Performance of Large-Scale Polling Systems with Branching-Type and Limited Service

T. Meyfroyt, M. Boon, S. Borst, O. Boxma

Performance of Large-Scale Polling Systems with Branching-Type and Limited Service

T.M.M. Meyfroyt, M.A.A. Boon, S.C. Borst, O.J. Boxma
tmeyfroyt@gmail.com, {m.a.a.boon, s.c.borst, o.j.boxma}@tue.nl

Department of Mathematics and Computer Science
Eindhoven University of Technology, P.O. Box 513
5600 MB Eindhoven, The Netherlands

February 7, 2018

Abstract

We analyze the performance of large-scale symmetric polling systems where the number of queues grows large, motivated by emerging Internet-of-Things (IoT) applications and smart building environments. We consider a scenario in which the total arrival rate is kept fixed and the individual switch-over time and service time distributions remain the same. This asymptotic regime leads to cycles of infinite length and queue lengths with non-trivial distributions. We show that for most traditional service policies the scaled cycle times converge to a deterministic value in the limit, which in turn implies that the queue lengths at the various nodes become asymptotically independent. Using these insights, we find that the behavior of individual queues simplifies to that of a discrete-time bulk service queue in the limit, so that the marginal queue length and waiting-time distributions become considerably easier to analyze. Additionally, we propose a new flexible k-limited service discipline aimed at striking a good balance between short mean queue lengths and predictable cycle times for deadline-critical applications.

Keywords: Polling models, queue lengths, cycle times, flexible k-limited service.

1 Introduction

In the present paper we investigate the performance of large-scale symmetric polling systems. Polling systems provide canonical models for evaluating the delay performance of systems where multiple queues contend for access to a shared resource and are served in an alternating manner. These models find applications in a wide range of domains, e.g., in computer-communications, production and transportation, and many results are available in the literature [3]. However, asymptotic regimes where the number of queues grows large appear to have received hardly any attention so far. Such scenarios are of strong interest in the context of emerging Internet-of-Things (IoT) applications and smart building environments. In building automation and control systems, efficient and economic networking solutions are required in order to transmit and exchange many kinds of monitoring, control, maintenance and management data through the network [12, 21]. Typically the number of devices and sensors in these scenarios is quite large, and it is critical that the end-to-end delay of the data transmitted satisfies predetermined deadlines in order for these systems to meet their performance and functional requirements.

The particular application that motivated the present study is the so-called *BACnet* (Building Automation and Control networks) protocol, which is specifically designed to meet the communication needs of the latter building automation and control systems [7]. BACnet relies on token-passing algorithms for its medium access control. Token-passing algorithms ensure orderly access to a communication channel by passing a token in a cyclic fashion along all nodes in the network. Only when

a node holds the token, it is allowed to send messages to other nodes, before it has to pass on the token to the next node. The token-passing algorithm determines when and to which node the token has to be passed on. Since there is only a single token in the network, only one device can be transmitting at any given time and no data collisions occur. Access to the network is thus guaranteed, and deadline-critical applications can be supported. Another well-known communication protocol that uses token-passing for its medium access control is the Token Ring mechanism [18]. The delay performance of such token-passing algorithms is typically analyzed using polling models.

In view of the usually sizeable number of nodes in BACnet deployments, we focus in the present paper on the performance of polling systems in an asymptotic regime where the number of queues grows large. Many-queue asymptotics of polling systems have been studied before, but only when the mean switch-over times go to zero as the number of queues grows large. In the limit, such a scenario behaves as a "continuous" spatial polling system where a single server moves at a constant rate along a closed tour, stopping to perform services wherever it encounters requests, which appear at locations independently and uniformly distributed over the tour [9, 13]. In this asymptotic regime, the mean cycle time remains finite and queues have either length zero or one with high probability. In contrast, we consider a regime where the distribution of the individual switch-over times remains the same when the number of queues grows large. This regime gives rise to cycles of infinite length and queue lengths with non-trivial distributions. Performance metrics of interest are the asymptotic queue length, cycle time and waiting-time distributions as the number of queues grows large.

We will first focus on so-called branching-type service disciplines, such as the exhaustive and gated service disciplines. In the exhaustive service policy the server only switches to the next queue if the current queue is empty. It is well known that the exhaustive service discipline minimizes the total amount of work in the system among all cyclic service policies [15], but the disadvantage is that it does not put any restrictions on the cycle times. For this reason, the exhaustive service discipline is ill-suited for deadline-critical applications.

A more suitable service policy for deadline-critical applications is a k-limited service policy, which will be our second focus. Such a policy ensures that the number of customers that are served at each queue does not exceed k, thus to some extent bounding the cycle time. However, a major drawback of a k-limited service policy is the fact that if the server reaches a very long queue, it will still serve at most k customers, even though it possibly did not have to serve any customers at previously visited queues. In order to overcome this drawback, we introduce a *flexible k-limited* service discipline, which exploits the fact that the server sometimes visits a queue which has less than k customers and essentially has time to spare.

As key contributions of this paper, we give explicit results for the covariance of queue lengths, the covariance of visit times and the variance of the cycle time for symmetric polling systems when the server uses a branching-type service discipline. Additionally, we derive the corresponding many-queue limits. Furthermore, we show that in the many-queue regime each individual queue behaves asymptotically as a discrete-time bulk service queue, which significantly simplifies the analysis of the marginal queue length and waiting-time distributions for many service disciplines, including the k-limited service discipline. Finally, we introduce the flexible k-limited service discipline and show how to approximate its performance.

The remainder of this paper is organized as follows. We provide a model description in Section 2. In Section 3 we analyze the many-queue asymptotic performance of branching-type service disciplines, such as the exhaustive, gated and binomial disciplines, which admit a detailed analysis. We then briefly consider general non-idling service policies in Section 4, with a focus on the 1-limited service discipline. In Section 5 we propose the flexible k-limited service policy and provide a way to approximate its asymptotic performance. Finally, in Section 6, we use simulations to compare the performance of the various service disciplines and we investigate how well our asymptotic results can approximate networks of finite size. We summarize our results in Section 7.

2 Model description and preliminaries

We consider $n \geq 1$ queues $Q_1, ..., Q_n$ being served by a single unit-rate server. The server visits the queues in a cyclic non-idling manner. We assume that customers arrive at each Q_i according to independent Poisson processes of rate λ_i. We assume that customers at Q_i have i.i.d. service times with first moment β_i, second moment $\beta_i^{(2)}$ and Laplace-Stieltjes transform (LST) $\mathscr{B}_i(\cdot)$. We define $\Lambda = \sum_{i=1}^{n} \lambda_i$, $\rho_i = \lambda_i \beta_i$ and $\rho = \sum_{i=1}^{n} \rho_i$. The switch-over times of the server for moving between Q_i and the next queue are i.i.d. random variables with first moment s_i, second moment $s_i^{(2)}$ and LST $\mathscr{S}_i(\cdot)$. All interarrival, service and switch-over times are assumed to be independent.

In addition to the Poisson arrivals, we also allow that after a customer receives service, it remains in the system and joins a new queue or that additional customers instantaneously join the system. Specifically, we assume that the service completion of a customer at Q_i leads to $M_{i,j}$ additional customers joining Q_j, where we assume the $M_{i,j}$ to be independent for all i and j. This allows that after a customer has received service, with some probability, it remains in the system and is routed to join another queue, which could arise for example in communication networks where some packets are routed in a multi-hop manner. Additionally, it allows that multiple customers simultaneously join the system, after a customer has received service. Again, one can think of communication networks where some packets may require a response from several nodes.

Let a cycle be defined as the time between two consecutive visits of the server to the same queue. We define the following random variables:

$V_i^{(k)}$: The length of the k'th server visit to Q_i.

$S_i^{(k)}$: The length of the k'th switch-over time between Q_i and Q_{i+1}.

$C_i^{(k)}$: The length of the k'th cycle *starting* at a visit beginning at Q_i, i.e.
$$C_i^{(k)} := \sum_{j=i}^{n} \left(V_j^{(k)} + S_j^{(k)} \right) + \sum_{j=1}^{i-1} \left(V_j^{(k+1)} + S_j^{(k+1)} \right).$$

$C_i^{*(k)}$: The length of the k'th cycle starting at a visit *completion* at Q_i, i.e.
$$C_i^{*(k)} := S_i^{(k)} + \sum_{j=i+1}^{n} \left(V_j^{(k)} + S_j^{(k)} \right) + \sum_{j=1}^{i-1} \left(V_j^{(k+1)} + S_j^{(k+1)} \right) + V_i^{(k+1)}.$$

$I_i^{(k)}$: The length of the k'th intervisit time of Q_i, i.e. $I_i^{(k)} := C_i^{(k)} - V_i^{(k)}$.

Additionally, let V_i, S_i, C_i, C_i^* and I_i denote their steady-state limits for $k \to \infty$, assuming these exist. For ease of notation we will write $V := V_1$, $S := S_1$, $C := C_1$, $C^* := C_1^*$ and $I := I_1$.

Lastly, let X_j^i denote the steady-state queue length of Q_j at a visit beginning at Q_i and write $\mathscr{F}_i(\boldsymbol{z})$, $\boldsymbol{z} = (z_1, ..., z_n)$, for the probability generating function (PGF) of the steady-state joint queue lengths at a visit beginning at Q_i.

3 Branching-type service disciplines

In this section we will consider polling systems using branching-type service disciplines. A service discipline is said to be of the branching-type when the system satisfies the following property [16]:

Property 1. *If there are k_i customers present at Q_i at the start of a server visit, then during the course of the visit, each of these k_i customers will effectively be replaced in an i.i.d. manner by a random population having PGF $h_i(z_1, ..., z_n)$.*

Note that while the exhaustive service discipline satisfies Property 1, the k-limited service discipline does not.

Consider now a subclass of polling systems which in addition to Property 1 satisfy the following

Property 2. $M_{i,j} = 0$ *for all $i \neq j$.*

Property 2 essentially puts certain restrictions on the branching functions $h_i(z_1, ..., z_n)$, as it implies the following

$$h_i(z_1, ..., z_{i-1}, 1, z_{i+1}, ..., z_n) = \theta_i \left(\sum_{j \neq i} \lambda_j (1 - z_j) \right), \tag{1}$$

where $\theta_i(\cdot)$ is the LST of the time that the server spends at Q_i due to the presence of one customer there, i.e. the time it spends serving that customer and its descendants during that visit, see [16] and [20] where this is referred to as a sub-busy period. For example, this prohibits that a customer, after it has received service, is routed to receive service at a different queue. However, it does permit that a customer immediately returns to the queue where it has just received service. Note that systems using a classical service discipline such as the exhaustive, gated and binomial service disciplines without routing all satisfy Property 2.

For systems satisfying Properties 1 and 2, we now relate the joint LST of $V_i + S_i$, $i = 1, ..., n$, to the joint PGF $\mathscr{F}_1(\cdot)$, extending Corollary 3.1 of [6], which deals with the LST of the cycle time $C = \sum_{i=1}^{n} (V_i + S_i)$. In the following subsections, we will use this result to determine the cycle time variance.

Proposition 1. *Write $\boldsymbol{\omega} = (\omega_1, ..., \omega_n)$. In case Properties 1 and 2 hold, the joint LST of $V_i + S_i$, $i = 1, ..., n$, is given by*

$$\mathbb{E} \left[e^{-\sum_{i=1}^{n} \omega_i (V_i + S_i)} \right] = \mathscr{F}_1(\theta_1(\gamma_1(\boldsymbol{\omega})), ..., \theta_n(\gamma_n(\boldsymbol{\omega}))) \prod_{i=1}^{n} \mathscr{S}_i(\gamma_i(\boldsymbol{\omega})), \tag{2}$$

where $\gamma_n(\boldsymbol{\omega}) = \omega_n$ and, for $1 \leq i \leq n-1$,

$$\gamma_i(\boldsymbol{\omega}) := \omega_i + \sum_{j=i+1}^{n} \lambda_j (1 - \theta_j(\gamma_j(\boldsymbol{\omega}))). \tag{3}$$

Proof. Let $\boldsymbol{X} = (X_1^1, ..., X_n^1)$ and $\boldsymbol{m} = (m_1, ..., m_n)$. It is sufficient to show that

$$\mathbb{E} \left[e^{-\sum_{i=1}^{n} \omega_i (V_i + S_i)} \,\middle|\, \boldsymbol{X} = \boldsymbol{m} \right] = \prod_{i=1}^{n} \mathscr{S}_i(\gamma_i(\boldsymbol{\omega})) \theta_i^{m_i}(\gamma_i(\boldsymbol{\omega})).$$

First, notice that, because of Properties 1 and 2, the PGF of the number of customers joining Q_j during a visit plus switch-over time of Q_i of length x, where $i \neq j$, is given by

$$\mathbb{E} \left[z^{X_j^{i+1} - X_j^i} \,\middle|\, V_i + S_i = x \right] = e^{-\lambda_j x (1 - z)},$$

which implies that the LST of the time the server spends at Q_j during its next visit because of these customers is given by $e^{-\lambda_j x (1 - \theta_j(\omega_j))}$.

