COMPUTING DISTRIBUTIONS AND MOMENTS IN POLLING MODELS BY NUMERICAL TRANSFORM INVERSION

by

Gagan L. Choudhury[1] *and Ward Whitt*[2]

AT&T Bell Laboratories

September 23, 1994

Revised: March 1, 1995

[1] AT&T Bell Laboratories, Room 1L-238, Holmdel, New Jersey 07733-3030; email: gagan@buckaroo.att.com

[2] AT&T Bell Laboratories, 2C-178, Murray Hill, New Jersey 07974-0636; email: wow@research.att.com

Abstract

We show that probability distributions and moments of performance measures in many polling models can be effectively computed by numerically inverting transforms (generating functions and Laplace transforms). We develop new efficient iterative algorithms for computing the transform values and then use our recently developed variant of the Fourier-series method to perform the inversion. We also show how to use this approach to compute moments and asymptotic parameters of the distributions. We compute a two-term asymptotic expansion of the tail probabilities, which turns out to be remarkably accurate for small tail probabilities. The tail probabilities are especially helpful in understanding the performance of different polling disciplines. For instance, it is known that the exhaustive discipline produces smaller mean steady-state waiting times than the gated discipline, but we show that the reverse tends to be true for small tail probabilities. The algorithms apply to describe the transient behavior of stationary or nonstationary models as well as the steady-state behavior of stationary models. We demonstrate effectiveness by analyzing the computational complexity and by doing several numerical examples for the gated and exhaustive service disciplines, with both zero and non-zero switchover times. We also show that our approach applies to other polling models. Our main focus is on computing exact tail probabilities and asymptotic approximations to them, which seems not to have been done before. However, even for mean waiting times, our algorithm is faster than previous algorithms for large models. The computational complexity of our algorithm is $O(N^{\alpha})$ for computing performance measures at one queue and $O(N^{1+\alpha})$ for computing performance measures at all queues, where N is the number of queues and α is typically between 0.6 and 0.8.

1. Introduction

Over the last thirty years a substantial literature on polling models has evolved, as can be seen from Coffman and Hofri [20], Grillo [31] and Takagi [43,44]. Polling models have generated great interest because they have many applications to the performance analysis of computer and telecommunications systems. These applications include data transfer from terminals on multi-drop lines to a central computer, token passing schemes in local and wide area networks, job scheduling in telephone switching systems and scheduling moving arms in secondary storage devices.

Multidimensional transforms of performance measures of interest have been derived for several basic polling models with Poisson arrivals, general service-time and general switchover-time distributions. These transform expressions have been successfully exploited to derive means and sometimes second moments, but we are unaware of them being used to calculate the distributions themselves, higher moments or asymptotic parameters. (However, distributions have been calculated by other methods by Blanc [12] and Leung [35], but these methods are computationally intensive so that they are feasible for only a few queues. Also approximate distributions have been calculated by Federgruen and Katalan [25].) One reason that these other characteristics have not previously been derived is that methods for transform inversion are not especially well known. Another reason is that usually the transforms are not available directly; instead the transform at one polling instant is expressed in terms of the transform at another polling instant. However, as early as 1972, Eisenberg [23] suggested that this recursive relation among the transforms could be exploited to compute the steady-state transform values.

In this paper we show that the approach suggested by Eisenberg [23] can indeed be carried out. We show that these polling transforms often can be quite easily computed and inverted numerically to calculate distributions, all moments and asymptotic parameters. We develop a new efficient recursive algorithm for computing transform values. Our operation count for computing moments and distributions is $O(N^{\alpha})$ for one queue and is $O(N^{1+\alpha})$ for all queues of an N-queue system, where α is typically in the range 0.6 to 0.8. It appears that our algorithm is faster than other available algorithms for mean waiting time and queue lengths. To demonstrate, in Section 6 we consider an asymmetric 1000-queue system and compute the mean waiting time in less than 5 seconds, and several moments and tail probability values in a few minutes, using a SUN SPARC-2 workstation.

We demonstrate the power of numerical inversion by specifically considering the gated and exhaustive service models with unlimited waiting space, but we also show how numerical inversion can be applied to many other polling models. In fact, from Resing [39], it is easy to see that our approach extends in a straightforward way to the rich class of polling models in which the vector queue-length process at polling instants of a fixed queue is a multitype branching process with immigration as in Athreya and Ney [10]. However, the generating function recursion needed by our approach does not readily extend to models outside of this class, such as ones with customer-limited or time-limited service or probabilistic server routing; see [14,33]. The discrete-Fourier-transform approach in Leung [35] and the power-series approach in Blanc [12] do apply to these more difficult models, but the computational complexity of their approach is much higher than ours.

To perform the numerical inversion, we use the Fourier-series method in Abate and Whitt [5], Choudhury, Lucantoni and Whitt [19], Choudhury and Lucantoni [17], and Abate, Choudhury, Lucantoni and Whitt [1]. These references contain algorithms for one-dimensional probability distributions, multi-dimensional functions, higher moments and asymptotic parameters. These algorithms effectively control numerical errors and are very fast. For example, simple one-dimensional transforms are usually inverted in a fraction of a second on a standard PC or workstation with 8–12 decimal places accuracy. For the sake of completeness, we include the relevant inversion formulas here in an Appendix.

Computing full distributions instead of only means is important because in many performance analysis applications we really want to know high percentiles, such as the 95^{th} or 99^{th}. In emerging high-speed communication networks there is even great interest in very small tail probabilities, such as 10^{-9}, in order to provide an appropriate quality of service. Hence, the ability to compute full distributions should significantly increase the usefulness of polling models. Moreover, simulation is usually quite effective for computing means, but simulation has difficulty computing small tail probabilities. The advantage of analytical models over simulation really shows up in computing small tail probabilities (although methods such as importance sampling can improve simulation estimates of small tail probabilities).

Note that we compute all moments by numerical inversion, so that our method is the same for the hundredth moment as it is for the first moment. By contrast, previous results for higher moments, such as by Ferguson [26] and Konheim, Levy and Srinivasan [34], have been via analytical differentiation of the transform, which leads to cumbersome expressions. So far, analytical differentiation has only provided results for the first two moments, but we can easily compute even the hundredth moment. Moreover, these higher moments are very useful for performing numerical asymptotic analysis of the tail probabilities, as we demonstrate.

Unlike for the first-in first-out (FIFO) discipline [4,6,18,19,29,38], for polling models computing transient performance measures tends to be easier than computing steady-state performance measures, because we compute steady-state transforms by computing transient transforms for a suitably large number of queue visits. Here we compute *both* transient and steady-state performance measures for polling models. The transient performance measures are computed only at polling instants, whereas the steady-state performance measures are computed at both polling instants and arbitrary times. In the transient case, the model need not be stationary; i.e., the arrival rate, service-time distribution and switchover-time distribution can change at polling instants. The nonstationary model allows us to study the effect of a sudden overload. We are unaware of any previous work on time-varying polling models.

We also point out that qualitative properties of tail probabilities for polling models are different from qualitative properties for means. To understand tail probabilities such as for a steady-state waiting time *W*, it is useful to exploit asymptotics such as

$$P(W > x) \sim \alpha x^{\beta} e^{-\eta x} \quad \text{as } x \to \infty , \tag{1.1}$$

where $f(x) \sim g(x)$ means that $f(x)/g(x) \to 1$ as $x \to \infty$. We studied asymptotics of the form (1.1) for the GI/G/1 queue with the first-in first-out (FIFO) service discipline in Abate, Choudhury and Whitt [2,3] and for the M/G/1 queue with the last-in first-out (LIFO) service discipline in Abate, Choudhury and Whitt [4]. Here we study the asymptotics (1.1) for polling models by *numerically* deriving the parameters α, β and η, using the moment approach in Choudhury and Lucantoni [17] and Abate, Choudhury, Lucantoni and Whitt [1]. We show that the asymptote (1.1) is often quite accurate and that it reveals important properties of the polling disciplines. Unlike FIFO, but like LIFO, our experience indicates that $\beta \neq 0$ for M/G/1 polling models. Indeed, the pure-exponential asymptotics with $\beta = 0$ typically holding with the FIFO discipline seems to be the exception rather than the rule for other disciplines. When $\beta \neq 0$, we also do two-term asymptotic expansions, which significantly improve the accuracy.

To indicate the kind of insights we can obtain, consider a symmetric system (gated or exhaustive) with *N* queues, mean switchover time *r* per queue and a fixed total server utilization. First, for $r = 0$, if we increase *N*, then the mean waiting time remains constant, but often the asymptotic parameter η in (1.1) decreases, and the tail probabilities increase. Next, for a fixed *N*, if we increase *r*, then the mean waiting time increases significantly, but the asymptotic decay rate η decreases (and the tail probabilities increase) only slightly. Finally, for $r > 0$ and $N > 1$, the mean waiting time is bigger for the gated discipline than for the exhaustive discipline. However, the reverse is true for higher-order moments and for small tail probabilities.

Here is how the rest of this paper is organized. In Section 2 we develop new efficient algorithms for computing transient and steady-state transforms for the gated service discipline. We also show how the computational complexity (in the case of mean waiting times) compares to that of previous algorithms. In Section 3 we briefly describe several other important polling models for which our approach is applicable. In Section 4 we show how to extend our approach to zero switchover times. There we derive a new iterative algorithm for computing the steady-state transform values. In Section 5 we show how to do asymptotic analysis using the moment computation algorithm, drawing on [1,17]. Finally, in Section 6 we provide several numerical examples to illustrate the power of the transform inversion approach for computing moments and distributions and for doing asymptotic analysis. We give a brief account of the numerical transform inversion algorithm in an Appendix.

