n \frac{(\lambda_j x)^n}{n!} e^{-\lambda_j x} \\ &= \prod_{j=1}^{N} e^{-\lambda_j x(1-z_j)}.\end{aligned}$$

Therefore, we have

$$\mathbb{E}\left[\prod_{j=1}^{N} z_j^{A_j(D_i)} \mid D_i\right] = e^{-D_i \sum_{j=1}^{N} \lambda_j (1-z_j)}. \tag{3.6}$$

Likewise, we obtain that

$$\mathbb{E}\left[\prod_{j\neq i} z_j^{A_j(V_i)} \mid V_i\right] = e^{-V_i \sum_{j\neq i} \lambda_j (1-z_j)}. \tag{3.7}$$

Moreover,

$$\begin{aligned}\mathbb{E}\left[z_i^{\text{Binom}(X_i^i,\, 1-p_i(V_i))} \mid X_i^i = k,\, V_i = x\right] &= \mathbb{E}\left[z_i^{\text{Binom}(k,1-p_i(x))}\right] \\ &= \sum_{\ell=0}^{k} z_i^{\ell} \binom{k}{\ell} (1-p_i(x))^{\ell} p_i(x)^{k-\ell} \\ &= (p_i(x) + z_i[1-p_i(x)])^k;\end{aligned}$$

in other words,

$$\mathbb{E}\left[z_i^{\text{Binom}(X_i^i,\, 1-p_i(V_i))} \mid X_i^i,\, V_i\right] = (p_i(V_i) + z_i[1-p_i(V_i)])^{X_i^i}. \tag{3.8}$$

For the last term of the right-hand side of (3.5) we have that

$$\begin{aligned}\mathbb{E}\left[z_i^{\text{Poisson}(\Lambda_i(V_i))} \mid V_i = x\right] &= \mathbb{E}\left[z_i^{\text{Poisson}(\Lambda_i(x))}\right] \\ &= \sum_{n=0}^{\infty} z_i^n \frac{(\Lambda_i(x))^n}{n!} e^{-\Lambda_i(x)} = e^{-\Lambda_i(x)(1-z_i)},\end{aligned}$$

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

which yields that

$$\mathbb{E}\left[z_i^{\text{Poisson}(\Lambda_i(V_i))} \mid V_i\right] = e^{-\Lambda_i(V_i)(1-z_i)}. \tag{3.9}$$

Substituting (3.6)–(3.9) into (3.5), we obtain

$$G_{i+1}(\mathbf{z}) = \mathbb{E}\left[\prod_{j\neq i} z_j^{X_i^j} e^{-V_i\sum_{j\neq i}\lambda_j(1-z_j)} e^{-D_i\sum_{j=1}^{N}\lambda_j(1-z_j)} (p_i(V_i)+z_i[1-p_i(V_i)])^{X_i^i} e^{-\Lambda_i(V_i)(1-z_i)}\right]. \tag{3.10}$$

Recall that for all j, the random variables X_i^j are independent of V_i and D_i. Therefore, (3.10) becomes

$$\begin{aligned} G_{i+1}(\mathbf{z}) = {} & \mathbb{E}\left[e^{-D_i\sum_{j=1}^{N}\lambda_j(1-z_j)}\right] \\ & \times \mathbb{E}\left[\prod_{j\neq i} z_j^{X_i^j} e^{-V_i\sum_{j\neq i}\lambda_j(1-z_j)} (p_i(V_i)+z_i[1-p_i(V_i)])^{X_i^i} e^{-\Lambda_i(V_i)(1-z_i)}\right]. \end{aligned} \tag{3.11}$$

Consequently,

$$\begin{aligned} G_{i+1}(\mathbf{z}) = {} & \tilde{D}_i\left(\sum_{j=1}^{N}\lambda_j(1-z_j)\right) \\ & \times \mathbb{E}\left[e^{-V_i\sum_{j\neq i}\lambda_j(1-z_j)} e^{-\Lambda_i(V_i)(1-z_i)} G_i(z_1,z_2,\ldots,z_{i-1},p_i(V_i)\right. \\ & \left. +[1-p_i(V_i)]z_i,z_{i+1},\ldots,z_N)\right], \end{aligned} \tag{3.12}$$

where $\tilde{D}_i(s) = \mathbb{E}[e^{-sD_i}]$ denotes the Laplace–Stieltjes transform of the random variable D_i. Evidently, if V_i follows a discrete distribution, the above expression simplifies significantly. Note that the mean queue length at a polling instant (3.3) can also be obtained by differentiating (3.12).

Remark 1: Applying similar techniques, we can also derive the probability generating function of the number of customers at the end of a visit at queue $i+1$. If we denote by Y_i^j the number of customers in queue $j = 1, \ldots, N$ at the moment when the service at queue $i = 1, \ldots, N$ is completed, then the law of motion describing the evolution of the system is given by

$$Y_{i+1}^j = \begin{cases} Y_i^j + A_j(D_i) + A_j(V_{i+1}), & j \neq i+1 \\ \text{Binom}(Y_i^j + A_j(D_i),\, 1 - p_j(V_j)) + \text{Poisson}(\Lambda_j(V_j)), & j = i+1. \end{cases} \tag{3.13}$$

Also note that the expected value of Y_i^j can be easily computed from (3.3) by observing that for all $j \neq i$, $Y_i^j = X_i^j + A_j(V_i)$, whereas for the ith queue, we have that $Y_i^i = \text{Binom}(X_i^i,\, 1 - p_i(V_i)) + \text{Poisson}(\Lambda_i(V_i))$.

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

4. SOJOURN TIME

Let the sojourn time of a customer at queue i be denoted by S_i. We compute its expected value (and thus, by Little's law, also the mean queue length of queue i at an arbitrary moment), and we derive the Laplace–Stieltjes transform of S_i. As stated earlier, for each queue we assume that if the service of a customer is not completed during a visit, then for the next visit at that queue, a new service time will be resampled for the same customer from the service-time distribution of that queue.

4.1. Mean Sojourn Time

Recall that the cycle time is given by $C = \sum_{i=1}^{N} (V_i + D_i)$. In order to derive the mean sojourn time of a customer arriving at queue i, we will need some further notation. Denote by V_i^{res} the residual visit time of the group of servers at queue i and by $C_{/i}$ the cycle time except the time spent serving queue i (i.e., $C_{/i} = C - V_i$). Similarly, $C_{/i}^{\text{res}}$ represents the residual cycle time excluding the visit time of queue i; that is, $C_{/i}^{\text{res}}$ measures the length of time from a random moment after leaving queue i until the next polling instant of queue i. Furthermore, let $\{C_m\}$ be a family of i.i.d. random variables distributed like C and let N_i be a (shifted) geometric random variable with success probability $p_i = \mathbb{E}[\,p_i(V_i)] = \mathbb{P}[B_i \leqslant V_i]$ (i.e., $\mathbb{P}[N_i = n] = (1 - p_i)^n p_i$), for all integer $n \geqslant 0$. One should observe here that $N_i + 1$ is a stopping time, as it is the first time when the service time of a customer in queue i is less than or equal to the visit time at that queue; that is, $N_i + 1 = \inf\{k : B_{i,k} \leqslant V_{i,k}\}$, where $B_{i,k}$ and $V_{i,k}$ are i.i.d. copies of B_i and V_i respectively. Similarly, a second index is added to a random variable, every time that we explicitly need to indicate that an independent copy is considered. Then the sojourn time of a customer at queue i is given by