This then allows us to write

$$\mathbb{E} \left[e^{-\sum_{i=1}^{n} \omega_i (V_i + S_i)} \,\middle|\, \boldsymbol{X} = \boldsymbol{m} \right]$$

$$= \mathscr{S}_n(\omega_n) \theta_n^{m_n}(\omega_n) \mathbb{E} \left[e^{-\sum_{i=1}^{n-1} (\omega_i + \lambda_n(1 - \theta_n(\omega_n)))(V_i + S_i)} \,\middle|\, \boldsymbol{X} = \boldsymbol{m} \right]$$

$$= \mathscr{S}_n(\omega_n) \theta_n^{m_n}(\omega_n) \mathscr{S}_{n-1}(\gamma_{n-1}(\boldsymbol{\omega})) \theta_{n-1}^{m_{n-1}}(\gamma_{n-1}(\boldsymbol{\omega}))$$

$$\cdot \mathbb{E} \left[e^{-\sum_{i=1}^{n-2} (\omega_i + \lambda_{n-1}(1 - \theta_{n-1}(\gamma_{n-1}(\boldsymbol{\omega}))) + \lambda_n(1 - \theta_n(\omega_n)))(V_i + S_i)} \,\middle|\, \boldsymbol{X} = \boldsymbol{m} \right].$$

Performing this procedure n times and deconditioning gives the desired result. $\qquad\square$

In Appendix A, we generalize Proposition 1 to systems where the restrictive Property 2 is relaxed. However, it is not at all straightforward to extend *all* other results of this section to this more general case.

Note that by substituting $\omega_i = u$ for $i = 1, ..., n$ in Equation (2), we readily find the LST of the cycle time C, as shown by the following corollary.

Corollary 1. *The LST of the cycle time is given by*

$$\mathbb{E}\left[e^{-uC}\right] = \mathscr{F}_1(\theta_1(\gamma_1(u, ..., u)), ..., \theta_n(\gamma_n(u, ..., u))) \prod_{i=1}^{n} \mathscr{S}_i(\gamma_i(u, ..., u)).$$

In the remainder of this paper, we will focus on symmetric systems, i.e. $\lambda_i = \lambda = \Lambda/n$, $s_i = s$, $s_i^{(2)} = s^{(2)}$, $\beta_i = \beta$ and $\beta_i^{(2)} = \beta^{(2)}$, $h_i(z_1, ..., z_n) = h_1(z_i, ..., z_n, z_1, .., z_{i-1})$ and $\theta_i(u) = \theta(u)$ for all $i = 1, ..., n$. Additionally, we assume that the $M_{i,j}$ are independent and identically distributed for all i and $j \neq i$, i.e. all queues that are not being served are treated equally. In Subsection 3.2 we will use Proposition 1 to derive $r_i := \mathbb{E}[V_i + S_i]$ and $r_{i,j} := \text{Cov}[V_i + S_i, V_j + S_j]$, and obtain an explicit expression for $\text{Var}[C]$. However, in preparation for this, we first determine in Subsection 3.1 explicit expressions for $\mathbb{E}[X_i^1]$ and $\text{Cov}[X_i^1, X_j^1]$.

3.1 Queue length covariance

In this subsection we will derive $l_i := \mathbb{E}[X_i^1]$ and $l_{i,j} := \mathbb{E}[X_i^1 X_j^1]$ for polling systems satisfying Property 1, which will give us $\text{Var}[X_i^1]$ and $\text{Cov}[X_i^1, X_j^1]$.

Recall that $\mathscr{F}_i(\boldsymbol{z})$ is the PGF of the joint queue lengths at a visit beginning at Q_i. In [16] it is shown that the following relation holds for all polling systems satisfying Property 1:

$$\mathscr{F}_{i+1}(\boldsymbol{z}) = \mathscr{F}_i(z_1, \ldots, z_{i-1}, h_i(\boldsymbol{z}), z_{i+1}, \ldots, z_n) \mathscr{S}_i\left(\sum_{j=1}^{n} \lambda_j(1 - z_j)\right).$$

For a symmetric system, this gives

$$\mathscr{F}_1(\boldsymbol{z}) = \mathscr{F}_1(h_1(z_n, z_1, \ldots, z_{n-1}), z_1, \ldots, z_{n-1}) \mathscr{S}\left(\sum_{j=1}^{n} \lambda(1 - z_j)\right). \tag{4}$$

We define

$$\frac{\partial}{\partial z_i} h_1(\boldsymbol{z})\Big|_{\boldsymbol{z}=(1,...,1)} := \begin{cases} \phi, & i = 1, \\ \psi, & i \neq 1. \end{cases}$$

In the polling literature $1 - \phi$ is called the *exhaustiveness* of a (branching-type) service discipline, introduced by Van der Mei and Levy [20].

Differentiating (4) with respect to z_i, we find

$$l_i = \begin{cases} l_{i+1} + \psi l_1 + \lambda s, & i \neq n, \\ \phi l_1 + \lambda s, & i = n. \end{cases}$$

Solving gives

$$l_i = \frac{ns\lambda}{1 - \phi - (n-1)\psi}\left(1 - \frac{i-1}{n}(1 - \phi + \psi)\right). \tag{5}$$

Note that the stability condition for the system is $\phi + (n-1)\psi < 1$, see [16]. For the most common service disciplines satisfying Properties 1 and 2 including (binomial) gated and (binomial) exhaustive, we have $\phi = y + \psi$ and $\psi = \mu\Lambda/n$ for $y := \mathbb{E}[M_{1,1}]$ (which is zero for standard gated and exhaustive service) and $\mu := -\frac{d}{du}\theta(u)|_{u=0}$. For the systems considered in this paper, neither y

nor μ depends on n, meaning that the total workload arriving to the system per time unit is equal to $\phi + (n-1)\psi = y + \mu\Lambda$, which does not depend on n either. It also follows that l_1 does not depend on n.

Now, define

$$\frac{\partial^2}{\partial z_i \partial z_j} h_1(z)\Big|_{z=(1,\ldots,1)} := \begin{cases} \phi^{(2)}, & i=1, j=1, \\ \psi_1^{(2)}, & i=j, i \neq 1, \\ \psi_2^{(2)}, & i \neq 1, j \neq 1, i \neq j, \\ \chi, & i=1, j \neq 1. \end{cases}$$

Differentiating Equation (4) with respect to z_i and z_j gives for $i \neq n$, $j \neq n$, $i \neq j$:

$$l_{i,j} = l_{i+1,j+1} + \psi(l_{1,i+1} + l_{1,j+1}) + \psi^2 l_{1,1} + \lambda s(l_{i+1} + l_{j+1}) \\ + (\psi_2^{(2)} - \psi^2 + 2\lambda s\psi)l_1 + \lambda^2 s^{(2)}; \tag{6}$$

for $i < n$ and $j = n$:

$$l_{i,n} = \phi l_{1,i+1} + \phi\psi l_{1,1} + \lambda s l_{i+1} + (\chi - \phi\psi + \lambda s(\phi + \psi))l_1 + \lambda^2 s^{(2)}; \tag{7}$$

for $i = j$ and $i < n$:

$$l_{i,i} = l_i - l_{i+1} + l_{i+1,i+1} + 2\psi l_{1,i+1} + \psi^2 l_{1,1} + 2\lambda s l_{i+1} \\ + (\psi_1^{(2)} - \psi^2 + 2\lambda s\psi)l_1 + \lambda^2 s^{(2)}; \tag{8}$$

and finally for $i = j = n$:

$$l_{n,n} = l_n + \phi^2 l_{1,1} + (\phi^{(2)} - \phi^2 + 2\lambda s\phi)l_1 + \lambda^2 s^{(2)}. \tag{9}$$

Consider Equation (6). By subtracting $l_{i,j+1}$, we find for $j+1 \neq n$ and $i \neq j+1$:

$$l_{i,j} - l_{i,j+1} = l_{i+1,j+1} - l_{i+1,j+2} + \psi(l_{1,j+1} - l_{1,j+2}) + \lambda s(l_{j+1} - l_{j+2}).$$

Notice that the last term is independent of j, suggesting that $l_{i,j}$ depends linearly on j for fixed $i < j$. Similarly, subtracting $l_{i+1,j}$ from both sides of Equation (6) gives

$$l_{i,j} - l_{i+1,j} = l_{i+1,j+1} - l_{i+2,j+1} + \psi(l_{1,i+1} - l_{1,i+2}) + \lambda s(l_{i+1} - l_{i+2}),$$

suggesting that $l_{i,j}$ also depends linearly on i for fixed $j > i$. Together with Equations (7) and (9), assuming $l_{i,j} = a_i + b_j(i-1) + c_i(j-1)$ then allows us to express b_i and c_i in terms of b_1, c_1 and $l_{1,1}$. Similarly, rewriting Equation (8), we have

$$l_{i,i} - l_{i+1,i+1} = l_i - l_{i+1} + 2\psi l_{1,i+1} + \psi^2 l_{1,1} + 2\lambda s l_{i+1},$$

which suggests that $l_{i,i}$ is a quadratic function in i. Writing $l_{i,i} = l_{1,1} + b(i-1) + c(i-1)^2$ and substituting then allows us to solve the resulting equations and determine the remaining unknowns, confirming the linear and quadratic relations alluded to above.

We find, after a substantial amount of algebraic manipulation and simplification, that the complete solution is of the form

$$\text{Var}[X_i^1] = l_{i,i} - l_i^2 = \alpha_1 + \alpha_2 - \alpha_3((i-1)(\phi-\psi) - (n-i+1)(1-(i-1)\psi)) - \alpha_4(i-1), \tag{10}$$

and, for $i < j$,

$$\text{Cov}[X_i^1, X_j^1] = l_{i,j} - l_i l_j = \alpha_1 - \alpha_3((i-1)(\phi-\psi) - (n-j+1)(1-(i-1)\psi)), \tag{11}$$

where

$$\alpha_1 = \frac{n\lambda^2(\phi - \psi)(s^{(2)} - s^2)}{(1+\phi)(1-\phi-(n-1)\psi)}$$

$$+ \frac{ns\lambda(\phi - \psi)\psi(\phi^{(2)} + (n-1)\psi_1^{(2)})}{(1+\phi)(1-\phi+\psi)(1-\phi-(n-1)\psi)^2}$$

$$+ \frac{ns\lambda\chi(1 - \phi^2 - (n-1)\psi^2 - n\psi(1-\phi))}{(1+\phi)(1-\phi+\psi)(1-\phi-(n-1)\psi)^2}$$

$$- \frac{ns\lambda\psi_2^{(2)}(1 - \phi^2 - (n-1)\psi^2 - n\phi(1-\phi))}{(1+\phi)(1-\phi+\psi)(1-\phi-(n-1)\psi)^2}, \tag{12}$$

$$\alpha_2 = \frac{ns\lambda(1+\psi)(1 + \phi^{(2)} - \chi + (n-1)\left(\psi_1^{(2)} - \psi_2^{(2)}\right))}{(1+\phi)(1-\phi+\psi)(1-\phi-(n-1)\psi)}$$

$$+ \frac{ns\lambda\phi(\psi - \phi + \psi_2^{(2)} - \chi)}{(1+\phi)(1-\phi+\psi)(1-\phi-(n-1)\psi)}, \tag{13}$$

$$\alpha_3 = \frac{ns\lambda\left(\psi^2\phi^{(2)} + 2(1-\phi)\psi\chi + (1-\phi)^2\psi_2^{(2)} + (n-1)\psi^2\left(\psi_1^{(2)} - \psi_2^{(2)}\right)\right)}{(1+\phi)(1-\phi+\psi)(1-\phi-(n-1)\psi)^2}$$

$$+ \frac{\lambda^2(1-\phi+\psi)(s^{(2)} - s^2)}{(1+\phi)(1-\phi-(n-1)\psi)}, \tag{14}$$

and

$$\alpha_4 = \frac{s\lambda\left(1 - \phi + \psi + n(\psi_1^{(2)} - \psi_2^{(2)})\right)}{1 - \phi - (n-1)\psi}.$$

3.2 Station-time covariance

We will now use Proposition 1 to derive the mean $r_i := \mathbb{E}[V_i + S_i]$ and covariance $r_{i,j} := \mathrm{Cov}[V_i + S_i, V_j + S_j]$ of the station time (sum of the visit time at a queue and switch-over time to the next queue), expressing them in terms of $\mathrm{Var}[X_i^1]$ and $\mathrm{Cov}[X_i^1, X_j^1]$, cf. Equations (10) and (11). This will then allow us to give an explicit expression for $\mathrm{Var}[C]$. Since we will rely on Proposition 1, we restrict ourselves to systems satisfying both Properties 1 and 2 for the remainder of this section.