2. The Gated Service Model

There are N queues indexed by i, $0 \leq i \leq N-1$. There is a single server that moves successively from queue i to queue $i+1 \mod N$, i.e., from i to $i+1$ for $i \leq N-2$ and from $N-1$ to 0. We assume that a *gated policy* is in use, so that the server serves all customers at a queue that it finds upon arrival there, but no new customers that arrive after the server.

Customers arrive at queue i according to a Poisson process with the rate λ_i. Each customer at queue i requires a service time that has *Laplace-Stieltjes transform* (LST) $\hat{B}_i(s) \equiv \int_0^\infty e^{-st} dB_i(t)$. The server has a switchover (reply) time to go from queue i to queue $i+1 \mod N$ with LST $\hat{R}_i(s)$. The customer service times, server switchover times and customer arrival processes are all mutually independent.

2.1 Transient Analysis

We assume that the system starts at time 0 with some initial distribution of customers at the various queues and with the server about to begin service at some queue. Let $L_{i,m} \equiv (L_{i,m,0}, \ldots, L_{i,m,N-1})$ represent the vector of queue lengths at the instant the server is about to begin service at queue i after having visited m queues. For analysis, it is significant that $\{L_{i,m} : m \geq 0\}$ is a discrete-time Markov chain. Let $p_{i,m} \equiv (p_{i,m}(k_0, \ldots, k_{N-1}))$ be the N-dimensional probability mass function (pmf) of $L_{i,m}$ and let $\hat{p}_{i,m}$ be its N-dimensional generating function, i.e.,

$$\hat{p}_{i,m}(\mathbf{z}) = E \prod_{j=0}^{N-1} z_j^{L_{i,m,j}} = \sum_{k_0 = 0}^{\infty} \cdots \sum_{k_{N-1} = 0}^{\infty} \prod_{j=0}^{N-1} z_j^{k_j} \; p_{i,m}(k_0, \ldots, k_{N-1}) \tag{2.1}$$

for an N-dimensional vector of complex variables $\mathbf{z} = (z_0, \ldots, z_{N-1})$. It is known that we can express $\hat{p}_{i,m}(\mathbf{z})$ in terms of $\hat{p}_{i-1,m-1}(\mathbf{z})$; e.g., p. 105 of Takagi [43]. The following is the transient analog of the steady-state equations given in (5.231) of [43], which hold by the same argument. For the statement, let w be the wrap-around function

$$w(k) = k \bmod N . \tag{2.2}$$

Proposition 2.1. *The generating function of the pmf $p_{i,m}$ is*

$$\begin{aligned} \hat{p}_{i,m}(\mathbf{z}) = {} & \hat{R}_{w(i-1)}\Big(\sum_{j=0}^{N-1} (\lambda_j - \lambda_j z_j)\Big) \\ & \times \hat{p}_{w(i-1),m-1}\Big(z_0, \ldots, z_{i-2}, \hat{B}_{w(i-1)}\Big(\sum_{j=0}^{N-1} (\lambda_j - \lambda_j z_j)\Big), z_i, \ldots, z_{N-1}\Big) . \end{aligned} \tag{2.3}$$

Using (2.3) recursively m times, we arrive at the following proposition. (The proof is by mathematical induction.)

Proposition 2.2. *The generating function in (2.3) can be expressed as*

$$\hat{p}_{i,m}(\mathbf{z}) = \prod_{k=1}^{m} \hat{R}_{w(i-k)} \left[\sum_{j=0}^{N-1} (\lambda_j - \lambda_j z_j^{(k-1)}) \right] \hat{p}_{w(i-m),0}(\mathbf{z}^{(m)}) \tag{2.4}$$

where

$$z_l^{(k)} = \begin{cases} z_l^{(k-1)} & \text{for} \quad l \neq w(i-k) \\ \hat{B}_l \left(\sum_{j=0}^{N-1} (\lambda_j - \lambda_j z_j^{(k-1)} \right) & \text{for} \quad l = w(i-k) \end{cases} \tag{2.5}$$

and $\mathbf{z}^0 = \mathbf{z}$.

The quantity $\hat{p}_{w(i-m),0}(\mathbf{z}^{(m)})$ in (2.4) is the transform of the initial distribution and $w(i-m)$ is the index of the first queue polled. If the system is empty at $t = 0$, then $\hat{p}_{w(i-m),0}(\mathbf{z}) = 1$ for all $\mathbf{z}$. If $\mathbf{k}_0 = (k_{0,0}, k_{0,1}, \ldots k_{0,N-1})$ represents a deterministic vector of initial queue lengths, then

$$\hat{p}_{w(i-m),0}(\mathbf{z}) = \prod_{j=0}^{N-1} z_j^{k_{0,j}} \ . \tag{2.6}$$

To compute each term of the m-fold product in (2.4), the operation count is $O(N)$ due to the presence of the two N-fold sums in (2.4) and (2.5). We show below that the operation count may be reduced to $O(1)$ by introducing the auxiliary variable

$$y^{(k)} = \sum_{j=0}^{N-1} (\lambda_j - \lambda_j z_j^{(k)}) \ . \tag{2.7}$$

We replace (2.4) and (2.5) by the following new set of recursions:

$$\hat{p}_{i,m}(\mathbf{z}) = \prod_{k=1}^{m} \hat{R}_{w(i-k)}(y^{(k-1)}) \hat{p}_{w(i-m),0}(\mathbf{z}^{(m)}) \ , \tag{2.8}$$

$$z_l^{(k)} = \begin{cases} z_l^{(k-1)} & \text{for } l \neq w(i-k) \\ \hat{B}_l(y^{(k-1)}) & \text{for } l = w(i-k) \ , \end{cases} \tag{2.9}$$

$$y^{(k)} = y^{(k-1)} + \lambda_{w(i-k)} (z_{w(i-k)}^{(k-1)} - z_{w(i-k)}^{(k)}) \ , \tag{2.10}$$

with initial conditions $\mathbf{z}^{(0)} = \mathbf{z}$ and

$$y^{(0)} = \sum_{j=0}^{N-1} (\lambda_j - \lambda_j z_j) \ . \tag{2.11}$$

Note that the computation of (2.10) and $\hat{R}_{w(i-k)} (y^{(k-1)})$ in (2.8) require $O(1)$ operations, and so does the computation of (2.9) since only one z variable changes. The total operation count in computing $\hat{p}_{i,m} (\mathbf{z})$ is thus $O(m)$ and the total storage requirement is $O(N)$ since we have to store the vector $\mathbf{z}^{(k)}$ and the scalar $y^{(k)}$ in each step. By contrast, the total operation count for the more straightforward recursions in (2.4) and (2.5) is $O(Nm)$ and storage requirement is $O(N)$.

Once we get $\hat{p}_{i,m} (\mathbf{z})$, we can compute distributions and moments by numerical inversion. In particular, by inverting $\hat{p}_{i,m} (1,1, \ldots , z_j ,\ldots 1)$ we get the marginal distribution at queue j. This is one-dimensional inversion and is pretty fast. So we can get marginal distributions at all queues even for pretty large N. (See Section 2.3 on the computational complexity). We can also compute joint distributions of, say, queues 0, 1 and 2 by inverting $\hat{p}_{i,m} (z_0 ,z_1 ,z_2 ,1,1,\ldots 1)$. However, for joint distributions the computational complexity grows exponentially with the number of dimensions [19].

For transient analysis we can even allow the system to be time-varying. In particular we can allow each of the service time LSTs $\hat{B}_l (\cdot)$, switchover time LSTs $\hat{R}_l (\cdot)$ and arrival rates λ_l to change as a function of m. We illustrate this via numerical examples in Section 6.

We can also get (and numerically invert) transforms related to $\hat{p}_{i,m} (\mathbf{z})$. For this purpose, let $\tilde{p}_{i,m} (z) = \hat{p}_{i,m} (\tilde{\mathbf{z}})$ where $\tilde{z}_k = 1$ for $k \neq i$ and $\tilde{z}_i = z$. Let $\hat{S}_{i,m} (s)$ be the LST of the time spent by the server at queue i after having visited m queues. By (5.39a) of [43],

$$\hat{S}_{i,m} (s) = \tilde{p}_{i,m} (\hat{B}_i (s)) \ . \tag{2.12}$$

Let $\hat{C}_{i,m} (s)$ represent the LST of a cycle time that ends just before a visit to queue i and after m

queue visits overall ($m > N$). Based on (5.39a) of [43],

$$\hat{C}_{i,m}(s) = \tilde{p}_{i,m}(1 - s/\lambda_i) \ . \tag{2.13}$$

2.2 Steady-state Analysis

The transient transforms of Section 2.1 converge to proper steady-state values as $m \to \infty$ if the stability condition for this model is satisfied. The stability condition is well known to be $\sum_{j=0}^{N-1} \rho_j < 1$, where $\rho_j \equiv \lambda_j b_j$ is the offered load at queue j and b_j is the mean service time; e.g., see [8,30,39]. The steady-state limit, if it exists, has to be independent of the transform of the initial distribution $\hat{p}_{w(i-m),0}(\mathbf{z}^{(m)})$. This is possible only if $\mathbf{z}^{(m)}$ approaches a vector of all 1's. We can apply Proposition 2.2 to obtain an explicit representation for the steady-state transform $\hat{p}_i(\mathbf{z}) \equiv \lim_{m \to \infty} \hat{p}_{i,m}(\mathbf{z})$.