$$S_i = \begin{cases} B_{i,0} & \text{(arrival during } V_i \text{ and } B_{i,0} \leqslant V_{i,0}^{\text{res}}) \\ V_{i,0}^{\text{res}} + \sum_{m=1}^{N_i} C_m + C_{/i} + B_{i,N_i+1} & \text{(arrival during } V_{i,0} \text{ and } B_{i,0} > V_{i,0}^{\text{res}}) \\ C_{/i}^{\text{res}} + \sum_{m=1}^{N_i} C_m + B_{i,N_i+1} & \text{(arrival during } C_{/i}). \end{cases} \tag{4.1}$$

Note that the probability of an arrival occurring during the visit time of queue i is $\mathbb{E}[V_i]/\mathbb{E}[C]$ (i.e., the expected visit time of queue i over the expected cycle time) and similarly for the other two events. Therefore, from (4.1) we obtain that the

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

expected sojourn time of a customer of queue i is given by

$$\begin{aligned}\mathbb{E}[S_i] &= \frac{\mathbb{E}[V_i]}{\mathbb{E}[C]}\mathbb{P}[B_i \leqslant V_i^{\text{res}}]\mathbb{E}[B_i \mid B_i \leqslant V_i^{\text{res}}] \\ &\quad + \frac{\mathbb{E}[V_i]}{\mathbb{E}[C]}\mathbb{P}[B_i > V_i^{\text{res}}]\mathbb{E}\left[V_{i,0}^{\text{res}} + \sum_{m=1}^{N_i} C_m + C_{/i} \right. \\ &\qquad\qquad \left. + B_{i,N_i+1} \mid B_{i,0} \geqslant V_{i,0}^{\text{res}}\right] \\ &\quad + \frac{\mathbb{E}[C_{/i}]}{\mathbb{E}[C]}\mathbb{E}\left[C_{/i}^{\text{res}} + \sum_{m=1}^{N_i} C_m + B_{i,N_i+1}\right]. \end{aligned} \tag{4.2}$$

In order to compute the second conditional expectation appearing at the right-hand side of the above equation, we think as follows. For N_i cycles, the service of the customer is not completed during that visit because for every visit $B_i > V_i$, while at the $(N_i + 1)$st visit, the service is completed within that cycle. Therefore, define

$$\bar{C}_m = C_{/i,m} + \min(B_{i,m}, V_{i,m})$$

and observe that

$$\mathbb{E}\left[\sum_{m=1}^{N_i+1} \bar{C}_m\right] = \mathbb{E}\left[\sum_{m=1}^{N_i} C_m + C_{/i} + B_i\right].$$

Thus,

$$\begin{aligned}&\mathbb{E}\left[V_{i,0}^{\text{res}} + \sum_{m=1}^{N_i} C_m + C_{/i} + B_{i,N_i+1} \mid B_{i,0} \geqslant V_{i,0}^{\text{res}}\right] \\ &\quad = \mathbb{E}[V_i^{\text{res}} \mid B_i \geqslant V_i^{\text{res}}] + \mathbb{E}\left[\sum_{m=1}^{N_i+1} \bar{C}_m\right] \\ &\quad = \mathbb{E}[V_i^{\text{res}} \mid B_i \geqslant V_i^{\text{res}}] + \mathbb{E}[N_i + 1]\mathbb{E}[\bar{C}_m] \\ &\quad = \mathbb{E}[V_i^{\text{res}} \mid B_i \geqslant V_i^{\text{res}}] + \mathbb{E}[N_i + 1](\mathbb{E}[C_{/i}] + \mathbb{E}[\min(B_i, V_i)]), \end{aligned} \tag{4.3}$$

where in the second equality we used Wald's equation.

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

For the third conditional expectation appearing at the right-hand side of (4.2), we have that

$$\begin{aligned}
&\mathbb{E}\left[C_{/i}^{\mathrm{res}} + \sum_{m=1}^{N_i} C_m + B_{i,N_i+1}\right] \\
&\quad = \frac{\mathbb{E}[C_{/i}^2]}{2\mathbb{E}[C_{/i}]} + \mathbb{E}\left[\sum_{m=1}^{N_i} (C_{/i,m} + V_{i,m}) + B_{i,N_i+1}\right] \\
&\quad = \frac{\mathbb{E}[C_{/i}^2]}{2\mathbb{E}[C_{/i}]} + \mathbb{E}\left[\sum_{m=1}^{N_i} C_{/i,m}\right] + \mathbb{E}\left[\sum_{m=1}^{N_i+1} \min(B_{i,m}, V_{i,m})\right] \\
&\quad = \frac{\mathbb{E}[C_{/i}^2]}{2\mathbb{E}[C_{/i}]} + \mathbb{E}[N_i]\mathbb{E}[C_{/i}] + \mathbb{E}[N_i + 1]\mathbb{E}[\min(B_i, V_i)]. \qquad \textbf{(4.4)}
\end{aligned}$$

Summarizing the above, we have that

$$\begin{aligned}
\mathbb{E}[S_i] = {} & \frac{\mathbb{E}[V_i]}{\mathbb{E}[C]} \mathbb{P}[B_i \leqslant V_i^{\mathrm{res}}] \mathbb{E}[B_i \mid B_i \leqslant V_i^{\mathrm{res}}] \\
& + \frac{\mathbb{E}[V_i]}{\mathbb{E}[C]} \mathbb{P}[B_i > V_i^{\mathrm{res}}] (\mathbb{E}[V_i^{\mathrm{res}} \mid B_i > V_i^{\mathrm{res}}] \\
& + \mathbb{E}[N_i + 1](\mathbb{E}[C_{/i}] + \mathbb{E}[\min(B_i, V_i)])) \\
& + \frac{\mathbb{E}[C_{/i}]}{\mathbb{E}[C]} \left(\frac{\mathbb{E}[C_{/i}^2]}{2\mathbb{E}[C_{/i}]} + \mathbb{E}[N_i]\mathbb{E}[C_{/i}] + \mathbb{E}[N_i + 1]\mathbb{E}[\min(B_i, V_i)] \right). \qquad \textbf{(4.5)}
\end{aligned}$$

In Section 5 we will illustrate through an example the effect of the first two moments of the visit time and the service time on the mean sojourn time of an arbitrary customer.