Define $\mu^{(2)} := \frac{\mathrm{d}^2}{\mathrm{d}u^2}\theta(u)\big|_{u=0}$. First, note that by differentiating (3), we have

$$\Gamma_i^j := \frac{\partial}{\partial\omega_i}\gamma_j(\boldsymbol{\omega})\bigg|_{\boldsymbol{\omega}=(0,\dots,0)} = \mathbb{1}[i=j] + \mu\lambda\sum_{k=j+1}^{i}\Gamma_i^k.$$

Solving gives

$$\Gamma_i^j = \begin{cases} 1, & i = j, \\ \mu\lambda(1+\mu\lambda)^{i-j-1}, & j < i, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Differentiating (2) with respect to ω_i yields, as expected,

$$r_i = \sum_{j=1}^{i}(s + \mu l_j)\Gamma_i^j = s + \mu l_1,$$

where the second equality follows from the fact that for service disciplines satisfying Property 2 we have $\psi = \mu\lambda$, see Equation (1). Additionally, it follows that

$$\mathbb{E}[C] = nr_i = \frac{ns(1-\phi+\psi)}{1-\phi-(n-1)\psi}.$$

Differentiating (2) with respect to ω_1 and ω_i and subtracting $r_1 r_i = \sum_{j=1}^{i} \Gamma_i^j (s + \mu l_1)(s + \mu l_j)$ from both sides gives

$$r_{1,i} = \Gamma_i^1 (s^{(2)} - s^2 + l_1(\mu^{(2)} - \mu^2)) + \mu^2 \sum_{j=1}^{i} \Gamma_i^j \left(l_{1,j} - l_1 l_j \right).$$

Using Equation (15), some manipulation yields

$$r_{1,i} = \begin{cases} s^{(2)} - s^2 + l_1(\mu^{(2)} - \mu^2) + \mu^2 \mathrm{Var}[X_1^1], & i = 1, \\ \mu \lambda r_{1,1} + \mu^2 \mathrm{Cov}[X_1^1, X_2^1], & i = 2, \\ (1 + \mu \lambda) r_{1,i-1} - \mu^2 \mathrm{Cov}[X_1^1, X_{i-1}^1 - X_i^1], & i > 2. \end{cases} \tag{16}$$

Equation (11) then tells us that $\mathrm{Cov}[X_1^1, X_{i-1}^1 - X_i^1] = \alpha_3$ for all $i > 2$, hence

$$r_{1,i} = \begin{cases} s^{(2)} - s^2 + \dfrac{n s \lambda (\mu^{(2)} - \mu^2)}{1 - \phi - (n-1)\psi} + \mu^2 (\alpha_1 + \alpha_2 + n\alpha_3), & i = 1, \\ \dfrac{\mu}{\lambda} \alpha_3 + \dfrac{\mu}{\lambda} (1 + \mu \lambda)^{i-2} (\lambda^2 r_{1,1} + \alpha_1 \mu \lambda - (1 - (n-1)\mu\lambda)\alpha_3), & i \ge 2. \end{cases} \tag{17}$$

Note that in general it may not hold that $r_{1,i} > 0$. As we will see in the next section however, $r_{1,i}$ will be positive for traditional branching-type service policies, such as the binomial-gated and binomial-exhaustive service disciplines, as n grows large.

Finally, we find an explicit formula for the variance of the cycle time as follows

$$\begin{aligned} \mathrm{Var}[C] &= \sum_{i=1}^{n} \sum_{j=1}^{n} r_{i,j} = n r_{1,1} + 2 \sum_{i=1}^{n-1} (n-i) r_{1,1+i} \\ &= 2(1 + \mu\lambda)^n \left(\frac{r_{1,1}}{\mu\lambda} + \frac{\alpha_1}{\lambda^2} - \frac{\alpha_3}{\mu\lambda^3}(1 - (n-1)\mu\lambda) \right) - \frac{2 + n\mu\lambda}{\mu\lambda} r_{1,1} \\ &\quad - \frac{2(1 + n\mu\lambda)}{\lambda^2} \alpha_1 + \frac{2 + \mu\lambda(2 - n(n-1)\mu\lambda)}{\lambda^3 \mu} \alpha_3. \end{aligned} \tag{18}$$

Remark 1. *For some service disciplines, for example exhaustive service (cf. [2, 22]), it is more natural to consider C^*, the time between successive visit completions at Q_1. Obviously it holds that $\mathbb{E}[C] = \mathbb{E}[C^*]$, but higher moments generally differ. For the variance, it is readily seen that*

$$\mathrm{Var}[C^*] = \mathrm{Var}[S_1 + V_2 + S_2 + \cdots + S_n + V_{n+1}] = \mathrm{Var}[C] + 2 \sum_{i=1}^{n} \mathrm{Cov}[S_1, V_{i+1}].$$

Following the exact same steps as in the proof of Proposition 1 it is possible to find the joint LST of $S_1, V_2, S_2, \ldots, S_n, V_{n+1}$ which, after differentiation, yields

$$\mathrm{Cov}[S_1, V_{i+1}] = (s^{(2)} - s^2)\lambda\mu(1 + \lambda\mu)^{i-1},$$

for $i = 1, \ldots, n$. Finally, we obtain

$$\mathrm{Var}[C^*] = \mathrm{Var}[C] + 2(s^{(2)} - s^2)\big((1 + \lambda\mu)^n - 1\big), \tag{19}$$

which is strictly greater than $\mathrm{Var}[C]$ unless $\mathrm{Var}[S] = 0$.

3.3 Many-queue asymptotics

We will now apply the results of the previous subsections to analyze the limiting behavior of symmetric polling systems satisfying Properties 1 and 2 as the number of queues grows large. To this end, we will consider a sequence of polling systems where the total arrival rate is kept constant, i.e.

we now assume $\lambda = \Lambda/n$ with Λ fixed, so $\lambda \downarrow 0$ as $n \to \infty$. The goal is to give an explicit expression for the limit of the scaled cycle time variance $n\mathrm{Var}[C/n]$.

First, define the following limiting values

$$\lim_{n\to\infty} \phi = \Phi, \quad \lim_{n\to\infty} n\psi = \Psi, \quad \lim_{n\to\infty} \mu = m, \quad \lim_{n\to\infty} \mu^{(2)} = m^{(2)}. \tag{20}$$

Additionally, define

$$\lim_{n\to\infty} \phi^{(2)} = \Phi^{(2)}, \quad \lim_{n\to\infty} n\psi_1^{(2)} = \Psi_1^{(2)}, \quad \lim_{n\to\infty} n^2\psi_2^{(2)} = \Psi_2^{(2)}, \quad \lim_{n\to\infty} n\chi = \mathscr{X}. \tag{21}$$

Note that the limits in (20) should exist by the stability condition $\phi + (n-1)\psi < 1$. However, the stability condition does not guarantee the existence of the limits in (21), but their existence is necessary for the existence of a finite limiting value of $n\mathrm{Var}[C/n]$.

Additionally, since we limit ourselves to service disciplines satisfying Property 2, we know from Equation (1) that $\Psi = m\Lambda$ and $\Psi_1^{(2)} = 0$.

Remark 2. *Note that $\Phi^{(2)}$, $\Psi_2^{(2)}$ and $\mathscr{X}$ need not be zero, as illustrated by the following example. Consider a system in which the queues get served according to the usual gated service discipline, but whenever a customer completes its service, Y new customers join the same queue at which the customer received its service, where Y is a non-negative integer-valued random variable with PGF $G_Y(z)$. The branching function is then given by*

$$h(\mathbf{z}) = G_Y(z_1)\mathscr{B}\left(\frac{\Lambda}{n}\sum_{i=1}^{n}(1-z_i)\right).$$

Consequently, for such a system, we find $\Phi^{(2)} = \mathbb{E}[Y(Y-1)]$, $\Psi_2^{(2)} = \Lambda^2 m^{(2)}$, and $\mathscr{X} = \Lambda m\mathbb{E}[Y]$ with stability condition $\mathbb{E}[Y] + \Lambda m < 1$.

We will now derive the limiting value of $n\mathrm{Var}[C/n]$ as n grows large. From Equations (3.2) and (5) we find

$$\lim_{n\to\infty} \mathbb{E}[C/n] = \frac{s}{1-\Phi-\Psi},$$

$$\lim_{n\to\infty} l_{\lfloor xn \rfloor} = \frac{\Lambda s}{1-\Phi-\Psi}(1-x(1-\Phi)).$$

Furthermore, write

$$A_1 := \lim_{n\to\infty} n\alpha_1 = \frac{\Lambda^2\Phi(s^{(2)}-s^2)}{(1+\Phi)(1-\Phi-\Psi)} + \frac{\Lambda s(1+\Phi-\Psi)\mathscr{X}}{(1+\Phi)(1-\Phi-\Psi)^2}$$

$$+ \frac{\Lambda s\Psi\Phi\Phi^{(2)}}{(1-\Phi^2)(1-\Phi-\Psi)^2} + \frac{\Lambda s\Phi\Psi_2^{(2)}}{(1+\Phi)(1-\Phi-\Psi)^2},$$

$$A_2 := \lim_{n\to\infty} \alpha_2 = \frac{\Lambda s}{1-\Phi-\Psi}\left(1 + \frac{\Phi^{(2)}}{1-\Phi^2}\right),$$

$$A_3 := \lim_{n\to\infty} n^2\alpha_3 = \frac{\Lambda^2(1-\Phi)(s^{(2)}-s^2)}{(1+\Phi)(1-\Phi-\Psi)} + \frac{2\Lambda s\Psi\mathscr{X}}{(1+\Phi)(1-\Phi-\Psi)^2}$$

$$+ \frac{\Lambda s\Psi^2\Phi^{(2)}}{(1-\Phi^2)(1-\Phi-\Psi)^2} + \frac{2\Lambda s\Psi_2^{(2)}(1-\Phi)}{(1+\Phi)(1-\Phi-\Psi)^2},$$

and

$$A_4 := \lim_{n\to\infty} n\alpha_4 = \frac{\Lambda s(1-\Phi)}{1-\Phi-\Psi}.$$

Then taking limits in Equations (10) and (11) we find

$$\lim_{n\to\infty} \mathrm{Var}[X^1_{\lfloor xn\rfloor}] = A_2 - xA_4,$$

and, for $x \neq y$,

$$\lim_{n\to\infty} n\mathrm{Cov}[X^1_{\lfloor xn\rfloor}, X^1_{\lfloor yn\rfloor}] = A_1 + (1 - x\Phi - y(1 - \Psi x))A_3.$$

Similarly,

$$R_{1,1} := \lim_{n\to\infty} r_{1,1} = s^{(2)} - s^2 + \frac{s\Lambda}{1 - \Phi - \Psi}(m^{(2)} - m^2) + m^2 A_2,$$

and

$$\lim_{n\to\infty} n r_{1,\lfloor xn\rfloor} = \frac{m}{\Lambda}A_3 + \frac{m}{\Lambda}(\Lambda^2 R_{1,1} + m\Lambda A_1 - (1 - m\Lambda)A_3)e^{m\Lambda x}.$$