Proposition 2.3. *If $L_{i,m}$ converges to a proper limit L_i as $m \to \infty$, then its probability generating function is*

$$\hat{p}_i(\mathbf{z}) = E \prod_{j=0}^{N-1} z_j^{L_{i,j}} = \prod_{k=1}^{\infty} \hat{R}_{w(i-k)}(y^{(k-1)}) \ , \tag{2.14}$$

$$z_l^{(k)} = \begin{cases} z_l^{(k-1)} & \text{for } l \neq w(i-k) \\ \hat{B}_l(y^{(k-1)}) & \text{for } l = w(i-k) \ , \end{cases} \tag{2.15}$$

$$y^{(k)} = y^{(k-1)} + \lambda_{w(i-k)}(z_{w(i-k)}^{(k-1)} - z_{w(i-k)}^{(k)}) \ .$$

with initial conditions $\mathbf{z}^{(0)} = \mathbf{z}$ and $y^{(0)} = \sum_{j=0}^{N-1} (\lambda_j - \lambda_j z_j)$. Furthermore,

$$\lim_{k \to \infty} z^{(k)} = (1, 1, \ldots, 1) \tag{2.17}$$

$$\lim_{k \to \infty} y^{(k)} = 0 \tag{2.18}$$

$$\lim_{k \to \infty} \hat{R}_{w(i-k)}(y^{(k-1)}) = 1 \ . \tag{2.19}$$

Explicit product-form expressions for transient and steady-state generating functions, such as in Propositions 2.2 and 2.3, are not routinely reported in the polling literature. One is given for the globally gated service discipline in [9], however. Following our approach and [39], it is possible to get explicit product-form expressions for all ‘‘multitype branching process’’ type polling models. In general, the special speeded-up recursion using the auxiliary variable $z_l^{(k)}$ that we get for gated systems may not extend to all such models, but such speed-ups are possible for many special cases (e.g., exhaustive systems).

To numerically compute $\hat{p}_i(\mathbf{z})$, we truncate the infinite product in (2.14) at $k = I$ using the stopping criterion

$$|\hat{R}_{w(i-I)}(y^{(I-1)}) - 1| < \varepsilon \tag{2.20}$$

for some suitably small ε. Other stopping criteria, based on (2.17) or (2.18) may also be used. Note that in general I will depend on the particular $\mathbf{z}$ value at which the transform is needed. As in the transient case, the computational complexity in computing each term in the product on the right hand side of (2.14) is $O(1)$ and the overall computational complexity in computing one transform value is $O(I)$.

In this paper we do all computations using double precision arithmetic and typically aim for about 8-to-10 place accuracy in the final answer. For this purpose, setting $\varepsilon = 10^{-12}$ or 10^{-13} in (2.20) is usually adequate.

Under the stability condition, the other transient transforms in (2.12) and (2.13) also approach their steady-state values as given below:

$$\hat{S}_i(s) = \tilde{p}_i(\hat{B}_i(s)) \tag{2.21}$$

and

$$\hat{C}_i(s) = \tilde{p}_i(1 - s/\lambda_i) , \tag{2.22}$$

where as before $\tilde{p}_i(z) = \hat{p}_i(\tilde{\mathbf{z}})$ with $\tilde{z}_k = 1$ for $k \neq i$ and $\tilde{z}_i = z$.

We also have some steady-state transforms which do not have transient counterparts. As on p. 109 of Takagi [43], the steady-state number of customers at queue i at an arbitrary time is

$$\hat{Q}_i(\mathbf{z}) = \frac{(1 - \sum_{k=0}^{N-1} \rho_k)}{\lambda_i \sum_{k=0}^{N-1} r_k} \cdot \frac{\hat{B}_i(\lambda_i - \lambda_i z)}{z - \hat{B}_i(\lambda_i - \lambda_i z)} [\tilde{p}_i(z) - \tilde{p}_i(\hat{B}_i(\lambda_i - \lambda_i z))] , \tag{2.23}$$

where r_i is the mean switchover time from queue i to queue $(i+1)$ mod N. As on p. 110 of [43], the LST of the waiting time of an arbitrary customer at queue i is

$$\hat{W}_i(s) = \frac{(1 - \sum_{k=0}^{N-1} \rho_k)}{\sum_{k=0}^{N-1} r_k} \cdot \frac{[\tilde{p}_i(\hat{B}_i(s)) - \tilde{p}_i(1 - s/\lambda_i)]}{s - \lambda_i + \lambda_i \hat{B}_i(s)} . \tag{2.24}$$

2.3 Computational Complexity and Comparison With Other Algorithms

Our inversion algorithm for the computation of distributions and moments requires computation of the transform at a number of complex values of **z**. From equation (A.13) and the discussion after equation (A.8) in the Appendix, we see that the computation of the n^{th} moment requires computation of the transform at $(nl + 1)$ values and the computation of one point of the distribution requires computation of the transform at about $40l$ values, where l is a parameter for roundoff error control. For 8-to-10-place accuracy we need $l = 2$ for the moment computation and $l = 1$ for the distribution computation.

In Sections 2.1 and 2.2 we have shown that the computational complexity for computing one transform value is $O(m)$ in the transient case and $O(I)$ for the steady-state case, where m is the number of queue visits and I is the number of iterations for convergence of the transform value, determined by the stopping criterion (2.20). Hence, for all our moment and distribution computations the complexity is $O(m)$ in the transient case and $O(I)$ in the steady-state case. Henceforth we only discuss the computational complexity of the steady-state case since the transient computation is faster.

A key measure of the computational complexity is the dependence of I upon N, the number of queues. In Section 6 we show that $I = O(N^{\alpha})$ where α is in the range 0.6 to 0.8. So for all our computations also the complexity is $O(N^{\alpha})$. This is for computing performance measures at one queue. For computation at all queues the complexity is $O(N^{1+\alpha})$. For the computation of mean waiting times at gated or exhaustive systems the algorithm by Sarkar and Zangwill [40] has been widely reported to be the fastest, e.g., see [44]. The computational complexity of that algorithm is $O(N^3)$ (same complexity for one queue and all queues) and hence is considerably slower than ours for large N (see Example 1 in Section 6).

Recently Konheim, Levy and Srinivasan [34] developed an algorithm for computing the mean and second moment of the waiting time with complexity $O(N)$ for one queue and $O(N^2)$ for all queues. Since $\alpha < 1$ we are somewhat faster than this algorithm as well. The algorithm in [34] may be extended to higher moments as well but would involve pretty cumbersome expressions. Of course, the main focus of our paper is on computing distributions and asymptotic parameters, for which no previous work has been reported.

It is to be noted that both our method and the method in [34] are iterative, so that the convergence becomes slow as the total server utilization ρ approaches 1. Dependence of I on ρ and the target error ε is discussed in Section 6.3. The iterative scheme in [34] has computational complexity $O(\ln\varepsilon / ln \rho)$. Our numerical experience discussed in Section 6.3 indicates that the above is also an upper bound on our computational complexity. By contrast, the computational complexity in [40] does not critically depend on ρ. Hence, for ρ sufficiently close to 1, the algorithm in [40] would be faster. However, our algorithm does work adequately even at $\rho = 0.95$.

3. Other Polling Disciplines

An advantage of our iterative-transform-computation and transform-inversion approach is that it is applicable not only to the gated case considered in Section 2 but to many other polling models as well. We could give details for other disciplines as we did in the gated case, but to save space we only provide brief descriptions. For these other models we can also readily compute distributions and an arbitrary number of moments of several steady-state, transient and time-varying performance measures of interest (such as waiting time, queue length and cycle time). Furthermore, as in the gated case, we can do the computations quickly and thus handle large models.

3.1 Exhaustive Service Discipline

The expressions for $\hat{p}_{i,m}(\mathbf{z})$ for this model may be readily obtained by replacing the service-time transform in (2.5) by an M/G/1 busy-period transform and by replacing the expression $\sum_{j=0}^{N-1} (\lambda_j - \lambda_j z_j^{(k-1)})$ in the same equation by $\sum_{\substack{j=0 \\ j \neq w(i-k)}}^{N-1} (\lambda_j - \lambda_j z_j^{(k-1)})$.

Other steady-state and transient transform expressions may also be readily obtained with minor modifications of the corresponding expressions in the gated case; see [43]. The busy-period transform may also be computed iteratively by starting the iteration with a transform value of 0 or 1; see [6]. However, due to this additional iteration, the computation in the exhaustive case is slower than for the gated case. However, there are other factors favoring the exhaustive case. For example, to compute the steady-state queue-length and waiting-time distributions, the gated case requires two computations of $\tilde{p}_i(\cdot)$ (see (2.23) and (2.24)) whereas the exhaustive case requires just one computation of $\tilde{p}_i(\cdot)$ (see (4.29) and (4.32) of [43]). We have implemented the exhaustive case and provide numerical examples in Section 6. Our observation is that typically the exhaustive case is about 5-to-10 times slower than the gated case.

3.2 Mixture of Gated and Exhaustive Disciplines

It is also not difficult to treat the more general model in which the exhaustive discipline is used at some queues, while the gated discipline is used at others. In an iterative step for computing $\hat{p}_{i,m}(\mathbf{z})$ use the equations for either the gated or the exhaustive case depending on the service discipline at the previous queue. For all other transform expressions use the gated expression if queue i has gated service and exhaustive expression if queue i has exhaustive service discipline.

3.3 Polling Table

Here the server does not visit the queues in a strictly cyclic fashion. There is a bigger cycle of length $M(>N)$ which is repeated and within the bigger cycle each queue is visited one or more times according to a fixed pattern or table [11,16]. A special case of the polling table, known as elevator polling, has also received considerable attention [7,20,45]; then the bigger cycle is of the form $1,2,\ldots,N,N,N-1,\ldots,2,1$. Our iterative method for transform computation readily extends to polling tables by noting that $w(i-k)$ in (2.4) and (2.5) has to be replaced by the index of the station appearing k steps earlier in the polling table. Of course there also has to be minor modifications in other transform expressions.