4.2. The Laplace–Stieltjes Transform

We now derive the Laplace–Stieltjes transform of the sojourn time of a customer of queue i. We first rewrite (4.1) in terms of the Laplace–Stieltjes transforms of all variables involved (cf. (4.5)), and thus we get that

$$\begin{aligned}
\mathbb{E}[e^{-sS_i}] = {} & \frac{\mathbb{E}[V_i]}{\mathbb{E}[C]} \mathbb{P}\left[B_i \leqslant V_i^{\mathrm{res}}\right] \mathbb{E}\left[e^{-sB_i} \mid B_i \leqslant V_i^{\mathrm{res}}\right] \\
& + \frac{\mathbb{E}[V_i]}{\mathbb{E}[C]} \mathbb{P}\left[B_i > V_i^{\mathrm{res}}\right] \mathbb{E}\left[e^{-sV_i^{\mathrm{res}}} \mid B_i > V_i^{\mathrm{res}}\right] \mathbb{E}\left[e^{-s\sum_{m=1}^{N_i+1} \bar{C}_m}\right] \\
& + \frac{\mathbb{E}[C_{/i}]}{\mathbb{E}[C]} \mathbb{E}\left[e^{-sC_{/i}^{\mathrm{res}}}\right] \mathbb{E}\left[e^{-s\sum_{m=1}^{N_i} C_{/i,m}}\right] \mathbb{E}\left[e^{-s\sum_{m=1}^{N_i+1} \min(B_{i,m}, V_{i,m})}\right]. \qquad \textbf{(4.6)}
\end{aligned}$$

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

We rewrite several of the terms appearing above as follows. The distribution function of V_i^{res} is given by

$$\mathbb{P}[V_i^{\text{res}} \leqslant x] = \frac{1}{\mathbb{E}[V_i]} \int_0^x \mathbb{P}[V_i > y]\, dy,$$

yielding

$$\mathbb{P}[B_i > V_i^{\text{res}}] = \frac{1}{\mathbb{E}[V_i]} \int_0^\infty \mathbb{P}[B_i > x]\mathbb{P}[V_i > x]\, dx. \tag{4.7}$$

Similarly, we have that

$$\mathbb{P}[C_{/i}^{\text{res}} \leqslant x] = \frac{1}{\mathbb{E}[C_{/i}]} \int_0^x \mathbb{P}[C_{/i} > y]\, dy,$$

which implies that

$$\mathbb{E}\left[e^{-sC_{/i}^{\text{res}}}\right] = \frac{1 - \widetilde{C}_{/i}(s)}{s\mathbb{E}[C_{/i}]}, \tag{4.8}$$

where $\tilde{C}_{/i}$ denotes the Laplace–Stieltjes transform of the random variable $C_{/i}$. Moreover,

$$\begin{aligned}
\mathbb{E}\left[e^{-s\sum_{m=1}^{N_i+1} \tilde{C}_m}\right] &= \sum_{n=0}^{\infty} \mathbb{E}\left[e^{-s\sum_{m=1}^{n+1} \tilde{C}_m}\right](1-p_i)^n p_i \\
&= \sum_{n=0}^{\infty} \mathbb{E}[e^{-s\tilde{C}}]^{n+1}(1-p_i)^n p_i \\
&= \frac{p_i \mathbb{E}[e^{-s\tilde{C}}]}{1-(1-p_i)\mathbb{E}[e^{-s\tilde{C}}]},
\end{aligned} \tag{4.9}$$

where $\tilde{C} = C_{/i} + \min(B_i, V_i)$. Likewise, we have that

$$\mathbb{E}\left[e^{-s\sum_{m=1}^{N_i} C_{/i,m}}\right] = \frac{p_i}{1-(1-p_i)\mathbb{E}[e^{-sC_{/i}}]} \tag{4.10}$$

and

$$\mathbb{E}\left[e^{-s\sum_{m=1}^{N_i+1} \min(B_{i,m}, V_{i,m})}\right] = \frac{p_i \mathbb{E}[e^{-s\min(B_i, V_i)}]}{1-(1-p_i)\mathbb{E}[e^{-s\min(B_i, V_i)}]}. \tag{4.11}$$

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

Substituting (4.7)–(4.11) into (4.6), we have that the Laplace–Stieltjes transform of the sojourn time of a customer of queue i is given by

$$\begin{aligned}\mathbb{E}[e^{-sS_i}] &= \frac{1}{\mathbb{E}[C]}\mathbb{E}\left[e^{-sB_i} \mid B_i \leqslant V_i^{\text{res}}\right]\int_0^\infty \mathbb{P}[B_i \le x]\mathbb{P}[V_i > x]\,dx \\ &\quad + \frac{1}{\mathbb{E}[C]}\mathbb{E}\left[e^{-sV_i^{\text{res}}} \mid B_i > V_i^{\text{res}}\right]\frac{p_i\mathbb{E}\left[e^{-s\tilde{C}}\right]}{1-(1-p_i)\mathbb{E}\left[e^{-s\tilde{C}}\right]} \\ &\quad \times \int_0^\infty \mathbb{P}[B_i > x]\mathbb{P}[V_i > x]\,dx \\ &\quad + \frac{1-\widetilde{C}_{/i}(s)}{s\mathbb{E}[C]}\frac{p_i}{1-(1-p_i)\widetilde{C}_{/i}(s)}\frac{p_i\mathbb{E}\left[e^{-s\min(B_i,V_i)}\right]}{1-(1-p_i)\mathbb{E}\left[e^{-s\min(B_i,V_i)}\right]}. \end{aligned} \tag{4.12}$$

Clearly, from the above expression, one can retrieve (4.5) for the mean sojourn time of a customer of queue i.

The transforms appearing in (4.12) might be cumbersome to compute when the service times or the visit times are generally distributed. However, when both B_i and V_i follow a phase-type distribution, all transforms can be computed explicitly since the class of phase-type distributions is closed under finite minima. Phase-type distributions are widely used in computations. The class of phase-type distributions is dense (in the sense of weak convergence) in the class of all distributions on $(0, \infty)$ (cf. [2, Props. 1 and 2]). As an example, we will derive the Laplace–Stieltjes transform of the sojourn time of a customer of queue i, as well as its mean, in the case that both the visit time and the service time at queue i are exponentially distributed.

4.3. A Special Case

Let the service time and the visit time at queue i be exponentially distributed with rates μ_i and γ_i, respectively. Then all terms appearing in (4.5) can be easily computed in terms of μ_i and γ_i. For example,

$$\mathbb{E}[B_i \mid B_i \leqslant V_i^{\text{res}}] = \frac{1}{\gamma_i+\mu_i}$$

and $\mathbb{P}[B_i > V_i^{\text{res}}] = \gamma_i/(\gamma_i+\mu_i)$. Thus, (4.5) becomes

$$\begin{aligned}\mathbb{E}[S_i] &= \frac{1}{\gamma_i\mathbb{E}[C]}\frac{\mu_i}{\gamma_i+\mu_i}\frac{1}{\gamma_i+\mu_i} \\ &\quad + \frac{1}{\gamma_i\mathbb{E}[C]}\frac{\gamma_i}{\gamma_i+\mu_i}\left(\frac{1}{\gamma_i+\mu_i} + \left(\frac{\gamma_i}{\mu_i}+1\right)\left(\mathbb{E}[C_{/i}] + \frac{1}{\gamma_i+\mu_i}\right)\right) \\ &\quad + \frac{\mathbb{E}[C_{/i}]}{\mathbb{E}[C]}\left(\frac{\mathbb{E}[C_{/i}^2]}{2\mathbb{E}[C_{/i}]} + \frac{\gamma_i}{\mu_i}\mathbb{E}[C_{/i}] + \left(\frac{\gamma_i}{\mu_i}+1\right)\frac{1}{\gamma_i+\mu_i}\right)\end{aligned}$$