Finally, for the scaled variance of the cycle time we find

$$\begin{aligned}
\lim_{n\to\infty} n\mathrm{Var}[C/n] = {} & \frac{1}{m\Lambda^3}((2 - m^2\Lambda^2)A_3 - \Lambda^2(2 + m\Lambda)R_{1,1} - 2m\Lambda(1 + m\Lambda)A_1) \\
& + \frac{2}{m\Lambda^3}(\Lambda^2 R_{1,1} + m\Lambda A_1 - (1 - m\Lambda)A_3)e^{m\Lambda}.
\end{aligned} \tag{22}$$

For C^*, the cycle time starting at a visit completion, taking $n \to \infty$ in Equation (19) gives

$$\lim_{n\to\infty} (\mathrm{Var}[C^*] - \mathrm{Var}[C]) = 2(s^{(2)} - s^2)(e^{m\Lambda} - 1),$$

from which it follows that

$$\lim_{n\to\infty} n\mathrm{Var}[C^*/n] = \lim_{n\to\infty} n\mathrm{Var}[C/n].$$

Remark 3. *It is noteworthy that for some service disciplines, the mean waiting time can be expressed in terms of the mean residual cycle time. For example, with $\rho_1 = \Lambda\beta/n$, we have*

$$\mathbb{E}[W_{gated}] = (1 + \rho_1)\left(\frac{\mathrm{Var}[C]}{2\mathbb{E}[C]} + \frac{\mathbb{E}[C]}{2}\right),$$

$$\mathbb{E}[W_{exhaustive}] = (1 - \rho_1)\left(\frac{\mathrm{Var}[C^*]}{2\mathbb{E}[C]} + \frac{\mathbb{E}[C]}{2}\right).$$

As $n \to \infty$, we see that in both cases the scaled mean waiting time W/n converges to a scaled mean residual cycle time:

$$\lim_{n\to\infty} \mathbb{E}[W_{gated}/n] = \lim_{n\to\infty} (1 + \Lambda\beta/n)\left(\frac{\mathrm{Var}[C/n]}{2\mathbb{E}[C/n]} + \frac{\mathbb{E}[C/n]}{2}\right) = \frac{s}{2(1 - \rho)},$$

$$\lim_{n\to\infty} \mathbb{E}[W_{exhaustive}/n] = \lim_{n\to\infty} (1 - \Lambda\beta/n)\left(\frac{\mathrm{Var}[C^*/n]}{2\mathbb{E}[C/n]} + \frac{\mathbb{E}[C/n]}{2}\right) = \frac{s}{2(1 - \rho)}.$$

3.4 Binomial service

As an example we will examine the binomial-gated and binomial-exhaustive service disciplines. Under the binomial-gated service discipline, when the server finds m customers present at the start of a visit period of a queue, it will serve $N \sim \mathrm{Bin}(m, p)$, $0 < p \leq 1$, of these customers before switching to the next queue. Under the binomial-exhaustive service discipline, the server not only serves these N customers but also any additional customers that arrive to the queue during its visit. The branching functions for the binomial-gated and binomial-exhaustive disciplines are given by

$$h(z) = (1 - p)z_1 + p\mathscr{B}\left(\frac{\Lambda}{n}\sum_{i=1}^{n}(1 - z_i)\right)$$

and

$$h(z) = (1-p)z_1 + p\mathscr{P}_n\left(\frac{\Lambda}{n}\sum_{i=2}^{n}(1-z_i)\right),$$

respectively. Here $\mathscr{B}(\cdot)$ is the LST of the service time distribution as before and $\mathscr{P}_n(\cdot)$ the LST of a busy-period distribution in an M/G/1 queue with arrival rate Λ/n and the same service time distribution.

We now make the observation that for both policies we have the following limits: $m = p\beta$, $m^{(2)} = p\beta^{(2)}$, $\Phi = 1-p$, $\Psi = p\Lambda\beta$, $\Phi^{(2)} = \Psi_1^{(2)} = \mathscr{X} = 0$ and $\Psi_2^{(2)} = p\Lambda^2\beta^{(2)}$. Hence, for the second moment of the cycle time, Equation (22) tells us that in the limit for $n \to \infty$ we see no distinction between the binomial-gated or binomial-exhaustive service disciplines.

We find, with $\rho = \Lambda\beta$,

$$R_{1,1} = s^{(2)} - s^2 + \frac{\Lambda s}{1-\rho}\beta^{(2)}$$

and

$$\lim_{n\to\infty} n\mathrm{Var}[C/n] = \frac{p}{(2-p)(1-\rho)}R_{1,1} + \frac{2(1-p)}{p(2-p)\rho}(e^{p\rho}-1)R_{1,1}. \tag{23}$$

Also note that, using Equation (16), one can show that, for $x > 0$,

$$\lim_{n\to\infty} nr_{1,\lfloor xn\rfloor} = \frac{1+(1-p)(1-\rho)e^{p\rho x}}{(2-p)(1-\rho)}\left(p\rho(s^{(2)}-s^2) + \frac{p\rho s\Lambda\beta^{(2)}}{1-\rho}\right).$$

This shows that in the special case $p = 1$ this limit does not depend on x and $\lim_{n\to\infty} nr_{1,i} = \lim_{n\to\infty} nr_{1,j}$ for all $i \neq 1$ and $j \neq 1$. For other values of p the covariance of the visit times increases as a function of x.

Besides giving an explicit expression for the limiting value of the scaled cycle time variance, Equation (23) has another interesting consequence. It reveals that, since $\mathrm{Var}[C/n]$ is of order $1/n$, the scaled cycle time C/n converges in probability to a deterministic value equal to $s/(1-\rho)$. Therefore, denoting by $X_i^i(j)$ the queue length of Q_i at the start of the j'th server visit to Q_i, this suggests that as $n \to \infty$:

$$X_i^i(j+1) \stackrel{d}{=} Y_i(X_i^i(j),j) + A_i(j),$$

where $Y_i(m_j,j) \sim \mathrm{Bin}(m_j,1-p)$ are independent binomially distributed random variables with parameters m_j and $1-p$, and $A_i(j)$ are independent Poisson distributed random variables with parameter $\Lambda s/(1-\rho)$, $j = 1,2,\dots$. In particular, letting $j \to \infty$, we conclude that the steady-state queue length X_i^i at the start of a visit to Q_i satisfies the distributional property

$$X_i^i \stackrel{d}{=} Y_i(X_i^i) + A_i, \tag{24}$$

where $Y_i(m) \sim \mathrm{Bin}(m,1-p)$ is a Binomially distributed random variable with parameters m and $1-p$, and A_i is a Poisson distributed random variable with parameter $\Lambda s/(1-\rho)$.

Now observe that if X is Poisson distributed with parameter α and given $X = x$, the random variable Y is binomially distributed with parameters x and q, then Y is Poisson distributed with parameter αq. Hence, we conclude from (24) that X_i^i is Poisson distributed with parameter γ satisfying $\gamma = (1-p)\gamma + \Lambda s/(1-\rho)$, i.e., $\gamma = \frac{\Lambda s}{p(1-\rho)}$:

$$X_i^i \sim \mathrm{Poi}\left(\frac{\Lambda s}{p(1-\rho)}\right). \tag{25}$$

In fact, as will also be seen in the following sections, this convergence of the scaled cycle time suggests that we can approximate the steady-state distribution of the queue length at the beginning of a visit for many service disciplines, not necessarily of branching-type, by examining a relation of the form

$$X_i^i \stackrel{d}{=} f(X_i^i) + A_i, \tag{26}$$

where the two terms on the right-hand side are independent, the (possibly random) function $f(\cdot)$ is determined by the actual service discipline, and A_i is a Poisson distributed random variable with parameter $\frac{\Lambda s}{1-\rho}$. For example, in the case of exhaustive service $f(x) = 0$ and in the case of k-limited service $f(x) = (x-k)^+$, as will be seen in Section 4.2.

This means that, in the limit, for many large-scale polling systems the analysis of the marginal queue length distribution simplifies to the analysis of a simpler discrete-time queueing model with i.i.d. arrivals. Often, such a model is much easier to analyze than the pre-limit polling system.

Remark 4. *In the case of branching-type service, where each customer is treated equally, one can write* $f(X_i^i) = \sum_{j=1}^{X_i^i} U_j$, *where the U_j are i.i.d. random variables. For this case, rewriting (26) in terms of PGFs $G_{X_i^i}(z)$ and $G_U(z)$ and iterating, gives*

$$G_{X_i^i}(z) = G_{X_i^i}(G_U(z)) e^{\frac{\Lambda s}{1-\rho}(z-1)} = \lim_{j\to\infty} G_{X_i^i}(G_U^{(j)}(z)) e^{\frac{\Lambda s}{1-\rho}\sum_{k=0}^{j-1}\left(G_U^{(k)}-1\right)},$$

where $G_U^{(j)}(z) = G_U(G_U^{(j-1)}(z))$ denotes the j'th iterate of $G_U(z)$. Assuming that $\mathbb{E}[U] < 1$, which is necessary for stability of the system, we have that $\lim_{j\to\infty} G_U^{(j)}(z) = 1$, which gives

$$G_{X_i^i}(z) = e^{\sum_{k=0}^{\infty} \frac{\Lambda s}{1-\rho}\left(G_U^{(k)}(z)-1\right)}.$$

We close this subsection by considering the mean waiting time for binomial-gated service. As shown in [14],

$$\mathbb{E}[W] = \frac{l_{1,1}}{l_1} \frac{n(2-p)+p\rho}{2\Lambda},$$

which gives the following limit

$$\lim_{n\to\infty} \mathbb{E}[W/n] = \left(\frac{1}{2} + \frac{1-p}{p}\right) \frac{s}{1-\rho}. \tag{27}$$

Indeed, in the limit for $n \to \infty$, the scaled waiting time is given by a residual cycle time plus an additional geometrically distributed number of cycle times. Note that the same limit holds for binomial-exhaustive service, since the probability that a customer joins a queue while the server is serving that queue, is negligible for large n.

Examining Equations (23) and (27), we see that when choosing p, there is actually a trade-off between a small variance of the cycle time and short mean waiting times. For example, consider the case in which both the service times and switch-over times are exponentially distributed with $\beta = s = 1$ and $\Lambda = 4/5$. Then setting $p = 2/3$ will give a mean waiting time of one full cycle compared to half a cycle for $p = 1$. However, it will also reduce the scaled variance of the cycle time from 48 to 28.4.

It is important to note, however, that this increase of the waiting times is of order n and the decrease of the standard deviation of the unscaled cycle time is of order $\sqrt{n}$. While in very large systems this trade-off might not seem so attractive, it could be of importance in somewhat smaller systems with deadline-critical applications, requiring a predictable return time of the server.

4 General non-idling service disciplines

We will now briefly consider general non-idling service disciplines, which are not necessarily of the branching-type. The goal of this section is to show that in general C/n may be expected to converge to a deterministic quantity for many polling systems, even if they do not satisfy Property 1.

We continue to concentrate on symmetric systems with $\lambda = \Lambda/n$ and throughout this section we assume that Property 2 holds, i.e. $M_{i,j} = 0$ for all i and j. Let L, L^b and L^e denote the steady-state

queue length at Q_1 at an arbitrary epoch, a visit beginning at Q_1 and a visit completion at Q_1, respectively. Furthermore, for a random variable U with CDF $F(u)$ and finite mean and second moment, let U^R denote the overshoot of U with PDF $(1 - F(u))/\mathbb{E}[U]$ and mean $\mathbb{E}[U^R] = \mathbb{E}[U^2]/(2\mathbb{E}[U])$.