3.4 Globally Gated Discipline

In this discipline [13] a gate is closed at the beginning of service at a specific queue (say queue 0). For the next N queue visits only those customers are served that arrived before the gate closing instant. This case may be treated just like the gated case, but instead of relating the N-dimensional generating functions of queue-length variables at polling instants in successive queues, we have to do that at successive gate closing instants. The parameter m in Section 2 representing the number of queues visited prior to the current polling instant would have to be replaced by a parameter representing the number of cycles prior to the current gate closing instant. In each term of the m-fold sum in (2.4), instead of one term corresponding to one switchover time, there has to be N terms corresponding to N switchover times between successive gate closing instants. Other equations need to be adjusted similarly, but clearly the basic method goes through without any difficulty.

3.5 Open Network Polling System

In this model [41] a customer, after getting service from queue i, leaves the system with probability p_{i0} and goes to queue j with probability p_{ij} ($j = i$ is allowed). Clearly $\sum_{j=0}^{N} P_{ij} = 1$. (To be consistent with the notation in [41], we now assume that the queue indices are $1, 2, \ldots, N$ instead of $0, 1, \ldots, N-1$, as we assume in the rest of the paper.) As is clear from (3.2a) and (3.2b) of [41], we get recursions very similar to those in Section 2.1 with the exception that a new generating function $P_i(\mathbf{z}) = \sum_{j=0}^{N} p_{ij} z_j$ is also involved, which of course is readily computable. So again the basic method goes through with appropriate modifications in the transform expressions as given in [41].

3.6 Closed Polling System

In this model [9] there are no external arrivals and the total number of customers stays the same, which is obtained by setting $p_{i0} = 0$ for $i = 1, 2, \ldots, N$ in the model of Section 3.5. The transient behavior may be studied similarly to that in Section 3.5. However [9] shows that all the steady-state transforms have explicit expressions, unlike nearly any other polling system. So our transform inversion method not only works, but it is very fast since no iterative step is required.

3.7 Other Models

Evidently many other models besides the six mentioned above can also be solved by transform inversion. Two other important kinds of polling models that evidently can be solved by the inversion method are models where a server stops whenever the system becomes empty [24] and models with certain types of server interruptions [15]. Also, as pointed out by Resing [39], for many polling systems (gated and exhaustive are special cases) the joint queue length process at polling instants of a fixed queue is a multitype branching process with immigration. For all these systems, joint generating functions at successive polling instants may be recursively related, so that our inversion method would work.

4. Zero Switchover Times

As the switchover times approach zero, the average cycle time approaches zero and many transient and steady-state quantities defined as averages at polling instants (such as number in system, server visit time, etc.) also approach zero since there are infinitely many polling instants in finite time whenever the system is empty. However, the steady-state queue length and waiting time (whose transforms are given by (2.23) and (2.24) in the gated case) remain non-trivial and well defined. We show that both these quantities can readily be computed by our iterative-transform calculation and transform-inversion method. We explicitly do it only for the waiting time but a very similar treatment is possible for the queue length. Also the same treatment should be applicable for all the other models in Section 3. There has been a considerable literature focusing on the case of zero switchover times and the relationship between zero and non-zero switchover times [16,21,22,28,31,37,42], but there has not been any previous attempt at transform inversion.

We assume that the mean switchover time $r_k = r$ and that $r \to 0$. In this section we consider strictly cyclic polling and so the steady-state waiting times and queue lengths in the limiting system as r_k approaches zero are well defined. However, for non-strictly-cyclic polling (such as polling tables) the limiting system is not well defined unless something more is specified. In [24] the extra specification is the stipulation that the server stops at a specified queue whenever the system is empty. In [16] the extra specification involves the fixed quantities $p_k \equiv \lim_{R \to 0} r_k / R$, where R is the total switchover time and r_k is the switchover time between pseudostations k and $k+1$. (Pseudostation k is the k^{th} station polled in the polling table. The same station may appear more than once in the polling table, each time as a different pseudostation). Note that the quantities p_k add up to one and essentially p_k represents the probability that at the instant of a new message arrival an an empty system the server is in transit between pseudostations k and $k+1$. See [24] and [16] for more details. We just point out that for non-strictly-cyclic cases we can use either the approach of [24] or [16].

For the cyclic polling system considered here, note that (2.24) may be written as

$$\hat{W}_i(s) \;=\; \frac{(1 - \sum_{k=0}^{N-1} \rho_k)}{s \,-\, \lambda_i \,+\, \lambda_i \hat{B}_i(s)} \; \frac{\tilde{p}_i(\hat{B}_i(s)) - \tilde{p}_i(1 - s/\lambda_i)}{Nr} \;. \tag{4.1}$$

As $r \to 0$, $\hat{R}_i(s) \to 1$ for all i and s. Therefore, from (2.14), $\hat{p}_i(\mathbf{z}) \to 1$ for all i and $\mathbf{z}$. So both the numerator and denominator of (4.1) approach zero. Applying L'Hospital's rule to (4.1), we get

$$\hat{W}_i(s) \;=\; \frac{(1 - \sum_{k=0}^{N-1} \rho_k)}{s \,-\, \lambda_i \,+\, \lambda_i \hat{B}_i(s)} \; \frac{\dfrac{\partial \tilde{p}_i(B_i(s))}{\partial r}\Big|_{r=0} \,-\, \dfrac{\partial \tilde{p}_i(1 - s/\lambda_i)}{\partial r}\Big|_{r=0}}{N} \;. \tag{4.2}$$

Taking logarithms on both sides of (2.14) yields

$$\log \hat{p}_i(\mathbf{z}) = \sum_{k=1}^{\infty} \log \hat{R}_{w(i-k)}(y^{(k-1)}) .$$

Taking derivatives on both sides with respect to r we get,

$$\frac{\dfrac{\partial \hat{p}_i(\mathbf{z})}{\partial r}}{\hat{p}_i(\mathbf{z})} = \sum_{k=1}^{\infty} \frac{\dfrac{\partial \hat{R}_{w(i-k)}(y^{(k-1)})}{\partial r}}{\hat{R}_{w(i-k)}(y^{(k-1)})} . \tag{4.3}$$

As $r \to 0$, $\hat{p}_i(\mathbf{z}) \to 1$ and $\hat{R}_{w(i-k)}(y^{(k-1)}) \to 1$. Hence, we get

$$\frac{\partial \hat{p}_i(\mathbf{z})}{\partial r}\Big|_{r=0} = \sum_{k=1}^{\infty} \frac{\partial \hat{R}_{w(i-k)}(y^{(k-1)})}{\partial r}\Big|_{r=0} . \tag{4.4}$$

Since we are going to let $r \to 0$, we specify how the switchover times depend on r. Let the switchover times be i.i.d. random variables, each distributed as r times a fixed random variable with mean 1, i.e., be $X_r \equiv rX_1$ where X_1 is a nonnegative random variable with mean 1 and finite higher-order moments. Since

$$\frac{\partial E e^{-srX_1}}{\partial r}\Big|_{r=0} = E(-sX_1 e^{-srX_1})\Big|_{r=0} = -s , \tag{4.5}$$

$$\frac{\partial \hat{R}_{w(i-k)}(s)}{\partial r}\Big|_{r=0} = -s . \tag{4.6}$$

Combining (4.4) and (4.6), we get

$$\frac{\partial \hat{p}_i(\mathbf{z})}{\partial r}\Big|_{r=0} = -\sum_{k=1}^{\infty} y^{(k-1)} . \tag{4.7}$$

Now we can use the same iterations given by (2.14)–(2.19) but replace (2.14) with (4.7) above to compute $\dfrac{\partial \hat{p}_i(\mathbf{z})}{\partial r}\Big|_{r=0}$, as needed in (4.2). Then the stopping criterion (2.20) has to be replaced by the following stopping criterion derived from (2.18): Stop at $k = I$ whenever

$$|y^{(k)}| < \varepsilon \tag{4.8}$$

for suitably small ε. Noting that $\dfrac{\partial \tilde{p}_i(z)}{\partial r} = \dfrac{\partial \hat{p}_i}{\partial r}(1,1, \ldots, 1,z,1, \ldots, 1)$ with z appearing in the i^{th} place, we can thus readily compute (4.2) which gives the steady-state waiting time transform.

We believe that the iteration we have just derived for obtaining steady-state transform values with zero switchover times is new. Note that the computational complexity in the zero switchover time case is the same as in the non-zero switchover time case and is faster than alternative algorithms (for mean waiting time and queue length computation).

5. Asymptotic Analysis

Even though we can exactly compute the steady-state waiting-time tail probabilities $P(W>x)$ by transform inversion, it is useful to seek an asymptotic form, because it provides insight, because it often provides a good simple approximation to small tail probabilities, and because it enables us to quickly compare two different distributions by looking at the asymptotic parameters instead of looking at the entire distributions.

We conjecture that the steady-state waiting-time distribution for an arbitrary customer at a queue in a polling system with gated or exhaustive service discipline has the following asymptotic form:

$$P(W > x) \sim \alpha x^{\beta} e^{-\eta x} \quad \text{as} \quad x \to \infty \;, \tag{5.1}$$

for $\eta > 0$ and $\alpha > 0$, where $f(x) \sim g(x)$ means that $f(x)/g(x) \to 1$ as $x \to \infty$, provided that the service-time and switchover-time distributions have finite moment generating functions, i.e., $\hat{B}_i(-s) < \infty$ and $\hat{R}_i(-s) < \infty$ for some positive real s. In general, the parameters α, β and η in (5.1) may depend on the queue index and the type of service discipline. This conjecture is supported by our previous work on GI/G/1 queues with the FIFO service discipline [2,3] and M/G/1 queues with the LIFO service discipline [4], but it remains to be proved mathematically.