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

or

$$\mathbb{E}[S_i] = \frac{(\gamma_i \mathbb{E}[C_{/i}] + 1)^2}{\gamma_i \mu_i \mathbb{E}[C]} + \frac{\mathbb{E}[C_{/i}^2]}{2\mathbb{E}[C]}.$$

Similarly, (4.12) reduces to

$$\begin{aligned}\mathbb{E}[e^{-sS_i}] = {} & \frac{1}{\mathbb{E}[C]} \frac{\gamma_i + \mu_i}{\gamma_i + \mu_i + s} \frac{\mu_i}{\gamma_i(\gamma_i + \mu_i)} \\ & + \frac{1}{\mathbb{E}[C]} \frac{\gamma_i + \mu_i}{\gamma_i + \mu_i + s} \frac{\mu_i \mathbb{E}[e^{-s\bar{C}}]}{\gamma_i + \mu_i - \gamma_i \mathbb{E}[e^{-s\bar{C}}]} \frac{1}{\gamma_i + \mu_i} \\ & + \frac{1 - \widetilde{C}_{/i}(s)}{s\mathbb{E}[C]} \frac{\mu_i}{\gamma_i + \mu_i - \gamma_i \widetilde{C}_{/i}(s)} \\ & \times \frac{\mu_i(\gamma_i + \mu_i)/(\gamma_i + \mu_i + s)}{\gamma_i + \mu_i - \gamma_i(\gamma_i + \mu_i)/(\gamma_i + \mu_i + s)}.\end{aligned}$$

Since $\mathbb{E}[e^{-s\bar{C}}] = \tilde{C}_{/i}(s)\mathbb{E}[e^{-s\min(B_i, V_i)}]$, we have that the previous expression reduces to

$$\begin{aligned}\mathbb{E}[e^{-sS_i}] = {} & \frac{1}{\mathbb{E}[C]} \frac{1}{\gamma_i + \mu_i + s} \frac{\mu_i}{\gamma_i} \\ & + \frac{1}{\mathbb{E}[C]} \frac{1}{\gamma_i + \mu_i + s} \frac{\mu_i \widetilde{C}_{/i}(s)}{\gamma_i + \mu_i + s - \gamma_i \widetilde{C}_{/i}(s)} \\ & + \frac{1 - \widetilde{C}_{/i}(s)}{s\mathbb{E}[C]} \frac{\mu_i}{\gamma_i + \mu_i - \gamma_i \widetilde{C}_{/i}(s)} \frac{\mu_i}{\mu_i + s}.\end{aligned}$$

Similar expressions can be easily derived in the case that both the visit times and the service times follow some phase-type distribution, such as Gamma, hyperexponential, or Coxian distributions.

5. NUMERICAL RESULTS

This section is devoted to some numerical results. In particular, we want to examine numerically the effect of the first two moments of the visit and service times on the sojourn time of an arbitrary customer. In all examples, we make the following assumptions. We consider a polling system with two queues. The arrival rate at the first queue is $\lambda_1 = 0.8$, and at the second queue, it is $\lambda_2 = 0.5$. The service time and the visit time at the first queue are exponentially distributed with rates $\mu_1 = 1$ and $\gamma_1 = 1$, respectively. Moreover, the total mean switch-over time is taken to be $\mathbb{E}[D] = 0.5$, and its second moment is assumed to be zero. In all figures that follow, we plot the mean sojourn time of an arbitrary customer, which is estimated by $(\lambda_1 \mathbb{E}[S_1] + \lambda_2 \mathbb{E}[S_2])/(\lambda_1 + \lambda_2)$.

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

In Figures 1 and 2 we investigate the effect of the first two moments of the service time at the second queue on the mean sojourn time of an arbitrary customer. For these plots, the visit time at the second queue is considered to be exponentially distributed with rate $\gamma_2 = 3/2$. For various values of the squared coefficient of variation of the service time at the second queue, which is denoted by $c_{B_2}^2$, we plot in Figure 1 the mean sojourn time of an arbitrary customer versus the mean service time $\mathbb{E}[B_2]$. The squared coefficient of variation of the service time is chosen to be comparable to the squared coefficient of variation of the (exponentially distributed) visit time, which is equal to 1. In Figure 2, we plot the mean sojourn time of an arbitrary customer versus $c_{B_2}^2$ for three values of $\mathbb{E}[B_2]$, which again are chosen to be comparable to $\mathbb{E}[V_2]$.

For each case of $c_{B_2}^2$, we fit a mixed Erlang or hyperexponential distribution to $\mathbb{E}[B_2]$ and $c_{B_2}^2$, depending on whether the squared coefficient of variation is less or greater than 1; see, e.g., Tijms [20]. So, if $1/n \leqslant c_{B_2}^2 \leqslant 1/(n-1)$ for some $n = 2, 3, \ldots$, then the mean and squared coefficient of variation of the mixed-Erlang distribution

$$G(x) = p\left(1 - e^{-\zeta x}\sum_{j=0}^{n-2}\frac{(\zeta x)^j}{j!}\right) + (1-p)\left(1 - e^{-\zeta x}\sum_{j=0}^{n-1}\frac{(\zeta x)^j}{j!}\right), \qquad x \geqslant 0,$$

matches with $\mathbb{E}[B_2]$ and $c_{B_2}^2$, provided the parameters p and ζ are chosen as

$$p = \frac{1}{1+c_{B_2}^2}\left(nc_{B_2}^2 - \sqrt{n(1+c_{B_2}^2) - n^2 c_{B_2}^2}\right), \qquad \zeta = \frac{n-p}{\mathbb{E}[B_2]}.$$

On the other hand, if $c_{B_2}^2 > 1$, then the mean and squared coefficient of variation of the hyperexponential distribution

$$G(x) = p(1 - e^{-\zeta_1 x}) + q(1 - e^{-\zeta_2 x}), \qquad x \geqslant 0,$$

match with $\mathbb{E}[B_2]$ and $c_{B_2}^2$, provided the parameters ζ_1, ζ_2, p, and q are chosen as

$$p = \frac{1}{2}\left(1 + \sqrt{\frac{c_{B_2}^2 - 1}{c_{B_2}^2 + 1}}\right), \qquad q = 1 - p,$$
$$\zeta_1 = \frac{2p}{\mathbb{E}[B_2]}, \qquad \zeta_2 = \frac{2q}{\mathbb{E}[B_2]}.$$

As is evident from the plot in Figure 1, the expected sojourn time of an arbitrary customer increases as the mean service time at the second queue increases. Moreover, the rate at which it increases is almost linear as $c_{B_2}^2$ grows and the effect of the second moment is less pronounced than the effect of $\mathbb{E}[B_2]$.