The pseudo-conservation law for cyclic service systems [5] gives:

$$\rho \mathbb{E}[W] = \frac{\rho \Lambda \beta^{(2)}}{2(1-\rho)} + \frac{\rho(s^{(2)} - s^2) + \rho n s^2}{2s} + \frac{\rho^2 n s}{2(1-\rho)} - \frac{\rho^2 s}{2(1-\rho)} + n\beta \mathbb{E}[L^e]. \tag{28}$$

Applying Little's law and simplifying shows

$$\mathbb{E}[L - L^e] = \frac{\Lambda}{n}\left(\frac{\rho}{1-\rho}\frac{\beta^{(2)}}{2\beta} + \frac{s^{(2)}}{2s} - \frac{s}{2(1-\rho)}\right) + \frac{\Lambda s}{2(1-\rho)}. \tag{29}$$

Consider now a tagged customer arriving at Q_1 during an intervisit period. We define L^I to be the queue length of Q_1 at an arbitrary epoch during an intervisit time and let $\tilde{L}^b$ and $\tilde{L}^e$ be the queue length of Q_1 at the start and end of the server visit contained in the cycle in which the customer arrives. Conditioning on the length of the intervisit period in which the tagged customer arrived, we find

$$\mathbb{E}[\tilde{L}^e] = \frac{1}{\mathbb{E}[I]}\int_{x=0}^{\infty} x \mathbb{E}[L^e \mid I = x]\,\mathrm{d}\mathbb{P}[I \leq x] = \frac{\mathbb{E}[L^e I]}{\mathbb{E}[I]} = \frac{\mathrm{Cov}[L^e, I]}{\mathbb{E}[I]} + \mathbb{E}[L^e]. \tag{30}$$

Also note that a similar relationship holds for $\mathbb{E}[\tilde{L}^b]$. The PASTA property implies that the expected queue length of Q_1 just before the arrival of the tagged customer equals $\mathbb{E}[L^I]$. Considering the change in the queue length of Q_1 since the last visit completion at Q_1, we can write

$$\mathbb{E}[L^I] = \mathbb{E}[\tilde{L}^e] + \frac{\Lambda}{n}\mathbb{E}[I^R] = \left(\frac{\mathrm{Cov}[L^e, I]}{\mathbb{E}[I]} + \mathbb{E}[L^e]\right) + \frac{\Lambda}{n}\mathbb{E}[I^R]. \tag{31}$$

Moreover, the Fuhrmann-Cooper decomposition [10] tells us that

$$\mathbb{E}[L^I] = \mathbb{E}[L] - \frac{\Lambda}{n}\frac{\Lambda\beta^{(2)}}{2n(1-\rho/n)}. \tag{32}$$

Equating (31) and (32) and multiplying by n, we find

$$n\mathbb{E}[L - L^e] - \frac{\Lambda}{n}\frac{\Lambda\beta^{(2)}}{2(1-\rho/n)} - \frac{\mathrm{Cov}[L^e, I]}{\mathbb{E}[I/n]} = \frac{\Lambda}{2}\left(\frac{n\mathrm{Var}[I/n]}{\mathbb{E}[I/n]} + \mathbb{E}[I]\right),$$

which, together with Equation (29), gives

$$\lim_{n\to\infty} n\mathrm{Var}[C/n] = \frac{\Lambda s}{(1-\rho)^2}\beta^{(2)} + \frac{s^{(2)} - s^2}{1-\rho} - \frac{2}{\Lambda}\lim_{n\to\infty}\mathrm{Cov}[L^e, I]. \tag{33}$$

Hence, we can express the asymptotic variance of the scaled cycle time in terms of $\mathrm{Cov}[L^e, I]$. Therefore, if one is able to show that the limit of $\mathrm{Cov}[L^e, I]$ exists, this implies that C/n will converge in probability to a deterministic value, as was the case for binomial service disciplines.

4.1 Binomial service

Consider again the binomial-gated and binomial-exhaustive service disciplines of Section 3.4. Combining Equations (23) and (33) we find

$$\lim_{n\to\infty}\mathrm{Cov}[L^e, I] = \Lambda\left(\frac{1-p}{(2-p)(1-\rho)} - \frac{1-p}{p(2-p)\rho}(e^{p\rho} - 1)\right)R_{1,1}.$$

4.2 Limited service

Consider now the well-known k-limited service discipline. Under this discipline, during a visit, the server keeps serving customers until either k customers have been served or the queue becomes empty, whichever occurs first.

A major benefit of the k-limited service discipline is that it in a way bounds the cycle time, which can be of vital importance for deadline-critical applications. However, the k-limited service discipline does not satisfy Property 1, making it notoriously difficult to analyze.

Indeed, also the evaluation of the limiting value of $\text{Cov}[L^e, I]$ in Equation (33) is difficult, even for $k = 1$, see Appendix B. However, while not giving explicit results, the calculations in Appendix B do suggest that also for the limited service discipline the limit of $n\text{Var}[C/n]$ exists. Consequently, it is plausible that in symmetric polling systems with a k-limited service discipline, the scaled cycle time C/n will again converge in probability to a deterministic value of $s/(1-\rho)$, as was also the case for the binomial service discipline.

Based on this result, we conjecture that as n grows large, the steady-state queue length at the start of a server visit satisfies the following relation, similar to Equations (24) and (26):

$$X_i^i \stackrel{d}{=} (X_i^i - k)^+ + A_i,$$

where A_i is a Poisson distributed random variable with parameter $\frac{\Lambda s}{1-\rho}$. We find that each individual queue asymptotically behaves as an $M/D/1$ queue with bulk service and fixed capacity as studied in [1].

For this $M/D/1$ bulk service queue, let π_i be the steady-state probability that the queue length at the start of a visit equals i and define $\Pi(z) = \sum_{i=0}^{\infty} \pi_i z^i$ and $v = \Lambda s/(1-\rho)$. Then in [1] it is shown that

$$\Pi(z) = \frac{(k-v)(z-1)\prod_{i=1}^{k-1}(z-z_i)/(1-z_i)}{z^k e^{v(1-z)} - 1}, \tag{34}$$

where the z_i are the zeros of the denominator within the unit circle. Moreover, one can deduce that

$$E[L^b] = \frac{k-(k-v)^2}{2(k-v)} + \sum_{i=1}^{k-1} \frac{1}{1-z_i}. \tag{35}$$

We conclude that analyzing the marginal queue length and waiting-time distributions for the k-limited service discipline becomes considerably easier in large polling systems. In Section 6 we will investigate through simulations how well this approach approximates the queue length distribution for finite n.

5   Flexible k-limited service

We have already discussed that the k-limited service discipline achieves a more predictable cycle time compared to the exhaustive and gated service disciplines, making it more suitable for deadline-critical applications. However, it is also known that the waiting times of a system with the k-limited service discipline can be large compared to the exhaustive and gated service disciplines. This is mostly caused by the fact that if the server reaches a very long queue, it will still serve at most k customers, even though it possibly did not have to serve any customers at the previously visited queues.

In order to reduce the effect of this drawback, we introduce a flexible k-limited service discipline. This discipline works the same as the k-limited service discipline, except that when the server serves less than k customers at a queue, we allow the server to use this 'lost' capacity during the visits to the next queues. This has the following benefit. If by chance there is a single abnormally large queue, while the queues before it are almost empty, the server is allowed to spend more time on the abnormally large queue, reducing waiting times.

More specifically, the flexible k-limited service discipline works as follows. Denote by $S(i, j)$ the number of served customers during the j'th visit of the server to $Q_{(i-1 \mod n)+1}$. Then upon reaching Q_i for the j'th time, the server will serve at most $K(i, j)$ customers, where

$$K(i,j) = k + \left(\ell k - \sum_{l=1}^{\ell} S(i-l, j - \mathbb{1}[i-l < 1]) \right)^+ ,$$

with $k \in \mathbb{N}$ and $\ell \in \{1, ..., n-1\}$. Note that the server will always be allowed to serve at least k and at most $(\ell + 1)k$ customers during a visit and that for $\ell = 0$ the flexible k-limited service policy coincides with the traditional k-limited policy. Furthermore, it is easily seen that during a cycle at most $(n + \ell)k$ customers will be served: the server serves $(\ell + 1)k$ customers at some queue and k customers at the following $n - 1$ queues.

In the following subsection we will analyze the flexible k-limited service policy and focus our attention on large systems with many queues.

5.1 Performance for large systems

Again let the arrival rate to each queue λ_i equal Λ/n. Our analysis in the previous sections suggests that $C/n \to s/(1-\rho)$ as $n \to \infty$, and therefore the number of customers that join a queue between two consecutive visit beginnings tends to a Poisson distributed random variable with mean $\Lambda s/(1-\rho)$ as n grows large. Therefore, we can argue that, under the flexible k-limited service discipline, each individual queue asymptotically behaves as an $M/D/1$ queue with bulk service and varying capacity. That is, again denoting by $L^b(i, j)$ the queue length of Q_i at the start of the j'th server visit, we have as $n \to \infty$:

$$L^b(i, j+1) = (L^b(i, j) - K(i, j))^+ + A(i, j), \tag{36}$$

where the $A(i, j)$ are i.i.d. $\text{Poi}(\Lambda s/(1-\rho))$ random variables.

In order to analyze the limiting behavior of Equation (36) as $j \to \infty$, we treat $\{K(i,j)\}_{j=1}^{\infty}$ as an i.i.d. sequence of random variables. Note that generally these variables will not be strictly independent, since a node having a low K value in a given cycle is more likely to have a low K value in the following cycle as well. However, if the sequence $\{K(i,j)\}_{j=1}^{\infty}$ is treated as i.i.d. with $p_m = \mathbb{P}[K = m]$ and $P_m = \mathbb{P}[K \geq m]$, we can determine the steady-state distribution of the Markov chain given by (36) as shown in [11].

Let π_i again denote the steady-state probability that the queue length at the start of a visit equals i and define $\Pi(z) = \sum_{i=0}^{\infty} \pi_i z^i$. Furthermore, write $\phi_i(z) = \sum_{j=k(\ell+1)-i}^{k(\ell+1)} p_j z^{-j}$. Writing again $\nu = \Lambda s/(1-\rho)$, we then have, see [11],

$$\Pi(z) = \frac{\sum_{i=0}^{k(\ell+1)-1} \pi_i (P_i - z^i \phi_{k(\ell+1)-i}(z))}{e^{\nu(1-z)} - \phi_{k(\ell+1)}(z)}.$$

Multiplying both the numerator and the denominator by $z^{k(\ell+1)}$, we know by Rouché's theorem that, in addition to the zero in $z = 1$, the denominator has exactly $k(\ell + 1) - 1$ zeros within the unit disk denoted by $z_1, ..., z_{k(\ell+1)-1}$. Moreover, since $\Pi(z)$ is analytic for $|z| < 1$, these zeros should also be the zeros of the numerator. Hence,

$$\Pi(z) = \frac{c(z-1) \prod_{i=1}^{k(\ell+1)-1} (z - z_i)}{z^{k(\ell+1)} (e^{\nu(1-z)} - \phi_{k(\ell+1)}(z))}.$$

Using the fact that $\Pi(1) = 1$, we find

$$c = (\mathbb{E}[K] - \nu) \prod_{i=1}^{k(\ell+1)-1} \frac{1}{1 - z_i}.$$

Here the probabilities p_i determining the function $\phi_{k(\ell+1)}(z)$ and the zeros z_i remain to be found.

5.1.1 The case $\ell = 1$

Consider the simplest case $\ell = 1$ in which the server is allowed to serve at most $2k$ customers at a queue. Supported by our analysis in Section 3, we assume asymptotic independence between the queue lengths of two neighboring queues. Since each time the server visits an empty queue it follows that the next queue is allowed to serve up to $2k$ customers, we should have $\pi_0 = p_{2k}$. More generally, we require $\pi_i = p_{2k-i}$, for $i = 0, ..., k-1$, and $p_k = 1 - \sum_{i=0}^{k-1} \pi_i$. Hence, this tells us that $q(z) = z^{2k} \phi_{2k}(z)$ is the moment generating function of $(2k - K)$. Substituting, we find

$$\Pi(z) = \frac{(\mathbb{E}[K] - \nu)(z-1) \prod_{i=1}^{2k-1} \frac{z - z_i}{1 - z_i}}{z^{2k} e^{\nu(1-z)} - q(z)}. \tag{37}$$

Since the coefficients of the function $q(z)$ are given in terms of the π_i, we can not directly find the roots z_i to determine the probabilities π_i. However, it is possible to determine them using the following iterative approach.

We start with initial estimates for π_i, $i = 0, ..., k-1$, which then also give estimates for the probabilities p_i and hence the generating function $q(z)$. Using these estimates, one can determine the roots z_i of the denominator of (37), giving a new estimate for the generating function $\Pi(z)$ and the probabilities π_i. Iterating this process will give estimates for the true values of π_i and p_i.

Implementing this approach gives good and fast results for reasonable values of k ($k < 20$) and ν not too close to k, converging after only a few iterations. In the next section we will investigate how well the actual queue length distribution is approximated using this approach.