We verify it in our examples numerically.

Clearly, among the three asymptotic parameters in (5.1) the most important one is η since it represents the exponential decay rate, the next most important one is β since it represents a polynomial growth/decay rate, and the least important is α since it represents just a constant multiplier.

In this section we show how to compute α, β and η numerically. From [1,17], we know that if (5.1) holds, then the n^{th} moment μ_n has the following asymptote

$$\mu_n \sim \frac{\alpha \Gamma(\beta + n + 1)}{\eta^{\beta + n}} \quad \text{as} \quad n \to \infty \ . \tag{5.2}$$

We can estimate α, β and η from moments μ_{n-2}, μ_{n-1} and μ_n for large n. Specifically, let

$$a_n = \mu_n / \mu_{n-1} \tag{5.3}$$

$$\eta_n = \frac{1}{a_n - a_{n-1}} \tag{5.4}$$

$$\beta_n = \frac{a_{n-1}}{a_n - a_{n-1}} - n + 1 \tag{5.5}$$

$$\alpha_n = \frac{\mu_n \eta^{\beta + n}}{\Gamma(\beta + n + 1)} \tag{5.6}$$

Then it can be shown [1] that, under additional conditions,

$$\alpha = \lim_{n \to \infty} \alpha_n \quad \beta = \lim_{n \to \infty} \beta_n \quad \text{and} \quad \eta = \lim_{n \to \infty} \eta_n \ . \tag{5.7}$$

Hence, if we compute α_n, β_n and η_n for increasing n and if these quantities converge to some limit, then we determine the asymptotic parameters. (For practical purposes, when we establish (5.7), we also justify (5.1), but in general the implication is only one way: (5.1) implies (5.2)–(5.7), but (5.2)–(5.7) does not imply (5.1). A counterexample is given in Section 2 of [1]. This numerical procedure is viable since we can compute the moments μ_n accurately even for very large n (100 or more) using the moment computation procedure of [17], which is described briefly in the Appendix. However, when $\beta \neq 0$ the estimators in (5.4)–(5.6) converge relatively slowly, in particular, at a rate of order $O(n^{-1})$. Much faster convergence can be obtained by extrapolating, as described in [1]. We use the extrapolation to advantage here. Specifically, a k^{th}-order extrapolation has a convergence rate of $O(n^{-k})$. Typically, using $k = 4$ we can estimate the asymptotic parameters more accurately using just the first 20 moments than we can with the simple estimators in (5.4)–(5.6) even using 100 moments. This significantly speeds up computation. Furthermore, using the more accurate extrapolation, we can even compute a two-term asymptotic expansion of the form

$$P(W > x) - \alpha_1 x^{\beta} e^{-\eta x} \sim \alpha_2 x^{\beta - 1} e^{-\eta x} \quad \text{as} \quad x \to \infty \ , \tag{5.8}$$

which provides a better approximation to the tail probabilities; see [1] for the details. Finally, we mention that the simple estimators (5.4) and (5.6) converge very rapidly when $\beta = 0$. Then the extrapolations are not needed (and do not help).

The most important asymptotic parameter η is also the dominant singularity (i.e., the singularity closest to the origin) of the LST of the waiting-time distribution. Also it is on the negative real axis. (In general we cannot preclude the existence of other singularities on the circle of convergence). The parameter η is often not easy to find by a search procedure since the dominant singularity need not be a simple pole and there may be other singularities nearby. However, once we find η by the moment method, we can check the transform and verify that it indeed represents the dominant singularity. The numerical method (5.2)–(5.7) thus clearly supports the weaker logarithmic asymptotic statement

$$\lim_{x \to \infty} x^{-1} \log P(W > x) = -\eta \ . \tag{5.9}$$

6. Numerical Examples

In this final section, we first discuss ways we verified our accuracy and then we present five illustrative examples.

6.1 Verification

Since we are unaware of any published results on distributions and higher moments for the models in this paper, an important issue is the verification of the accuracy of our numerical computations. We could adapt the algorithms in [12,35] to check accuracy for small models and we could check accuracy for larger models approximately using simulation or the approach in [25]. We did not do these checks, but did take the following steps:

1. We implemented the algorithms in Ferguson and Aminetzah [27] and Sarkar and Zangwill [40] and verified our mean computation in all cases except the largest example with 1000 queues, for which case the algorithms in [27] and [40] are too slow. Even for the mean computation we use transform inversion. Hence, the fact that the transform inversion is accurate for means leads us to believe that it is accurate in other cases as well. The original algorithms in [27] and [40] are only for non-zero switchover times, whereas our examples are for both zero and non-zero switchover times. To get the zero-switchover-time cases, we used the appropriately modified versions of [27] and [40] as described in Choudhury [16] and Garner [29].

2. For steady-state performance measures the fact that the transforms converge with an error specification of about $10^{-12} - 10^{-13}$ shows that the transforms are being computed accurately. Our experience with the transform inversion algorithms in other contexts [5,17,19] tells us that if the transform is being computed accurately, it is most likely inverted accurately as well.

3. In the special case of $N = 1$ and zero-switchover time, both the gated and exhaustive disciplines should match exactly an M/G/1 queue with FIFO service for which alternate Pollaczek-Khintchine transform expressions are available. We did observe that this is the case.

4. As explained in Section 5, we conjecture that $P(W>x) \sim \alpha x^{\beta} e^{-\eta x}$ as $x \to \infty$ and the parameters α, β, η may be computed from three successive high order moments. We do observe that the parameters α, β, η estimated from increasing higher-order moments converge. Furthermore, since η is the location of the dominant (real) singularity of waiting time LST it can be found alternatively by searching on the negative real-axis. We observe that the two methods for finding η agree. This provides an indirect accuracy verification of the moment computation algorithm. Next, the computed distribution has the computed asymptote mentioned above. Thus the moment inversion and the distribution inversion serve as checks on each other. In difficult cases, we can exploit the two-term asymptote in (5.8) to obtain a more accurate check.

5. In example 3 we show that the transient cycle-time distribution approaches the steady-state values. We observed this convergence in all other cases as well, showing that there are no significant error involved in these computations.

6. As explained in the Appendix, there is a parameter l for controlling round-off error. Typically $l=1$ is sufficient for the distribution computation and $l=2$ is sufficient for moment computation. Different values of l correspond to different contours on the complex plane. Hence, if two computations are done using different values of l and they agree up to, say 8 decimal places, then it is very likely that they are both accurate up to that many places. We verify all computations in this paper using $l=1$ and 2 for distributions and $l=2$ and 3 for moments. In all cases the agreement is good.

For these reasons, we are confident that the accuracy of all computations to be shown are high (8 or more decimal places), even though we do not show all the numerics up to that many places.

All computations in this paper are done on a SUN SPARC-2 workstation using double-precision arithmetic.

6.2 Example 1: 1000 Queues

We first consider an asymmetric 1000-queue system with gated service discipline. All queues are assumed to have the same service time distribution, a two-stage Erlang (E_2) distribution with mean 1 and coefficient of variation 0.5. All switchover times are assumed to be zero. The arrival rates at queues $0, 1, \ldots, 999$ are

$$\lambda_i = \begin{cases} 2\times10^{-4} + i\times2\times10^{-6} & \text{for } 0 \le i \le 500 \\ 2\times10^{-4} + (1000-i)\times2\times10^{-6} & \text{for } 501 \le i \le 999 \ . \end{cases} \tag{6.1}$$

Formula (6.1) makes the smallest arrival rate 2×10^{-4} at queue 0, the largest arrival rate 1.2×10^{-3} at queue 500, and the overall server utilization $\rho = 0.7$.

We show below the first five moments and 3 points of the steady-state waiting-time distribution seen by an arbitrary customer at queue 0.

Performance Measure	1st Moment	2nd Moment	3rd Moment	4th Moment	5th Moment
Value	3.497866	45.64751	1008.372	31725.30	1300039.0
Performance Measure	$Pr(W > 10)$	$Pr(W > 20)$	$Pr(W > 30)$		
Value	0.1023567	0.02536360	0.00706867		

Table 1. Numerical results for the steady-state waiting-time distribution in the 1000-queue example.

The mean was computed in less than 5 seconds, all 5 moments were computed in a little over 1 minute, and the distribution values were computed in about 5 minutes. By contrast, we estimated that the algorithm in [40] would have needed 3.5 hours to compute just the mean and the algorithm in [27] would have taken substantially longer, again to compute just the mean.

6.3 Example 2: Rate of Convergence to Steady State

All our steady-state computations have a complexity $O(I)$ where I is the number of iterations needed to satisfy (2.20) in computing one transform value. In this example we study numerically how I depends on the number of queues, N, server utilization, ρ, and error criterion, ε. The number I also depends on several other parameters, such as the service-time and switchover-time distributions, whether the computation is for mean or for higher moments, and so on. However, those factors have been observed to be less critical than N, ρ and ε.

We consider the average number of iterations needed per transform value in computing the mean steady-state waiting time in a symmetric system with gated service discipline, zero switchover time and a gamma service-time distribution with mean 1 and squared coefficient of variation 2. Figure 1 shows how I grows with N for fixed ρ and ε; here $\rho = 0.7$ and $\varepsilon = 10^{-12}$. We see that the dependence is slower than linear. We also show how I would have grown if the dependence were linear. It appears that $I = O(N^{\alpha})$ where $\alpha \approx 0.77$. We experimented with several other cases and in the range of 1 to 100 queues we have found that consistently $I = O(N^{\alpha})$ for α between 0.6 and 0.8.