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

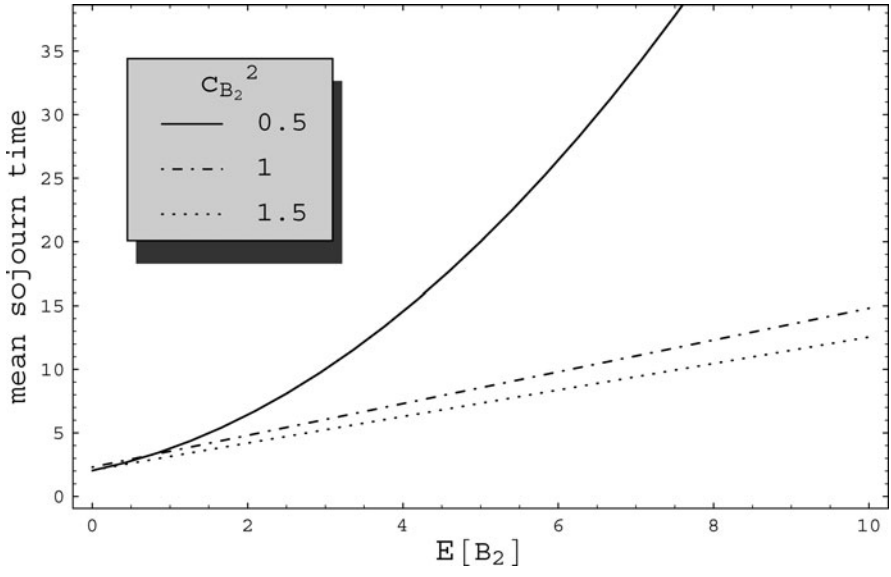


Figure 1. Mean sojourn time of an arbitrary customer against the mean service time $\mathbb{E}[B_2]$.

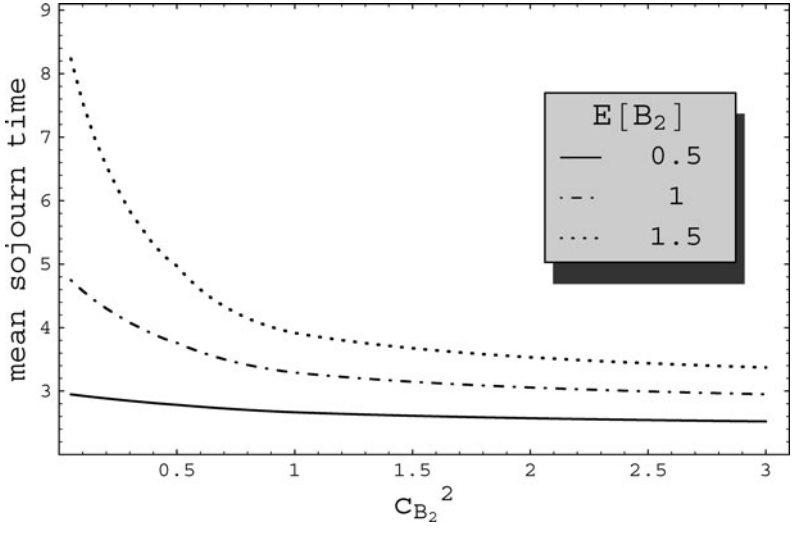


Figure 2. Mean sojourn time of an arbitrary customer against the squared coefficient of variation of the service time B_2.

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

In Figure 2, one observes that the mean sojourn time of an arbitrary customer decreases as the squared coefficient of variation of the service time increases, contrary to the case for the $M/G/1$ queue. This result is due to the fact that the service time of a customer that did not complete his service during one visit time is resampled for the following visit time. Therefore, the larger the variability in the service times, the larger is the probability that during the next visit time, this particular customer will complete his service. Recall that the mean visit time at the second queue is equal to $2/3$ and observe that in the case that $\mathbb{E}[B_2]$ is less than $\mathbb{E}[V_2]$, the effect of the second moment of the service time on the mean sojourn time of an arbitrary customer is almost negligible.

In Figures 3 and 4 we now investigate the effects of the first two moments of the visit time at the second queue on the mean sojourn time of an arbitrary customer. For these plots, we now take the service time at the second queue to be exponentially distributed with rate $\mu_2 = 3/2$. For various values of the squared coefficient of variation of the visit time at the second queue, which is denoted by $c^2_{V_2}$, we plot in Figure 3 the mean sojourn time of an arbitrary customer versus the mean visit time $\mathbb{E}[V_2]$. As earlier, the squared coefficient of variation of the visit time is chosen to be comparable with the squared coefficient of variation of the (exponentially distributed) service time, which is equal to 1. In Figure 4, we plot the mean sojourn time of an arbitrary customer versus $c^2_{V_2}$ for three values of $\mathbb{E}[V_2]$, which again are chosen to be comparable with $\mathbb{E}[B_2]$.

The plot in Figure 3 is interesting. Evidently, when $\mathbb{E}[V_2]$ is significantly smaller than $\mathbb{E}[B_2]$, only a very small number of customers will be served during a visit. As the

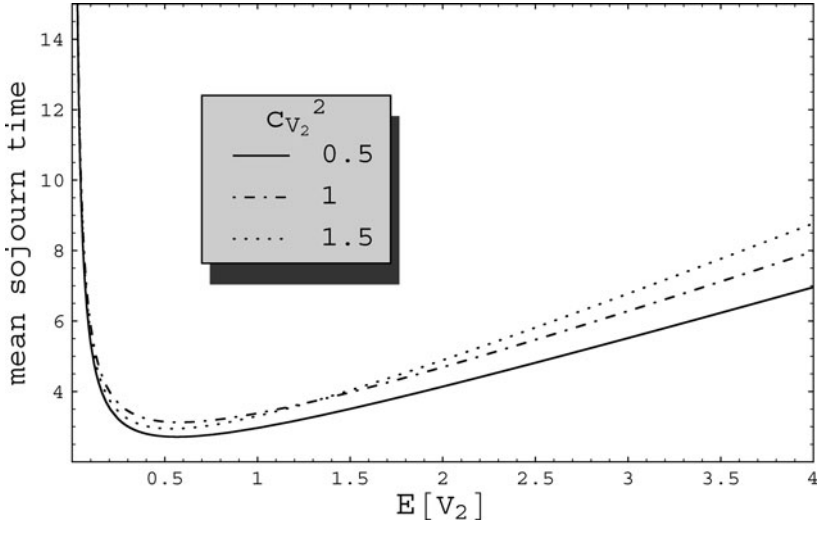


FIGURE 3. Mean sojourn time of an arbitrary customer against the mean visit time $\mathbb{E}[V_2]$.