Finally, note that the simplest case $k = 1$ allows a more direct treatment. Using the fact that $\Pi(0) = \pi_0 = p_2$ allows us to find the relation

$$p_2^2 = -(\mathbb{E}[K] - \nu) \frac{z_1}{1 - z_1}. \tag{38}$$

Using Equation (38), we can write z_1 as a function of p_2 as follows

$$z_1(p_2) := \frac{p_2^2}{p_2^2 - (\mathbb{E}[K] - \nu)} = \frac{p_2^2}{p_2^2 - (1 + p_2 - \nu)}.$$

Therefore, p_2 can be determined as the zero of

$$z_1(p_2)^2 e^{\nu(1 - z_1(p_2))} - (1 - p_2) z_1(p_2) - p_2,$$

also giving p_1. Additionally, one can deduce that

$$\mathbb{E}[L^b] = \frac{2 - (2 - \nu)^2}{2(1 + p_2 - \nu)} + \frac{1}{1 - z_1}. \tag{39}$$

5.1.2 The case $\ell > 1$

Consider now the case $\ell > 1$. This case can be treated similarly to the case $\ell = 1$, but it does require solving some combinatorial problems. For example, consider the case $\ell = 2$ and $k = 1$. Let X_{-3}, X_{-2} and X_{-1} denote the number of customers the server found at the last three queues that it visited and let K_{-3}, K_{-2} and K_{-1} denote the K values at those queues during those visits. If we again assume that queue lengths of neighboring queues are independent as $n \to \infty$, then the only way the current queue can have $K = 3$ is if its two predecessors did not have any customers waiting, hence $p_3 = \mathbb{P}[X_{-2} = X_{-1} = 0] = \pi_0^2$. Similarly, the probability that a queue has $K = 2$ is given by

$$\begin{aligned} p_2 =\ & \mathbb{P}[X_{-2} + X_{-1} = 1] + \mathbb{P}[X_3 > 1, K_{-3} > 1, X_{-2} = 0, X_{-1} > 1] \\ & + \mathbb{P}[X_{-2} > 1, K_{-2} = 1, X_{-1} = 0] \\ =\ & 2\pi_0 \pi_1 + (1 - \pi_0 - \pi_1)(1 - p_1)\pi_0(1 - \pi_0 - \pi_1) + (1 - \pi_0 + \pi_1)p_1 \pi_0. \end{aligned}$$

One can imagine that as ℓ and k become larger, expressing the probabilities p_i in terms of the π_j becomes more difficult. However, in practice, using moderate values of ℓ should suffice.

Note that in order to avoid dependencies between queue lengths and K's of neighboring queues and to simplify the analysis, one could also consider a randomized flexible k-limited service policy. By randomized we mean that instead of looking at how many customers were served during the last ℓ visits, the service limit $K(i,j)$ is determined by how many customers were served at ℓ random visits during the last cycle. Specifically, let $X^{i,j} = \{X_1^{i,j}, ..., X_\ell^{i,j}\}$ be i.i.d. randomly chosen subsets of $\{1, ..., n\} \backslash \{i\}$, then

$$K(i,j) = k + \left(\ell k - \sum_{l=1}^{\ell} S(X_l^{i,j}, j - \mathbb{1}[X_l^{i,j} > i]) \right)^+ .$$

Such a policy admits an easier analysis, since the probabilities p_i can be expressed more easily in terms of the π_j. Moreover, there should be weaker dependencies between $K(i,j)$ and $K(i,j+1)$.

6 Numerical results

In this section we will evaluate how well the limiting many-queue results of the previous sections for the marginal queue length distribution can serve as approximations for polling systems of finite size. We focus on the binomial-gated, k-limited and flexible k-limited service disciplines. Additionally, we will compare these disciplines with respect to the mean queue length and the cycle time variance.

6.1 Marginal queue length distribution convergence

We first consider the binomial-gated service discipline by investigating how well Equation (25) approximates the steady-state queue length at the start of a server visit for finite n. To this end, we simulate a polling system with $n = 10, 20, 50, 100, 200, 500$ for 10^8 cycles. The results for binomial-gated service are obtained by an exact analysis, using the PGF and LST inversion algorithms by Choudhury and Whitt [8]. We assume that service and switch-over times are exponentially distributed with mean $2/3$ and 1, respectively. Additionally, we set $\Lambda = 1$.

n	$p = 1/4$	$p = 1/2$	$p = 3/4$	$p = 1$
10	5.3×10^{-4}	1.0×10^{-3}	1.7×10^{-3}	$3. \times 10^{-3}$
20	1.5×10^{-4}	2.9×10^{-4}	5.1×10^{-4}	9.2×10^{-4}
50	2.6×10^{-5}	5.1×10^{-5}	9.2×10^{-5}	1.7×10^{-4}
100	6.7×10^{-6}	1.3×10^{-5}	2.4×10^{-5}	4.4×10^{-5}
200	1.7×10^{-6}	3.4×10^{-6}	6.1×10^{-6}	1.1×10^{-5}
500	2.7×10^{-7}	5.4×10^{-7}	9.9×10^{-7}	1.9×10^{-6}

Table 1: Binomial-gated service: Squared error between limiting and exact queue length distributions.

In Table 1 we show the squared error between the PDF of the limiting queue length distribution at cycle beginnings, given by Equation (25), and the corresponding exact results. The squared error between two PDFs $f(\cdot)$ and $g(\cdot)$ is defined as $\sum_{i=0}^{\infty} (f(i) - g(i))^2$. We find that even for small n the limiting distribution given by Equation (25) already approximates the steady-state queue length distribution well. In fact, as we can expect from the analysis of Section 3, the squared error is of order $1/n^2$. Additionally, we see that as p becomes smaller, the approximation tends to become more accurate.

Consider now the traditional k-limited service discipline without flexibility. In Table 2 we show the squared error between the PDF of the limiting queue length defined by the PGF of Equation (34) and the corresponding simulation results for several values of k, where we have used the same

n	$k = 4$	$k = 5$	$k = 6$
10	3.18×10^{-3}	3.88×10^{-3}	4.60×10^{-3}
20	9.42×10^{-4}	1.13×10^{-3}	1.33×10^{-3}
50	1.70×10^{-4}	2.01×10^{-4}	2.37×10^{-4}
100	4.46×10^{-5}	5.25×10^{-5}	6.15×10^{-5}
200	1.21×10^{-5}	1.38×10^{-5}	1.57×10^{-5}
500	2.02×10^{-6}	2.26×10^{-6}	2.57×10^{-6}

Table 2: k-limited service: Squared error between limiting and simulated queue length distributions.

parameter settings as before. Also here we find fast convergence to the limiting queue length distribution. For this case, however, the choice of k does not appear to have a big influence on the convergence rate.

n	$k = 4$	$k = 5$	$k = 6$
10	3.62×10^{-3}	2.93×10^{-3}	2.91×10^{-3}
20	1.13×10^{-3}	8.96×10^{-4}	8.97×10^{-4}
50	2.22×10^{-4}	1.66×10^{-4}	1.66×10^{-4}
100	6.56×10^{-5}	4.41×10^{-5}	4.34×10^{-5}
200	2.00×10^{-5}	1.14×10^{-5}	1.12×10^{-5}
500	4.95×10^{-6}	1.74×10^{-6}	1.93×10^{-6}

Table 3: Flexible k-limited service with $\ell = 1$: Squared error between limiting and simulated queue length distributions.

Finally, we consider the flexible k-limited service discipline with $\ell = 1$. We compare simulation results with the PDF of the limiting queue length defined by the PGF of Equation (37), which we determine by the iterative procedure described in Section 5.1. Results can be found in Table 3. As before, we find fast convergence to the limiting queue length distribution. To get a better impression of how well the actual queue length distribution is approximated by the limiting distribution, we have created a plot to compare them. Figure 1 shows the queue length distributions for systems with $n = 10$ and $n = 50$ queues, respectively, with flexible k-limited service ($k = 4$ and $\ell = 1$). Comparing these probabilities with the limiting distribution, also visualized in Figure 1, we see that the general shape is quite similar for $n = 10$, but the individual probabilities are not very close yet. For $n = 50$ the differences have almost disappeared and the limiting probabilities would be a very good approximation for the true distribution.

6.2 Comparison of service disciplines

In the previous subsection, we used simulations to examine how well asymptotic results can be used to approximate the performance of polling systems of finite size. In this subsection, we will compare the performance of the binomial-gated, k-limited and flexible k-limited service discipline.

We focus our attention on the cycle time variance and the mean queue length. In practice, one seeks a service discipline which minimizes the mean queue length, as this will also decrease the mean waiting time of customers. However, as we shall see, lowering the mean queue length usually comes at the price of increased cycle time variance. This trade-off is particularly important for deadline-critical applications, where upon returning to a queue, the server might find a high-priority customer which needs to be served immediately. In such a system a predictable return time of the server is necessary in order to guarantee that these high-priority customers with high probability do not have to wait longer than some predefined threshold.
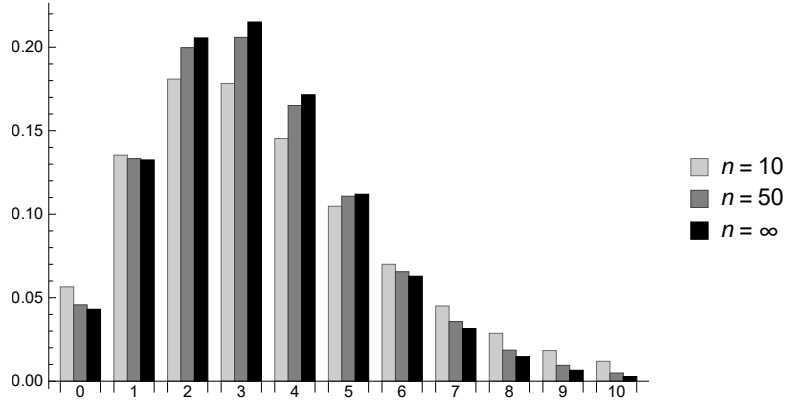
Figure 1: Queue-length distributions for flexible k-limited service with $k = 4$ and $\ell = 1$.

n	$p = 1/4$	$p = 1/2$	$p = 3/4$	$p = 1$
10	4.91	6.37	8.12	10.3
20	4.95	6.47	8.31	10.6
50	4.98	6.53	8.42	10.9
100	4.98	6.55	8.46	10.9
200	4.99	6.56	8.48	11.0
500	4.99	6.56	8.50	11.0
∞	4.99	6.57	8.50	11.0

Table 4: Binomial-gated service: Scaled variance of the cycle time distribution.

n	$p = 1/4$	$p = 1/2$	$p = 3/4$	$p = 1$
10	12	6	4	3
∞	12	6	4	3

Table 5: Binomial-gated service: Mean queue length at the start of a server visit.

In Tables 4 and 5 we show the variance of the cycle time and the mean queue length for the binomial-gated service discipline obtained by an exact analysis. The last row of both tables corresponds to the exact asymptotic results from Section 3.4. We find that, as predicted by Equation (23), we can actually decrease the cycle time variance by decreasing p. As expected (see the remark below Equation (5)), Table 5 shows that the *mean* queue length does not depend on the number of queues.

	$\ell = 0$			$\ell = 1$		
n	$k = 4$	$k = 5$	$k = 6$	$k = 4$	$k = 5$	$k = 6$
10	5.03	7.37	9.01	5.82	8.05	9.27
20	5.19	7.70	9.34	6.26	8.75	9.93
50	5.30	7.93	9.57	6.61	9.28	10.38
100	5.33	8.02	9.65	6.75	9.49	10.54
200	5.35	8.06	9.69	6.82	9.60	10.62
500	5.36	8.09	9.71	6.86	9.68	10.67

Table 6: k-limited and flexible k-limited service: Scaled variance of the simulated cycle time distribution.

Consider now the k-limited and the flexible k-limited service disciplines. In Tables 6 and 7 we

19

	$\ell = 0$			$\ell = 1$		
n	$k=4$	$k=5$	$k=6$	$k=4$	$k=5$	$k=6$
10	4.21	3.25	2.98	3.72	3.19	3.06
20	4.00	3.22	3.02	3.49	3.11	3.03
50	3.88	3.20	3.04	3.36	3.07	3.01
100	3.84	3.20	3.05	3.31	3.06	3.01
200	3.82	3.20	3.05	3.29	3.05	3.01
500	3.81	3.20	3.06	3.28	3.05	3.01
∞	3.80	3.20	3.06	3.26	3.04	3.01

Table 7: k-limited and flexible k-limited service: Simulated mean queue length at the start of a server visit.

show estimates of the variance of the cycle time and the mean queue length for both disciplines. Recall that the setting $\ell = 0$ corresponds to the k-limited service discipline without flexibility. Comparing with the results in Tables 4 and 5, we find that the k-limited service discipline is better at reducing the cycle time variance than the binomial-gated service discipline, without increasing the mean queue length too much when compared with the gated service discipline, i.e. $p = 1$.