Next we fix N at 10, ε at 10^{-12} and show in Table 2 how I grows with ρ.

ρ	0.2	0.4	0.6	0.7	0.8	0.9	0.95
I	73	118	196	270	413	824	1604

Table 2. Number of iterations required as a function of the traffic intensity ρ.

Clearly I approaches ∞ as ρ approaches 1. The growth rate is slower than $-1/\log \rho$ and faster than $1/(1-\rho)$. This appears to be the case in other examples as well.

Finally, we fix N at 10, ρ at 0.8 and show in the table below how I grows with ε. I appears to grow roughly linearly with $-\log \varepsilon$.

ε	10^{-8}	10^{-10}	10^{-12}	10^{-14}
I	172	297	413	530

Table 3. Number of iterations as a function of the specified error ε in the stopping criterion.

6.4 Example 3: Time-Varying Behavior

In this example we consider transient and time-varying behavior of a polling system. This example is motivated by the fact that in many polling systems (e.g., AT&T's 4ESS Switching System [32]), some overload control action is taken whenever the cycle time (or the weighted sum of several successive cycle times) exceeds a threshold. The threshold is chosen large enough so that it is unlikely to be exceeded under normal traffic conditions. We do not study the effect of any overload-control mechanism, but rather study how the transient cycle times behave under a sudden surge of overload and specifically how the probability of exceeding a large threshold changes during and after the overload. We consider a symmetric 10-queue system with a gated service discipline. The switchover time is a 4-stage Erlang (E_4) distribution with mean 0.1 and squared coefficient of variation 0.25. The service time is a 2-stage Erlang (E_2) distribution with mean 1 and squared coefficient of variation 0.5. The arrival rate is time-varying and it changes from cycle to cycle so that the instantaneous offered load ρ (the product of total arrival rate and mean service time) changes with cycle number as shown in Figure 2.

We let the threshold be 40. When the offered load is stationary with $\rho = 0.8$, the threshold is 8 times the mean cycle time, and the probability of exceeding it would be 8.862472×10^{-4} (this value we obtain from the steady-state analysis). Let C_n represent the length of the n^{th} cycle. In Figure 3 we show how the time-dependent cycle-time tail probabilities $P(C_n > 40)$ evolve with n under the time-varying load shown in Figure 2. Both during and after the overload, the probability of exceeding the threshold remains much higher than the steady-state value. This demonstrates that the observed cycle time should be a good indicator of an overload.

6.5 Example 4: Asymptotic Parameters

In this example we show how to get insights by computing asymptotic parameters numerically as explained in Section 5. We consider symmetric gated systems with a two-stage Erlang (E_2) service-time distribution with mean 1 and squared coefficient of variation 0.5 at each queue. The server utilization is 0.81. We first consider zero switchover times, and see how the asymptotics depends on N. We consider 3 cases: $N = 1$, 3 and 9. The results for the steady-state waiting-time distribution are given in Table 4. In the cases $N = 3$ and $N = 9$ we give results both for the estimators in (5.3)–(5.6) and the better results obtained by extrapolation. Extrapolation is not used when $N = 1$.

For $N = 1$, the three estimators in (5.3)–(5.6) converge to limits by the 11^{th} moment and $\beta = 0$. This is no surprise since this case is identical to an M/G/1 queue with the FIFO service discipline. In that setting it is known [3] that there is a pure-exponential asymptote and the distribution approaches this limit pretty quickly. However, for $N = 3$ and 9, the convergence to the asymptote is not nearly as quick. Moreover, significantly, $\beta \neq 0$ when $N > 1$. Without using the extrapolation in [1], even by the hundredth moment, η has only about 4 significant digits, and α and β have only 1 significant digit. The error in η propogates to β and α, so that the estimates of β and α without extrapolation tend to converge to the wrong limits. From the numerical results, it would appear that α_n and β_n have 3 significant digits by $n = 100$, but this is not the case, as can be seen from the extrapolation. The extrapolation does much better with 20 moments than the estimators in (5.3)–(5.6) do with 100 moments. However, since α and β are less important parameters, we are pretty close to the true asymptote by approximating η, α and β by their values obtained from the 100^{th} moment. With extrapolation, all digits given are significant.

We plot the true tail probabilities and the asymptotes in Figure 4. For $N = 1$, the asymptote and the true distribution are nearly indistinguishable. For $N = 3$ and $N = 9$, the asymptotic approximation gets quite good for tail probabilities 10^{-3} or lower.

Table 4 has important implications for system performance. Even though the mean waiting times are the same in all cases, the asymptotic exponential decay rates η are quite different, resulting in considerably bigger tail probabilities with increasing N. A second observation is that the second asymptotic parameter β evidently converges to -1. We remark that $\beta = -1.0$ corresponds to a logarithmic singularity in the transform. We observed this phenomenon for several other examples with zero switchover times as well (when $N > 1$).

Next we consider the effect of non-zero switchover times. For this purpose, we fix N at 3. We look at three cases: Case 1 has zero switchover times; Case 2 has switchover times with mean 0.1 and an exponential distribution with squared coefficient of variation 1; Case 3 has relatively long and highly variable switchover times; they have a gamma distribution with mean 1 and squared coefficient of variation 10.

The results are shown in Table 5 and Figure 5. The degree of convergence of the asymptotic parameters by the 100^{th} moment is about the same as in Table 4. As in Table 4, the extrapolation technique in [1] provides significantly greater accuracy. It is remarkable that Case 2 has a considerably bigger mean than Case 1, but both cases evidently have the same asymptotic exponential decay rate and therefore their small tail behaviors are not too different as can be seen in Figure 5. However, in Case 3, both the mean and the small tail probabilities are much larger than in Case 1 (η is considerably smaller). We observe that unlike between Cases 1 and 2, η changes significantly in Case 3. The change evidently occurs in Case 3 because the asymptotic decay rate η for the switchover-time distribution in Case 3 is 0.1, which is smaller than the asymptotic decay rate η for the waiting-time distribution in Case 1. Evidently, in general, the η for the waiting time has to be smaller than the η for switchover time. By contrast, the η for the switchover time in Case 2 is 10, which is much larger than the η for waiting time in Case 1. For all cases we did verify by a search method that the dominant singularity of the waiting-time LST is indeed located very near $-\eta$ for η in Tables 4 and 5.

From Figure 5, we also observe that the asymptotic approximation is quite good in Cases 1 and 2 but it is not so good in Case 3. For Case 3 in Figure 5, we also show the two-term asymptotic approximation

$$F^c(x) \approx \alpha_1 x^\beta e^{-\eta x} + \alpha_2 x^{\beta-1} e^{-\eta x} , \tag{6.2}$$

Case 1, $N = 1$

Moment Order	Moment	α_n	η_n	β_n
1	3.197368	—	—	—
3	285.9021	0.815720	0.2602494	0.011225
11	0.9299524×10^{14}	0.831021	0.2593260	0
12	0.4303243×10^{16}	0.831021	0.2593260	0

Case 2, $N = 3$

Moment Order	Moment	α_n	η_n	β_n
1	3.197368	—	—	—
3	354.8475	1.10836	0.2060491	–0.272482
98	$0.1146892 \times 10^{227}$	6.52631	0.1758953	–0.981276
99	$0.6391126 \times 10^{229}$	6.53219	0.1758950	–0.981445
100	$0.3597829 \times 10^{232}$	6.53862	0.1758947	–0.981630
	extrapolation	7.3487	0.175878	–1.00003

Case 3, $N = 9$

Moment Order	Moment	α_n	η_n	β_n
1	3.197368	—	—	—
3	450.4123	1.37951	0.1661709	–0.465349
98	0.185638×10^{234}	5.76511	0.1480554	–0.980175
99	$0.1229012 \times 10^{237}$	5.77150	0.1480551	–0.980377
100	$0.8219661 \times 10^{239}$	5.77755	0.1480548	–0.980567
	extrapolation	6.5744	0.148039	–1.0005

Table 4. The asymptotic parameters of the steady-state waiting-time distribution with zero switchover times.

where α_1 is the old α, as described in [1]. For Case 3, $\alpha_2 = 312$. From Figure 5, we see that the second term provides noticeable improvement for Case 3. Multi-term asymptotic approximations were obtained in other cases, but they are not shown since they are indistinguishable from the exact tail probabilities in Figure 5.

Case 1. Switchover Time = 0

Moment Order	Moment	α_n	η_n	β_n
1	3.197368	—	—	—
3	354.8475	1.10836	0.2060491	–0.272482
98	0.1146892×10^{227}	6.52631	0.1758953	–0.981276
99	0.6391126×10^{229}	6.53219	0.1758950	–0.981445
100	0.3597829×10^{232}	6.53862	0.1758947	–0.981630
	extrapolation	7.3487	0.175878	–1.00003

Case 2. Switchover Time: Mean = 0.1, $c_b^2 = 1$

Moment Order	Moment	α_n	η_n	β_n
1	4.25000	—	—	—
3	563.6498	1.1387	0.1995206	–0.148503
98	0.4163636×10^{227}	5.0027	0.1758908	–0.735567
99	0.2326088×10^{230}	5.0068	0.1758905	–0.735720
100	0.1312734×10^{233}	5.0099	0.1758903	–0.7358837
	extrapolation	5.4534	0.175878	–0.7492

Case 3. Switchover Time: Mean = 1, $c_b^2 = 10$

Moment Order	Moment	α_n	η_n	β_n
1	18.22368	—	—	—
3	20856.95	0.2815	0.0900686	0.587808
98	0.2172506×10^{272}	10.9981	0.0605776	–0.938639
99	0.3516792×10^{275}	10.9650	0.0605779	–0.938166
100	0.5750940×10^{278}	10.9331	0.0605782	–0.937710
	extrapolation	7.644	0.060593	–0.8894

Table 5. Asymptotics of the steady-state waiting time with non-zero switchover times, $N = 3$.