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

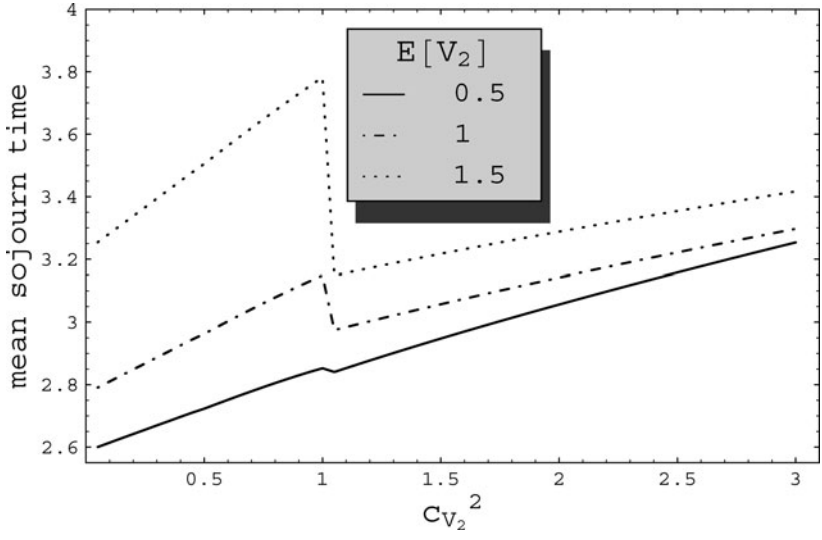


FIGURE 4. Mean sojourn time of an arbitrary customer against the squared coefficient of variation of the visit time V_2.

mean visit time increases, more customers are served during a visit and the mean sojourn time of an arbitrary customer is reduced. However, as the mean visit time continues to increase, this trend is reversed after the mean sojourn time of an arbitrary customer reaches a global minimum. In other words, there is an optimal value for the mean visit time to some queue that minimizes the mean sojourn time of an arbitrary customer; beyond that value, the mean sojourn time of an arbitrary customer increases at an almost linear rate. This indicates that the polling system under consideration can be optimized in expectation by controlling the visit time to each queue. In the following section, we will develop a policy that minimizes the mean sojourn time of an arbitrary customer in the system.

The plot in Figure 1 is also interesting. As is explained earlier, we fit either a mixture of Erlang distributions or a hyperexponential distribution to each pair of the first two moments of the visit time, depending on the value of the squared coefficient of variation. For every value of $c^2_{V_2}$, we obtain a different visit time distribution. Note that the jump in Figure 1 occurs when the distribution we fit to the first two moments of the visit time shifts from a mixture of Erlang distributions to a hyperexponential distribution. This indicates that the shape of the visit time distribution is important; for example, hyperexponential distributions are always unimodal, which is not the case for mixed Erlang distributions. Consequently, the first two moments cannot capture sufficiently the effect of the visit time distribution on the mean sojourn time of an arbitrary customer; one needs to know the exact distribution.

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

6. DYNAMIC CONTROL OF SERVERS' VISITS

A basic question that arises when planning efficient polling systems concerns the order of visits performed by the servers. As it is suggested by Figure 3, the polling system we are considering can be optimized in some way so that the mean sojourn time of an arbitrary customer is minimized. Rather than identifying the value of the minimum mean sojourn time for the cyclic processing (visiting) order considered so far, we first investigate whether there exists a fixed static order that the servers visit the various queues so that the mean (weighted) sojourn time of an arbitrary customer is minimized. As is evident from (4.5), the mean sojourn time of a customer of queue i *does not depend* on the order the queues have been visited and, thus, neither does the weighted sum thereof. Since the mean sojourn time of an arbitrary customer remains unaffected when altering the processing order, it does not constitute a practical performance measure of our system.

An appealing approach that leads to a simple and tractable rule is to develop a semidynamic control scheme. The idea is to dispatch the group of servers to perform Hamiltonian tours, each tour being possibly different from the previous one, depending on the state of the system at the beginning of the tour, so as to optimize some performance measure. An adequate performance measure is the *throughput* of the system, namely the number of customers served per cycle, as the throughput can be measured per cycle, and the sojourn time of a customer spans over a random number of cycles. The goal is to maximize the throughput of the system for each cycle.

Specifically, suppose that at the beginning of a cycle $n = (n_1, n_2, \ldots, n_N)$ is the state of the system, where n_i is the number of jobs waiting in queue i. We will compute the expected value of the number of customers served per cycle under a specific processing order and, consequently, identify the optimal processing order per cycle. The following theorem summarizes the result.

THEOREM 1: *For the Hamiltonian-tour approach, the optimal visiting order is* independent *of the number of customers present at the various queues at the start of the cycle and is given by the index-type rule*

$$\frac{\lambda_i p_i}{\mathbb{E}[V_i] + \mathbb{E}[D_i]}$$

in the sense that the throughput is maximized if and only if the visiting order is arranged according to an increasing *sequence of this rule.*

Before proceeding with the proof, we point out that having more information regarding the system, such as the number of customers present at all queues, has no effect on the optimal strategy; thus, it does not improve the performance of the system. This stands in contrast to many other polling systems where, typically, more information regarding a system leads to decisions that increase the efficiency

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

of the system; see, for example, [7, 8, 26]. This conclusion stems from the fact that no matter what cycle order is used, the number of customers served in a cycle among those initially present will have the same distribution and will have no effect on the number of others served (or the number receiving partial service) during the cycle. In other words, what happens to those initially present at the beginning of a tour is unaffected by what ordering is used.

PROOF: The proof follows from an interchange argument. Consider the processing order $\pi_0 = (1, 2, \ldots, N)$. Denote by θ_i the throughput of queue i under this processing order (i.e., the number of customers of queue i that are served during a cycle) and denote by θ the total throughput of the system (i.e., the sum of all θ_i). Given n (i.e., the state of the system at the beginning of a cycle), we will compute the expected value of θ.

The number of customers served after completing a visit at queue i is equal to the portion of customers that where present at polling instant i and successfully completed their service plus the number of customers that arrived during the visit time of queue i and completed their service within that visit. In other words,

$$\theta_i = \text{Binom}\left(n_i + A_i\left(\sum_{k=1}^{i-1}(V_k + D_k)\right), p_i(V_i)\right) + \big(A_i(V_i) - \text{Poisson}(\Lambda_i(V_i))\big).$$

As a result,

$$\mathbb{E}[\theta_i] = \left(n_i + \lambda_i \sum_{k=1}^{i-1}(\mathbb{E}[V_k] + \mathbb{E}[D_k])\right)p_i + \lambda_i \mathbb{E}[V_i] - \mathbb{E}[\Lambda_i(V_i)],$$

which yields

$$\mathbb{E}[\theta] = c + \sum_{i=1}^{N} \lambda_i p_i \sum_{k=1}^{i-1}(\mathbb{E}[V_k] + \mathbb{E}[D_k]), \tag{6.1}$$

where

$$c = \sum_{i=1}^{N}(n_i p_i + \lambda_i \mathbb{E}[V_i] - \mathbb{E}[\Lambda_i(V_i)]).$$

Observe that the constant c that appears in (6.1) does not depend on π_0, whereas the second term at the right-hand side of (6.1) does.