If we now compare the results for the flexible k-limited service discipline with $k = 4$ with the results for the k-limited service discipline without flexibility and the gated service discipline, we find that the flexible k-limited service discipline is able to achieve the best of both worlds. Compared to the gated service discipline, it achieves almost a 40% decrease in scaled cycle time variance, while only increasing the mean queue length by roughly 10%. Compared to the k-limited service discipline without flexibility, it reduces the mean queue length by 15%, while only increasing the scaled cycle time variance by approximately 25%.
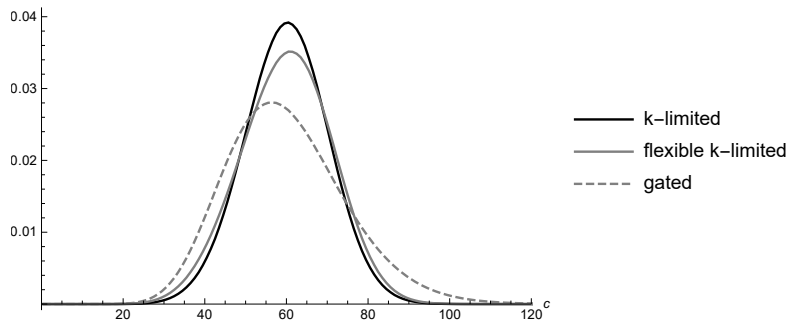

Figure 2: Cycle time PDFs for the k-limited, flexible k-limited and gated service disciplines for $n = 20$ and $k = 4$.

In Figure 2 we visualize this trade-off by plotting the cycle time PDFs of the gated service discipline and (simulations for) the k-limited service discipline with and without flexibility and $k = 4$. Indeed, the flexible k-limited service discipline is able to guarantee a server return time smaller than 85 time units for 99% of the cycles, while the gated service discipline can only guarantee this for 95% of the cycles. Finally, we see that the distributions for the k-limited service disciplines with and without flexibility are very similar.

7 Conclusion

We have studied the performance of symmetric polling systems in an asymptotic regime with a large number of queues. In contrast to "continuous" polling models, the distribution of the individual

switch-over times is assumed to remain the same as the number of queues grows large. We showed that for most classical service policies the scaled cycle times converge to a deterministic value in the limit, which in turn implies that the queue lengths at the various nodes become asymptotically independent. Using these insights, we demonstrated that each individual queue behaves as a discrete-time bulk-service queue in the limit, so that the analysis of the marginal queue length and waiting-time distributions greatly simplifies. Finally, we proposed and analyzed a new flexible k-limited service discipline aimed at achieving a good trade-off between short mean queue lengths and predictable cycle times. Our results provide insight in the delay performance of token-passing algorithms such as the BACnet protocol in deployment scenarios with a large number of nodes and deadline-critical applications which require short delays and guaranteed token return times.

Acknowledgments

Appendix

A   Extension of Proposition 1

In this appendix we generalize Proposition 1 to find the cycle time for branching-type polling systems that do *not* satisfy Property 2. Denote by $C_{i,j}(z)$ the PGF of $M_{i,j}$ as introduced in Section 2. For simplicity, we assume that the $M_{i,j}$ are independent, but at the end of this section we briefly discuss the case with dependencies. A visit period at Q_i starting with m customers present at Q_i is considered to consist of m sub-busy periods, each having LST $\theta_i(\cdot)$. Additionally, let $\tilde{\theta}_i(z, \omega)$ denote the joint transform of the number of customers that are served during a sub-busy period visit to Q_i and the length of the sub-busy period. Lastly, we define L_i as the number of service completions at Q_i during a visit in steady state.

Proposition 2. *Write $\mathbf{z} = (z_1, ..., z_n)$ and $\boldsymbol{\omega} = (\omega_1, ..., \omega_n)$. In case Property 1 holds, the joint transform of L_i and $V_j + S_j$, $i = 1, ..., n$, $j = 1, ..., n$, is given by*

$$\mathbb{E}\left[\left(\prod_{i=1}^{n} z_i^{L_i}\right) e^{-\sum_{j=1}^{n} \omega_j(V_j + S_j)}\right] = \mathscr{F}_1(\tilde{\kappa}_1(\mathbf{z}, \boldsymbol{\omega}), ..., \tilde{\kappa}_n(\mathbf{z}, \boldsymbol{\omega})) \prod_{i=1}^{n} \mathscr{S}_i(\tilde{\gamma}_i(\mathbf{z}, \boldsymbol{\omega})), \qquad (40)$$

where $\tilde{\gamma}_n(\mathbf{z}, \boldsymbol{\omega}) = \omega_n$, $\tilde{\kappa}_n(\mathbf{z}, \boldsymbol{\omega}) = \tilde{\theta}_n(z_n, \omega_n)$ and, for $1 \leq i \leq n-1$,

$$\tilde{\gamma}_i(\mathbf{z}, \boldsymbol{\omega}) = \omega_i + \sum_{j=i+1}^{n} \lambda_j(1 - \tilde{\theta}_j(z_j, \tilde{\gamma}_j(\mathbf{z}, \boldsymbol{\omega}))), \qquad (41)$$

and

$$\tilde{\kappa}_i(\mathbf{z}, \boldsymbol{\omega}) = \tilde{\theta}_i\left(z_i \prod_{j=i+1}^{n} C_{i,j}\left(\tilde{\kappa}_j(\mathbf{z}, \boldsymbol{\omega})\right), \tilde{\gamma}_i(\mathbf{z}, \boldsymbol{\omega})\right). \qquad (42)$$

Proof. We define

$$Z_i := \text{extra work introduced at } Q_i \text{ during } V_1 + S_1,$$
$$H_i := \text{extra customers introduced at } Q_i \text{ during } V_1 + S_1.$$

Consider first the 2-queue case. Then

$$\mathbb{E}\left[z_1^{L_1}z_2^{L_2}\exp\left(-\omega_1(V_1+S_1)-\omega_2(V_2+S_2)\right)\mid X=(m_1,m_2)\right]$$

$$=\tilde{\theta}_2(z_2,\omega_2)^{m_2}\mathscr{S}_2(\omega_2)\mathbb{E}\left[z_1^{L_1}z_2^{H_2}\exp\left(-\omega_1(V_1+S_1)-\omega_2 Z_2\right)\mid X=(m_1,m_2)\right]$$

$$=\tilde{\theta}_1\left(z_1 C_{1,2}\left(\tilde{\theta}_2(z_2,\omega_2)\right),\omega_1+\lambda_2(1-\tilde{\theta}_2(z_2,\omega_2))\right)^{m_1}\tilde{\theta}_2(z_2,\omega_2)^{m_2}$$

$$\cdot\,\mathscr{S}_1(\omega_1+\lambda_2(1-\tilde{\theta}_2(z_2,\omega_2)))\mathscr{S}_2(\omega_2),$$

where for the second equality, we have used that

$$\mathbb{E}\left[z_2^{H_2}e^{-\omega_2 Z_2}\mid L_1=l\right]$$

$$=\tilde{\theta}_1\left(C_{1,2}\left(\tilde{\theta}_2(z_2,\omega_2)\right),\lambda_2(1-\tilde{\theta}_2(z_2,\omega_2))\right)^l\mathscr{S}_1(\lambda_2(1-\tilde{\theta}_2(z_2,\omega_2))).$$

It follows that

$$\mathbb{E}\left[z_1^{L_1}z_2^{L_2}e^{-\omega_1(V_1+S_1)-\omega_2(V_2+S_2)}\right]$$

$$=\mathscr{F}_1\left(\tilde{\theta}_1\left(z_1 C_{1,2}\left(\tilde{\theta}_2(z_2,\omega_2)\right),\omega_1+\lambda_2(1-\tilde{\theta}_2(z_2,\omega_2))\right),\tilde{\theta}_2(z_2,\omega_2)\right)$$

$$\cdot\,\mathscr{S}_1(\omega_1+\lambda_2(1-\tilde{\theta}_2(z_2,\omega_2)))\mathscr{S}_2(\omega_2).$$

Consider now n queues. Generalizing the analysis of the 2-queue example, we deduce

$$\mathbb{E}\left[\left(\prod_{i=1}^{n}z_i^{L_i}\right)e^{-\sum_{j=1}^{n}\omega_j(V_j+S_j)}\,\middle|\,X=m\right]$$

$$=\mathscr{S}_n(\tilde{\gamma}_n(z,\omega))(\tilde{\kappa}_n(z_n,\omega_n))^{m_n}$$

$$\cdot\,\mathbb{E}\left[\left(\prod_{i=1}^{n-1}\left(z_i C_{i,n}(\tilde{\kappa}_n(z,\omega))\right)^{L_i}\right)e^{-\sum_{i=1}^{n-1}(\omega_i+\lambda_n(1-\tilde{\theta}_n(z_n,\omega_n)))(V_i+S_i)}\,\middle|\,X=m\right]$$

$$=\mathscr{S}_n(\tilde{\gamma}_n(z,\omega))(\tilde{\kappa}_n(z,\omega))^{m_n}\mathscr{S}_{n-1}(\tilde{\gamma}_{n-1}(z,\omega))(\tilde{\kappa}_{n-1}(z,\omega))^{m_{n-1}}$$

$$\cdot\,\mathbb{E}\Bigg[\left(\prod_{i=1}^{n-2}\left(z_i C_{i,n-1}(\tilde{\kappa}_{n-1}(z,\omega))C_{i,n}(\tilde{\kappa}_n(z,\omega))\right)^{L_i}\right)$$

$$\cdot\,e^{-\sum_{i=1}^{n-2}(\omega_i+\lambda_{n-1}(1-\tilde{\theta}_{n-1}(z_{n-1},\tilde{\gamma}_{n-1}(z,\omega)))+\lambda_n(1-\tilde{\theta}_n(z_n,\omega_n)))(V_i+S_i)}\,\middle|\,X=m\Bigg].$$

Iterating, we find

$$\mathbb{E}\left[\left(\prod_{i=1}^{n}z_i^{L_i}\right)e^{-\sum_{j=1}^{n}\omega_j(V_j+S_j)}\,\middle|\,X=m\right]=\prod_{i=1}^{n}\tilde{\kappa}_i(z,\omega)^{m_i}\mathscr{S}_i(\tilde{\gamma}_i(z,\omega)).$$

Consequently,

$$\mathbb{E}\left[\left(\prod_{i=1}^{n}z_i^{L_i}\right)e^{-\sum_{j=1}^{n}\omega_j(V_j+S_j)}\right]=\mathscr{F}_1(\tilde{\kappa}_1(z,\omega),...,\tilde{\kappa}_n(z,\omega))\prod_{i=1}^{n}\mathscr{S}_i(\tilde{\gamma}_i(z,\omega)).$$

$$\square$$

In particular, this implies that the LST of the cycle time is given by

$$\mathbb{E}\left[e^{-uC}\right]=\mathscr{F}_1(\kappa_1(u),...,\kappa_n(u))\prod_{i=1}^{n}\mathscr{S}_i(\gamma_i(u,...,u)),\tag{43}$$

where $\kappa_n(u) = \theta_n(u)$ and, for $1 \leq i \leq n-1$,

$$\kappa_i(u) = \tilde{\theta}_i\Big(\prod_{j=i+1}^{n} C_{i,j}\big(\kappa_j(u)\big), \gamma_i(u, \ldots, u)\Big). \tag{44}$$

Remark 5. *As indicated, we have assumed that the $M_{i,j}$'s are independent. However, the results can easily be extended to the case with dependencies by introducing $C_i(\boldsymbol{z})$ as the joint PGF of $M_{i,1}, \ldots, M_{i,n}$. It can be verified that the only difference would be that in (42) and (44), the product needs to be replaced by this joint PGF. For example, in (44) one would obtain*

$$\kappa_i(u) = \tilde{\theta}_i\big(C_i(1, 1, \ldots, 1, \kappa_{i+1}(u), \kappa_{i+2}(u), \ldots, \kappa_n(u)), \gamma_i(u, \ldots, u)\big).$$

In fact, by taking $C_i(\boldsymbol{z}) = p_{i,0} + \sum_{j=1}^{n} p_{i,j} z_j$ we obtain a polling system with customer routing, as studied in [4, 17], where $p_{i,j}$ denotes the probability that a customer finishing service at Q_i is routed to Q_j. For that particular choice of $C_i(\boldsymbol{z})$, our result agrees with the cycle time LST found in [4, Proposition 3.1].