6.6 Example 5: Comparing Gated and Exhaustive Disciplines

It is well known that for symmetric polling systems

$$E(W_{FCFS}) \le E(W_{\text{Exhaustive}}) \le E(W_{\text{Gated}}) \ , \tag{6.3}$$

with equality holding in the case of zero switchover times. In this example we see how the corresponding high-order moments and tail probabilities compare. However, instead of the entire tail distribution we consider only the asymptotic exponential decay rate η.

We consider symmetric 10-queue systems with gated or exhaustive service disciplines (either all gated or all exhaustive), $\rho = 0.8$, service times distributed as gamma with mean 1 and squared coefficient of variation 2. Table 6 below describes the results with zero switchover times.

Note that the means are the same, but for all other moments (we only show 2nd and 3rd), exhaustive is larger than gated. Of course both are larger than FIFO. Also the asymptotic exponential decay rate is smaller for exhaustive than gated, resulting in bigger tail probabilities.

Next we consider two cases of non-zero switchover times distributed as two-stage Erlang with squared coefficient of variation 0.5. The mean in the first case is 0.1 and in the second case 1.0. The results are displayed in Table 7.

Service Discipline	1st Moment	2nd Moment	3rd Moment	η
FIFO	6.0	92.00	2121.00	0.13007
Gated	6.0	100.68	3391.36	0.07369
Exhaustive	6.0	118.31	3954.61	0.06713

Table 6. A comparison of the steady-state waiting times for three disciplines with zero switchover times.

Case 1. Switchover Time: Mean = 0.1, $c_b^2 = 0.5$

Service Discipline	1st Moment	2nd Moment	3rd Moment	η
Gated	8.725	181.303	6074.64	0.07369
Exhaustive	8.325	180.576	6482.96	0.06713

Case 2. Switchover Times: Mean 1, $c_b^2 = 0.5$

Service Discipline	1st Moment	2nd Moment	10th Moment	11th Moment	η
Gated	33.2500	1654.51	0.352483×10^{20}	0.623732×10^{22}	0.07369
Exhaustive	29.2500	1396.69	0.351401×10^{20}	0.652084×10^{22}	0.06713

Table 7. A comparison of the steady-state waiting times for gated and exhaustive disciplines with small and large switchover times.

In Case 1 the first moment is smaller for exhaustive than for gated, as is well known. The second moment is also smaller, but the third and all subsequent moments are larger for exhaustive than gated. In Case 2, the first 10 moments are smaller for exhaustive, but the 11th and higher moments are larger for exhaustive. Also, in both Cases 1 and 2 the asymptotic exponential decay rate is smaller for exhaustive compared to gated, which implies that if we go sufficiently far out in the tail, the exhaustive tail probability will have bigger values than the gated ones.

Also note that in Table 7 the asymptotic decay rates η are unchanged going from Case 1 to Case 2. This is consistent with Cases 1 and 2 in Table 5. Evidently the asymptotic decay rate with switchover times is the same as without switchover times, provided that the switchover time contribution does not dominate the zero-switchover contribution. This phenomenon can be understood from decomposition results; see Fuhrmann [24] and references therein.

It is well known that if the objective is to minimize the mean delay, then we should prefer the exhaustive discipline to the gated discipline. However, this example shows that if the object is to minimize higher moments or tail probabilities, and if the switchover times are not too large, then gated might be the discipline of choice.

Acknowledgement

We would like to thank Uri Yechiali for a discussion about closed polling systems which led us to undertake this work, and Kin Leung, Martin Eisenberg and the referees for further helpful comments.

References

[1] J. Abate, G. L. Choudhury, D. M. Lucantoni and W. Whitt, ‘‘Asymptotic analysis of tail probabilities based on the computation of moments,’’ (1994) submitted for publication.

[2] J. Abate, G. L. Choudhury and W. Whitt, ‘‘Waiting-time tail probabilities in queues with long-tail service-time distributions,’’ *Queueing Systems* 16 (1994) 311-338.

[3] J. Abate, G. L. Choudhury and W. Whitt, ‘‘Exponential approximations for tail probabilities in queues, I: waiting times,’’ *Oper. Res.,* to appear.

[4] J. Abate, G. L. Choudhury and W. Whitt, ‘‘Calculating the M/G/1 busy-period density and LIFO waiting-time distribution by direct numerical transform inversion,’’ (1994) submitted for publication.

[5] J. Abate and W. Whitt, ‘‘The Fourier-series method for inverting transforms of probability distributions,’’ *Queueing Systems* 10 (1992) 5-88.

[6] J. Abate and W. Whitt, ‘‘Solving probability transform functional equations for numerical inversion,’’ *Oper. Res. Letters* 12 (1992) 245-251.

[7] E. Altman, A. Khamisy and U. Yechiali, ‘‘On Elevator polling with globally gated regime,’’ *Queueing Systems* 11 (1992) 85-90.

[8] E. Altman, P. Konstantopoulos and Z. Liu, ‘‘Stability, monotonicity and invariant quantities in general polling systems,’’ *Queueing Systems* 11 (1992) 35-57.

[9] E. Altman and U. Yechiali, ‘‘Polling in a closed network,’’ *Prob. in the Eng. Inf. Sci.* 8 (1994) 327-343.

[10] K. B. Athreya and P. E. Ney, *Branching Processes,* Springer-Verlag, Berlin, 1972.

[11] J. F. Baker and I. Rubin, ‘‘Polling with a general service order table,’’ *IEEE Trans. Commun.,* COM-35 (1987) 283-288.

[12] J. P. C. Blanc, ‘‘A numerical approach to cyclic-service queueing models,’’ *Queueing Systems* 6 (1990) 173-178.

[13] O. J. Boxma, H. Levy and U. Yechiali, ‘‘Cyclic reservation schemes for efficient operation of multiple-queue single-server systems,’’ *Ann. Oper. Res.* 35 (1992) 187-208.

[14] O. J. Boxma and J. A. Weststrate, ‘‘Waiting time in polling systems with Markovian server routing,’’ *Messung, Modellierung und Bevertung von Rechensysteme,* Proc. Conference Braunschweig, September 1989, Springer-Verlag, Berlin.

[15] O. J. Boxma, J. A. Weststrate and U. Yechiali, ‘‘A globally gated polling system with server interruptions, and applications to the repairman problem,’’ *Prob. Eng. and Inf. Sc.*, 7 (1993) 187-208.

[16] G. L. Choudhury, ‘‘Polling with a general service order table: gated service,’’ *Proc. IEEE INFOCOM ’90,* San Francisco, 1990, 268-276.

[17] G. L. Choudhury and D. M. Lucantoni,‘‘Numerical computation of the moments of a probability distribution from its transform,’’ *Oper. Res.,* to appear.

[18] G. L. Choudhury, D. M. Lucantoni and W. Whitt, ‘‘Numerical solution of $M_t/G_t/1$ queues,’’ *Oper. Res.*, to appear.

[19] G. L. Choudhury, D. M. Lucantoni and W. Whitt, ‘‘Multidimensional transform inversion with applications to the transient M/G/1 queue,’’ *Ann. Appl. Prob.* 4 (1994) 719-740.

[20] E. G. Coffman, Jr., and M. Hofri, ‘‘Queueing models of secondary storage devices,’’ *Stochastic Analysis of Computer Communication Systems,* H. Takagi, ed., Elsevier-Science Publishers, Amsterdam, 1990, 549-588.

[21] R. B. Cooper, ‘‘Queues served in cyclic order: waiting times,’’ *Bell System Tech. J.* 49 (1970) 399-413.

[22] R. B. Cooper and G. Murray, ‘‘Queues served in cyclic order,’’ *Bell System Tech. J.* 48 (1969) 675-689.

[23] M. Eisenberg, ‘‘Queues with periodic service and changeover times,’’ *Oper. Res.* 20 (1972) 440-451.

[24] M. Eisenberg, ‘‘The polling system with a stopping server,’’ *Queueing Systems*, 18 (1994) 387-431.

[25] A. Federgruen and Z. Katalan, ‘‘Approximating queue size and waiting time distributions in general polling models,’’ *Queueing Systems* 18 (1994) 353-386.

[26] M. J. Ferguson, ‘‘Computation of the variance of the waiting time for token rings,’’ *IEEE J. Sel. Areas Commun.* SAC-4 (1986) 775-782.

[27] M. J. Ferguson and Y. J. Aminetzah, ‘‘Exact results for nonsymmetric token ring systems,’’ *IEEE Trans. on Commun.*, Vol. COM-33 (1985) 223-231.

[28] S. W. Fuhrmann, ‘‘A decomposition result for a class of polling models, *Queueing Systems* 11, 109-120.

[29] G. M. Garner, ‘‘Implementation of an efficient algorithm for waiting times in a nonsymmetric cyclic queueing system,’’ AT&T internal technical memorandum, Holmdel NJ, 1988.

[30] L. Georgiadis and W. Szpankowski, ‘‘Stability of token passing rings,’’ *Queueing Systems* 11 (1992) 7-33.