Consider now the processing order $\pi_1 = (1, 2, \ldots, j-1, j+1, j, j+2, \ldots, N)$, where the visit order of queues j and $j+1$ is interchanged and denote by θ_i' and θ' the throughput of queue i and of the whole system under π_1, respectively. We promptly

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

have that $\mathbb{E}[\theta_i] = \mathbb{E}[\theta'_i]$ for all $i \neq j, j+1$ and that

$$\begin{aligned}\mathbb{E}[\theta'_j] &= \left(n_j + \lambda_j\left(\sum_{k=1}^{j-1}(\mathbb{E}[V_k] + \mathbb{E}[D_k]) + \mathbb{E}[V_{j+1}] + \mathbb{E}[D_{j+1}]\right)\right)p_j \\ &\quad + \lambda_j\mathbb{E}[V_j] - \mathbb{E}[\Lambda_j(V_j)], \\ \mathbb{E}[\theta'_{j+1}] &= \left(n_{j+1} + \lambda_{j+1}\sum_{k=1}^{j-1}(\mathbb{E}[V_k] + \mathbb{E}[D_k])\right)p_{j+1} \\ &\quad + \lambda_{j+1}\mathbb{E}[V_{j+1}] - \mathbb{E}[\Lambda_{j+1}(V_{j+1})].\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}[\theta'] = c &+ \sum_{i\neq j,\, j+1}\lambda_i p_i\sum_{k=1}^{i-1}(\mathbb{E}[V_k] + \mathbb{E}[D_k]) \\ &+ \lambda_j p_j\left(\sum_{k=1}^{j-1}(\mathbb{E}[V_k] + \mathbb{E}[D_k]) + \mathbb{E}[V_{j+1}] + \mathbb{E}[D_{j+1}]\right) \\ &+ \lambda_{j+1}p_{j+1}\sum_{k=1}^{j-1}(\mathbb{E}[V_k] + \mathbb{E}[D_k]).\end{aligned}$$

Therefore, we have that $\mathbb{E}[\theta] \leqslant \mathbb{E}[\theta']$ if and only if

$$\begin{aligned}&\lambda_j p_j\sum_{k=1}^{j-1}(\mathbb{E}[V_k] + \mathbb{E}[D_k]) + \lambda_{j+1}p_{j+1}\sum_{k=1}^{j}(\mathbb{E}[V_k] + \mathbb{E}[D_k]) \\ &\leqslant \lambda_j p_j\sum_{k=1}^{j-1}(\mathbb{E}[V_k] + \mathbb{E}[D_k]) + \lambda_j p_j(\mathbb{E}[V_{j+1}] + \mathbb{E}[D_{j+1}]) \\ &\quad + \lambda_{j+1}p_{j+1}\sum_{k=1}^{j-1}(\mathbb{E}[V_k] + \mathbb{E}[D_k])\end{aligned}$$

or

$$\lambda_{j+1}p_{j+1}(\mathbb{E}[V_j] + \mathbb{E}[D_j]) \leqslant \lambda_j p_j(\mathbb{E}[V_{j+1}] + \mathbb{E}[D_{j+1}]).$$

In other words, we get that the optimal processing order is by visiting the queues according to an *increasing* order of $\lambda_i p_i/(\mathbb{E}[V_i] + \mathbb{E}[D_i])$. ■

Roughly stated, this rule arranges the visit order according to the ratio between new arrivals per unit time that will successfully complete their service (i.e., $\lambda_i p_i$) and the mean duration of a visit there (i.e., $\mathbb{E}[V_i] + \mathbb{E}[D_i]$). It is intuitively clear that if the mean visit and switch time for a queue are relatively long, then one should visit this queue early on. In this way, the number of customers at the other queues during this

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

cycle will also be relatively high, and as a result, the throughput will be increased since all customers are served simultaneously by an infinite number of servers.

This is an extremely simple rule, which can be directly implemented. Moreover, suppose that, for one reason or another, the objective is to minimize the throughput of the system for each cycle. Then the index rule that determines the order of visits to the queues is simply *reversed*: The servers complete a Hamiltonian tour arranged in a decreasing order of $\lambda_i p_i/(\mathbb{E}[V_i] + \mathbb{E}[D_i])$. Observe that, under this strategy, the servers also visit the queues that are empty at the beginning of the cycle.

One expects that the throughput of the system in the long run is improved when these queues are not visited within a cycle; namely it might be more efficient to avoid queues that are empty at the beginning of the cycle in order to allow them to build up.

According to the way the system is designed, even if the servers do not visit a queue that at the beginning of the cycle was empty, the switch time associated with this queue (i.e., the time to switch from this queue to the following one) is still incurred. Therefore, as the number of queues that will not be visited in a cycle grows, the servers spend an increasing amount of time being essentially idle (as they switch between queues).

A possibly more efficient system design is as follows. Rather than envision the group of servers moving from one queue to another, we can think of a central point to which the servers always return after each completion of a visit to a queue. The return time to the central point after visiting queue i is denoted by R_i. The servers depart from that central point and move to the following queue that will be served. The total time from the moment the servers leave the central point until they enter queue i is denoted by E_i. According to this design, the total time to go from queue i to queue j is given by $R_i + E_j$ for any $i \neq j$. The question that arises is whether there exists a semidynamic control of this system. As earlier, it emerges that a Hamiltonian-tour approach leads to a *static* processing order according to an index rule.

THEOREM 2: *For the polling system with a central point, the Hamiltonian-tour approach leads to a fixed optimal visiting order, which is* independent *of the number of customers present at the various queues at the start of the cycle. The throughput of the system for each cycle is maximized if and only if the visiting order is arranged according to an* increasing *sequence of the index-type rule*

$$\frac{\lambda_i p_i}{\mathbb{E}[E_i] + \mathbb{E}[V_i] + \mathbb{E}[R_i]}. \tag{6.2}$$

PROOF: As earlier, let $n = (n_1, n_2, \ldots, n_N)$ be the state of the system at the start of the tour and denote by L the number of *nonempty* queues, $0 < L \leqslant N$, at the beginning of the cycle. The throughput of queue i during a Hamiltonian cycle that visits only the

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

nonempty queues according to the order $\pi_0 = (1,2,\ldots,L)$ is given by

$$\theta_i = \text{Binom}\left(n_i + A_i\left(\sum_{k=1}^{i-1}(E_k + V_k + R_k) + E_i\right), p_i(V_i)\right) + \big(A_i(V_i) - \text{Poisson}(\Lambda_i(V_i))\big).$$

Consequently,

$$\mathbb{E}[\theta_i] = \left(n_i + \lambda_i\left(\sum_{k=1}^{i-1}(\mathbb{E}[E_k] + \mathbb{E}[V_k] + \mathbb{E}[R_k]) + \mathbb{E}[E_i]\right)\right)p_i + \lambda_i\mathbb{E}[V_i] - \mathbb{E}[\Lambda_i(V_i)],$$

which yields

$$\mathbb{E}[\theta] = c' + \sum_{i=1}^{N}\lambda_i p_i \sum_{k=1}^{i-1}(\mathbb{E}[E_k] + \mathbb{E}[V_k] + \mathbb{E}[R_k]), \tag{6.3}$$

where

$$c' = \sum_{i=1}^{N}((n_i + \lambda_i\mathbb{E}[E_i])p_i + \lambda_i\mathbb{E}[V_i] - \mathbb{E}[\Lambda_i(V_i)]).$$

Applying an interchange argument, we have that the optimal processing order is constructed by an increasing sequence of the index rule given by (6.2). ■

As earlier, the optimal tour does not depend on the number of customers present at the beginning of the cycle. This is a direct consequence of the fixed visit times and the underlying $M/G/\infty$ process at each queue.