[31] D. Grillo, ‘‘Polling mechanism models in communication systems – some application examples,’’ *Stochastic Analysis of Computer and Communication Systems*, H. Takagi, ed.,

Elsevier-Science Publishers, Amsterdam, 1990, 659-698.

[32] D. J. Houck, K. Meier-Hellstern, F. Saheban and R. A. Skoog, ‘‘Failure and congestion propagation through signalling controls,’’ *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks, Proceedings of the 14th International Teletraffic Congress*, J. Labetoulle and J. Roberts (eds.), Elsevier, Amsterdam, 1994, 367-376.

[33] L. Kleinrock and H. Levy, ‘‘The analysis of random polling systems,’’ *Oper. Res.* 36 (1988) 716-732.

[34] A. G. Konheim, H. Levy and M. M. Srinivasan, ‘‘Descendant set: an efficient approach for the analysis of polling systems,’’ *IEEE Trans. Commun.* 42 (1994) 1245-1253.

[35] K. K. Leung, ‘‘Waiting time distributions for token-passing systems with limited-*K* service via discrete Fourier transform,’’ *Performance ‘90,* King, Mitrani and Pooley (eds.), Elsevier, Amsterdam, 1990, 333-345.

[36] H. Levy, ‘‘Delay computation and dynamic behavior of non-symmetric polling systems,’’ *Perf. Eval.* 10 (1989) 35-51.

[37] H. Levy and L. Kleinrock, ‘‘Polling systems with zero switch-over periods: a general method for analyzing expected delay,’’ *Perf. Eval.*, 13 (1991) 97-107.

[38] D. M. Lucantoni, G. L. Choudhury and W. Whitt, ‘‘The transient BMAP/G/1 queue,’’ *Stochastic Models* 10 (1994) 145-182.

[39] J. A. C. Resing, ‘‘Polling systems and multitype branching processes,’’ *Queueing Systems* 13 (1993) 409-426.

[40] D. Sarkar and W. I. Zangwill, ‘‘Expected waiting time for nonsymmetric cyclic queueing systems — exact results and applications,’’ *Management Science* 35 (1989) 1463-1474.

[41] M. Sidi, H. Levy and S. W. Fuhrmann, ‘‘A queueing network with a single cyclically roving server,’’ *Queueing Systems* 11 (1992) 121-144.

[42] M. M. Srinivasan, S-C. Niu and R. B. Cooper, ‘‘Relating polling models with zero and nonzero switchover times,’’ *Queueing Systems*, to appear.

[43] H. Takagi, *Analysis of Polling Systems*, MIT Press, 1986.

[44] H. Takagi, ‘‘Queueing analysis of polling models: an update,’’ *Stochastic Analysis of Computer and Communication Systems*, H. Takagi, ed., Elsevier Science Publishers, Amsterdam, 1990, 267-318.

[45] H. Takagi and M. Murata, ‘‘Queueing analysis of scan type TDM and polling systems,’’ *Computer Networking and Performance Evaluation*, T. Hasegawa, H. Takagi and Y. Takahasi (eds.), Elsevier Science Publishers, Amsterdam, 1986, 199-211.

Appendix

In this appendix we provide a brief summary of the inversion formulas; see [5,17,19] for more details. We only show one-dimensional formulas, which is all we use in our examples here, but it is also possible to do multi-dimensional inversions (e.g., to compute multivariate distributions), as explained in [19]. Multi-dimensional inversions are essentially iterated one-dimensional inversions.

A.1 Distribution Values

Let X be a random variable with a cdf $F(x) \equiv P(X \leq x)$ and complementary cdf $F^c(x) \equiv 1 - F(x)$. Then the Laplace-Stieltjes transform (LST) of F is

$$\hat{f}(s) \equiv Ee^{-sX} = \int_0^\infty e^{-sx} dF(x) \tag{A.1}$$

and the Laplace transform of $F^c(x)$ is

$$\hat{F}^c(s) \equiv \int_0^\infty e^{-sx} F^c(x)\, dx = \frac{1 - \hat{f}(s)}{s} \ , \tag{A.2}$$

where s is a complex variable with $Re(s) > 0$. We calculate $F^c(x)$ using the inversion formula

$$F^c(x) = \frac{\exp(A/2l)}{2lx} \left[\hat{F}^c \left[\frac{A}{2xl} \right] + 2 \sum_{k_1=1}^{l} \sum_{k=0}^{\infty} (-1)^k Re\left[e^{-ik_1\pi/l} \hat{F}^c \left[\frac{A}{2lx} - \frac{ik_1\pi}{lx} - \frac{ik\pi}{x} \right] \right] \right] \tag{A.3}$$

where $i \equiv \sqrt{-1}$ and A and l are parameters for controlling errors.

The *aliasing error* is given by

$$E_a = \sum_{j=1}^{\infty} e^{-Aj} F^c(x + 2jlx) \ . \tag{A.4}$$

Since $|F^c(x + 2jlx)| < 1$, the aliasing error is bounded by

$$|E_a| \leq \frac{e^{-A}}{1-e^{-A}} \ . \tag{A.5}$$

The aliasing error is controlled by choosing a suitably large A and the roundoff error is controlled by choosing a suitably large l.

The infinite sum in (A.3) is computed using Euler summation. To describe Euler summation, let

$$S = \sum_{k=0}^{\infty} (-1)^k a_k \tag{A.6}$$

and

$$S_j = \sum_{k=0}^{j} (-1)^k a_k \ . \tag{A.7}$$

The infinite sum S in (A.6) is approximated by the *Euler sum*

$$E(m,n) = \sum_{k=0}^{m} \binom{m}{2} 2^{-m} S_{n+k} \ . \tag{A.8}$$

For the computations in this paper, we obtained at least 8-place accuracy by choosing $l = 1$, $A = 21$, $n = 30$ and $m = 11$.

Next, let X represent a discrete random variable with probability mass function p_k and probability generating function $\hat{P}(z) = \sum_{k=0}^{\infty} p_k z^k$. Then we calculate p_k from $\hat{P}(z)$ using the inversion formula

$$p_k = \frac{1}{2lkr^k} \left[\hat{P}(r) + (-1)^k \hat{P}(-r) + 2 \sum_{j=1}^{lk-1} Re[\hat{P}(re^{\pi ij/lk}) e^{-\pi ij/l}] \right] , \tag{A.9}$$

where r and l are parameters for controlling errors. Let $r = 10^{-\gamma/2lk}$. Then the aliasing error in the inversion formula (A.9) is

$$E_a = \sum_{j=1}^{\infty} 10^{-\gamma j} p_{k+2nlk} \ . \tag{A.10}$$

Since $|p_{k+2nlk}| < 1$, the aliasing error is bounded by

$$|E_a| \leq \frac{10^{-\gamma}}{1-10^{-\gamma}} \ . \tag{A.11}$$

The aliasing error is controlled by choosing a bigger γ and the roundoff error is controlled by choosing a bigger l. To get at least 8-place accuracy, a choice of $l=1$ and $\gamma=10$ is adequate.

A.2 Moments

Let $\hat{M}(z)$ represent the moment generating function of a random variable X. We assume $\hat{M}(z)$ to be analytic at $z = 0$. If X is a random variable with LST $\hat{f}(s)$ as in (A.1), then

$$\hat{M}(z) \equiv E(e^{zX}) = \hat{f}(-z) \ . \tag{A.12}$$

If X is a discrete variable with probability generating function $\hat{P}(z)$, then

$$\hat{M}(z) = \hat{P}(e^z) \ . \tag{A.13}$$

Let μ_n represent the n^{th} moment with $\mu_0 \equiv 1$. Then μ_n appears as the n^{th} coefficient in the power-series representation of $\hat{M}(z)$, i.e.,

$$\hat{M}(z) = \sum_{n=0}^{\infty} \mu_n z^n \ . \tag{A.14}$$

To compute the moments, we first compute $\hat{M}(z)$, using either (A.12) or (A.13), and then obtain μ_n using the inversion formula

$$\begin{aligned}\mu_n = \frac{n!}{2nlr^n \alpha^n} \Big[& \hat{M}(\alpha r) + (-1)^n \hat{M}(-\alpha r) \\ & + 2 \sum_{j=1}^{nl-1} Re[\hat{M}(\alpha r e^{\pi ij/l}) e^{-\pi ij/l}] \Big] \ ,\end{aligned} \tag{A.15}$$

where

$$r = 10^{-\gamma/2nl} \tag{A.16}$$

and

$$\alpha = (n-1)\mu_{n-2}/\mu_{n-1} \ . \tag{A.17}$$

Note that (A.15) differs in form from (A.9) because (A.15) has the dynamic scale parameter α in (A.17) computed from the last two moment values. Note that (A.17) may be used to compute α only for $n \geq 3$. To compute μ_1 and μ_2, at first a rough estimate is obtained using $\alpha = 1$. Next α is obtained from μ_1 and μ_2 which then is used to recompute a more accurate α. High accuracy is typically obtained by choosing $\gamma = 11$ and $l = 2$.

Figure 1. The number I of iterations of the transform required to reach steady state as a function of the number N of queues in a symmetric polling model with the gated discipline and zero switchover times. Here $\rho = 0.7$ and $\varepsilon = 10^{-12}$.

Figure 3. The time-dependent cycle-time tail probability (in log scale) for the time-varying load in Figure 2.

Figure 4. Steady-state waiting-time tail probabilities and asymptotes (in log scale) in a symmetric gated model with zero switchover times for the cases of $N = 1,$ 3 and 9.

Figure 5. Steady-state waiting-time tail probabilities and asymptotes (in log scale) in a symmetric gated model with different switchover times.