Index rules appear regularly when optimizing polling systems. Browne and Yechiali [7,8] were the first to obtain dynamic control policies for single-server systems under the exhaustive, gated, or mixed-service regimes. The mechanics of the system are as described here: At the beginning of each cycle, the server decides on a new Hamiltonian tour and visits the channels accordingly. The authors showed that if the objective is to optimize the cycle duration under these policies, then an index-type rule applies, which is similar to the one described here. The main difference is that the index rule that is optimal for these policies depends on the state of the system at the beginning of a cycle, *contrary to the results obtained for the fixed-visit-time policy* studied in this article. The result derived by Browne and Yechiali [7,8] is a surprising result, as the index rule does *not* include the service times at the various channels. It is also surprising that the same index rule holds for both the gated and the exhaustive disciplines although the duration of a cycle starting from the same state is *different* for the two regimes. For a further discussion on other types of index-rule policies, see Yechiali [26], van der Wal and Yechiali [23], and references therein.

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

Acknowledgments

The authors thank Professor Onno Boxma for several insightful remarks. The authors also acknowledge the hospitality and support of EURANDOM while carrying out this research. The research of the first author was also supported by the Aristotle University of Thessaloniki (full scholarship from the legacy of L. Athanasoula). Part of the second author's work was carried out while visiting EURANDOM in his capacity as Beta Chair.

References

1. Altman, E. & Kushner, H.J. (2002). Control of polling in presence of vacations in heavy traffic with applications to satellite and mobile radio systems. *SIAM Journal on Control and Optimization* 41(1): 217–252.
2. Asmussen, S. (2000). Matrix-analytic models and their analysis. *Scandinavian Journal of Statistics: Theory and Applications* 27(2): 193–226.
3. Bernabéu-Aubán, J.M., Ammar, M.H., & Ahamad, M. (1995). Optimizing a generalized polling protocol for resource finding over a multiple access channel. *Computer Networks and ISDN Systems* 27(10): 1429–1445.
4. Borst, S.C. (1996). *Polling systems*. CWI Tract. Vol. 115. Amsterdam: Stichting Mathematisch Centrum Centrum voor Wiskunde en Informatica.
5. Boxma, O.J. & Groenendijk, W.P. (1987). Pseudo conservation laws in cyclic service systems. *Journal of Applied Probability* 24: 949–964.
6. Boxma, O.J., Levy, H., & Yechiali, U. (1992). Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. *Annals of Operations Research* 35(3): 187–208.
7. Browne, S. & Yechiali, U. (1989). Dynamic priority rules for cyclic-type queues. *Advances in Applied Probability* 21(2): 432–450.
8. Browne, S. & Yechiali, U. (1988). Dynamic routing in polling systems. In M. Bonatti (ed.), *Teletraffic science for new cost-effective systems, networks, and services*. Amsterdam: Elsevier Science, pp. 1455–1466.
9. Cooper, R.B. & Murray, G. (1969). Queues served in cyclic order. *Bell System Technical Journal* 48: 675–689.
10. Eliazar, I. & Yechiali, U. (1998). Polling under the randomly timed gated regime. *Communications in Statistics: Stochastic Models* 14(1–2): 79–93.
11. Levy, H. & Sidi, M. (1990). Polling systems: Applications, modeling, and optimization. *IEEE Transactions on Communications* 38(10): 1750–1760.
12. Mack, C. (1957). The efficiency of *N* machines uni-directionally patrolled by one operative when walking time is constant and repair times are variable. *Journal of the Royal Statistical Society, Series B, Methodological* 19: 173–178.
13. Mack, C., Murphy, T., & Webb, N.L. (1957). The efficiency of *N* machines uni-directionally patrolled by one operative when walking time and repair times are constants. *Journal of the Royal Statistical Society, Series B, Methodological* 19: 166–172.
14. Nakdimon, O. & Yechiali, U. (2003). Polling systems with breakdowns and repairs. *European Journal of Operational Research* 149(3): 588–613.
15. Stidham, S., Jr. (1969). Optimal control of a signalized intersection. Part I: Introduction. structure of intersection models. Part II: Determining the optimal switching policies; Part III: Descriptive stochastic models. Technical reports 94, 95, and 96, Department of Operations Research, Cornell University, Ithaca, NY.
16. Takács, L. (1962). *Introduction to the theory of queues*. University Texts in the Mathematical Sciences. New York: Oxford University Press.
17. Takagi, H. (1986). *Analysis of polling systems*. Series In Research Reports and Notes: Computer Systems Series. Cambridge, MA: MIT Press.

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press

18. Takagi, H. (1990). Queueing analysis of polling models: An update. In J.H. Dshalaow (ed.), *Stochastic analysis of computer and communication systems*. Amsterdam: North-Holland, pp. 267–318.
19. Takagi, H. (1997). Queueing analysis of polling models: Progress in 1990–1994. In H. Takagi (ed.), *Frontiers in queueing*, Probabability and Stochastics Series. Boca Raton, FL: CRC, pp. 119–146.
20. Tijms, H.C. (2003). *A first course in stochastic models*. Chichester: Wiley.
21. Twu, D.-C. & Chen, K.-C. (1996). A novel MAC protocol for broadband communication over CATV-based MANs. *Computer Communications* 19(11): 888–900.
22. van der Heijden, M., van Harten, A., & Ebben, M. (2001). Waiting times at periodically switched one-way traffic lanes: a periodic, two-queue polling system with random setup times. *Probability in the Engineering and Informational Sciences* 15(4): 495–517.
23. van der Wal, J. & Yechiali, U. (2003). Dynamic visit-order rules for batch-service polling. *Probability in the Engineering and Informational Sciences* 17(3): 351–367.
24. Van Vuuren, M. & Winands, E.M.M. (2006). Iterative approximation of *k*-limited polling systems. Technical report 2006-06, Eindhoven University of Technology. Available from http://www.win.tue.nl/math/bs/spor/.
25. Wang, Y.T. & Morris, R.J.T. (1985). Load sharing in distributed systems. *IEEE Transactions on Computers* C-34(3): 204–217.
26. Yechiali, U. (1993). Analysis and control of polling systems. In L. Donatiello & R. Nelson (eds.), *Performance evaluation of computer and communication systems*. Lecture Notes in Computer Science, Vol. 729. Berlin: Springer-Verlag, pp. 630–650.

https://doi.org/10.1017/S0269964808000065 Published online by Cambridge University Press