B 1-limited service

We consider the 1-limited service discipline. The goal of this appendix is to analyze the limiting value of $\mathrm{Cov}[L^e, I]$ as n grows large and demonstrate why this is difficult. Furthermore, we intend to provide arguments that support the conjecture that the variance of the scaled cycle time is of order $1/n$.

Let $g := \mathbb{P}[L^b > 0] = \Lambda s/(1-\rho)$. Note that, since switch-over times during an intervisit time are independent of L^e, by additivity of the covariance function

$$\mathrm{Cov}[L^e, I] = \sum_{i=2}^{n} \mathrm{Cov}[L^e, V_i].$$

Furthermore, we can write

$$\mathbb{E}[L^e V_i] = \mathbb{E}[L^e]\beta - \mathbb{P}[Q_i \text{ is found empty}]\mathbb{E}[L^e \mid Q_i \text{ is found empty}]\beta.$$

Notice that $\mathbb{P}[Q_i \text{ is found empty}] = 1 - g$. Using $\mathbb{E}[V_i] = g\beta$, we conclude

$$\mathrm{Cov}[L^e, I] = \beta(1-g)\sum_{i=2}^{n}\big(\mathbb{E}[L^e] - \mathbb{E}[L^e \mid Q_i \text{ is found empty}]\big). \tag{45}$$

Hence, it remains to find $\mathbb{E}[L^e]$ and $\mathbb{E}[L^e \mid Q_i \text{ is found empty}]$.

Write $\mathscr{F}(z)$ for the PGF of the joint queue lengths at the start of a visit to Q_1. For 1-limited service it is well known (see for example [19]) that the following relation holds

$$\mathscr{F}(z_1, \ldots, z_n) = \left(\frac{1}{z_n}\left(\mathscr{F}(z_n, z_1, \ldots, z_{n-1}) - \mathscr{F}(0, z_1, \ldots, z_{n-1})\right)\mathscr{B}\left(\sum_{j=1}^{n}\frac{\Lambda}{n}(1-z_j)\right)\right.$$
$$\left. + \mathscr{F}(0, z_1, \ldots, z_{n-1})\right)\mathscr{S}\left(\sum_{j=1}^{n}\frac{\Lambda}{n}(1-z_j)\right). \tag{46}$$

Differentiating once gives

$$l_i = \begin{cases} l_{i+1} + g\rho/n + s\Lambda/n = l_{i+1} - g/n, & i \neq n, \\ l_1 - g(1-\rho/n) + s\Lambda/n = l_1 - (1-1/n)g, & i = n. \end{cases} \tag{47}$$

Summing gives

$$\sum_{i=1}^{n} l_i = nl_1 - (n-1)g/2. \tag{48}$$

Setting $z_i = z$ for all i in Equation (46), we find

$$\mathscr{F}(z,...,z) = \frac{\mathscr{S}(\Lambda(1-z))(z-\mathscr{B}(\Lambda(1-z)))}{z-\mathscr{S}(\Lambda(1-z))\mathscr{B}(\Lambda(1-z))}\mathscr{F}(0,z,...,z). \tag{49}$$

We define $\mathscr{G}_i(z) := \mathscr{F}(1,...1,z,1,...,1)$ and $\mathscr{G}_i^0(z) := \mathscr{F}(0,1,...,1,z,1,...,1)$, where z is the i'th argument and $\mathscr{G}_1^0 := 1-g$, and let $l_i^0 := \frac{\partial}{\partial z_i}\mathscr{G}_i^0(z)\big|_{z=(1,...,1)}$. Differentiating (49) then gives

$$\sum_{i=1}^{n} l_i = \frac{1}{1-g}\sum_{i=1}^{n} l_i^0 + \frac{g}{1-g} + \frac{\Lambda^2}{(1-g)(1-\rho)}\left(s^{(2)} - s^2 + \frac{1}{2}(g\beta^{(2)} - s^{(2)})\right). \tag{50}$$

For a third equation, summing Equation (46) over all i, we find

$$\sum_{i=1}^{n} \mathscr{G}_i(z) = \frac{\mathscr{S}\left(\frac{\Lambda}{n}(1-z)\right)\left(1-\mathscr{B}\left(\frac{\Lambda}{n}(1-z)\right)\right)}{1-\mathscr{S}\left(\frac{\Lambda}{n}(1-z)\right)\mathscr{B}\left(\frac{\Lambda}{n}(1-z)\right)}\sum_{i=1}^{n} \mathscr{G}_i^0(z)$$
$$- \left(1-\frac{1}{z}\right)\frac{\mathscr{S}\left(\frac{\Lambda}{n}(1-z)\right)\mathscr{B}\left(\frac{\Lambda}{n}(1-z)\right)}{1-\mathscr{S}\left(\frac{\Lambda}{n}(1-z)\right)\mathscr{B}\left(\frac{\Lambda}{n}(1-z)\right)}(\mathscr{G}_1(z) - \mathscr{G}_1^0(z)).$$

Differentiating and simplifying gives

$$\sum_{i=1}^{n} l_i = \frac{1}{\rho(1-g)+g}\left(nl_1 + \sum_{i=1}^{n} l_i^0\right) - \frac{\Lambda^2(s^{(2)} - 2s^2 + g\beta^{(2)}) - 2g\rho + 2ng}{2(\rho(1-g)+g)}. \tag{51}$$

Finally, solving Equations (48), (50) and (51), we find (see also [19])

$$\sum_{i=1}^{n} l_i = \frac{1}{1-g}\left(\frac{\Lambda^2 s^{(2)} - 2\Lambda^2 s^2 + (n+1)\Lambda s}{2(1-\rho)} + \frac{\Lambda^3 s\beta^{(2)}}{2(1-\rho)^2}\right),$$

$$\sum_{i=2}^{n} l_i^0 = \frac{(n-1)\Lambda s}{2(1-\rho)} = (n-1)\frac{g}{2},$$

and

$$l_i = \frac{g(2-g)}{2(1-g)} + \frac{1}{n}\frac{\Lambda^2(s^{(2)} - s^2) + \Lambda^2 g(\beta s + \beta^{(2)})}{2(1-\rho)(1-g)} - \frac{i-1}{n}g.$$

Additionally, we find

$$l_i^0 = \frac{g(2-g)}{2} - \frac{i-1}{n}g(1-g).$$

It is then easily seen that

$$\mathbb{E}[L^e] = l_1 - (1-\rho/n)g.$$

Consider now a special cycle $C^0 = \sum_{i=1}^{n}\left(V_i^0 + S_i^0\right)$ in which the server at the end of the cycle returns to find Q_1 empty. Then, by symmetry, we can write

$$\mathbb{E}[L^e \mid Q_{n-i+2} \text{ is found empty}] = \frac{1}{1-g}l_i^0 - \frac{\Lambda}{n}\mathbb{E}\left[S_i^0 + \sum_{j=i+1}^{n}\left(V_j^0 + S_j^0\right)\right].$$

Now, write

$$\mathbb{E}[S_i^0] = s - s^0(i) \quad \text{and} \quad \mathbb{E}[V_i^0] = g\beta - v^0(i).$$

Then combining all results we find

$$\text{Cov}[L^e, I] = \frac{n-1}{2n}\left(\frac{\Lambda^2(s^{(2)} - s^2 + g\beta^{(2)})}{1-\rho} + \rho g^2\right)$$

$$+ (1-g)\frac{\Lambda}{n}\left(\sum_{i=2}^{n}(i-1)s^0(i) + \sum_{j=3}^{n}(j-2)v^0(j)\right),$$

and, taking limits and simplifying even further,

$$\lim_{n\to\infty} n\text{Var}[C/n] = s^{(2)} - s^2 + g\beta^{(2)} - g^2\beta^2$$

$$- (1-g)\lim_{n\to\infty}\frac{2}{n}\left(\sum_{i=2}^{n}(i-1)s^0(i) + \sum_{j=3}^{n}(j-2)v^0(j)\right).$$

Disregarding the last term, this result already has the desired property that, as $g \uparrow 1$, the scaled variance converges to $\text{Var}[S] + \text{Var}[B]$ and, as $g \downarrow 0$, the scaled variance converges to $\text{Var}[S]$.

It remains to determine $s^0(i)$ and $v^0(j)$, which turns out to be non-trivial. However, since both $s^0(i)$ and $v^0(j)$ are most likely of order $1/n$, we conjecture that the variance of the scaled cycle time will also be of order $1/n$. Hence, also for 1-limited service, the scaled cycle time will converge in probability to a deterministic value, which we can exploit to determine the steady-state queue length distribution at visit beginnings.

References

[1] N. T. J. Bailey. On queueing processes with bulk service. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(1):80–87, 1954.

[2] M. A. A. Boon, I. J. B. F. Adan, and O. J. Boxma. A polling model with multiple priority levels. *Performance Evaluation*, 67:468–484, 2010.

[3] M. A. A. Boon, R. D. van der Mei, and E. M. M. Winands. Applications of polling systems. *Surveys in Operations Research and Management Science*, 16(2):67–82, 2011.

[4] M. A. A. Boon, R. D. van der Mei, and E. M. M. Winands. Waiting times in queueing networks with a single shared server. *Queueing Systems*, 74(4):403–429, 2013.

[5] O. J. Boxma and W. P. Groenendijk. Pseudo-conservation laws in cyclic-service systems. *Journal of Applied Probability*, 24(4):949–964, 1987.

[6] O. J. Boxma, J. Bruin, and B. Fralix. Sojourn times in polling systems with various service disciplines. *Performance Evaluation*, 66(11):621–639, 2009.

[7] S. T. Bushby. BACnetTM: A standard communication infrastructure for intelligent buildings. *Automation in Construction*, 6(5):529–540, 1997.

[8] G. L. Choudhury and W. Whitt. Computing distributions and moments in polling models by numerical transform inversion. *Performance Evaluation*, 25(4):267–292, 1996.

[9] E. Coffman and E. Gilbert. A continuous polling system with constant service times. *IEEE Transactions on Information Theory*, 32(4):584–591, 1986.

[10] S. W. Fuhrmann and R. B. Cooper. Stochastic decompositions in the M/G/1 queue with generalized vacations. *Operations Research*, 33(5):1117–1129, 1985.

[11] N. K. Jaiswal. A bulk-service queueing problem with variable capacity. *Journal of the Royal Statistical Society. Series B (Methodological)*, 23(1):143–148, 1961.

[12] W. Kastner, G. Neugschwandtner, S. Soucek, and H. Newmann. Communication systems for building automation and control. *Proceedings of the IEEE*, 93(6):1178–1203, 2005.

[13] D. P. Kroese and V. Schmidt. A continuous polling system with general service times. *Annals of Applied Probability*, 2(4):906–927, 1992.

[14] H. Levy. Binomial-gated service: a method for effective operation and optimization of polling systems. *IEEE Transactions on Communications*, 39(9):1341–1350, 1991.

[15] H. Levy, M. Sidi, and O. J. Boxma. Dominance relations in polling systems. *Queueing Systems*, 6(1):155–171, 1990.

[16] J. A. C. Resing. Polling systems and multitype branching processes. *Queueing Systems*, 13(4): 409–426, 1993.

[17] M. Sidi, H. Levy, and S. W. Fuhrmann. A queueing network with a single cyclically roving server. *Queueing Systems*, 11:121–144, 1992.

[18] N. C. Strole. The IBM token-ring network - A functional overview. *IEEE Network*, 1(1):23–30, 1987.

[19] H. Takagi. Mean message waiting times in symmetric multi-queue systems with cyclic service. *Performance Evaluation*, 5(4):271–277, 1985.

[20] R. D. van der Mei and H. Levy. Polling systems in heavy traffic: Exhaustiveness of service policies. *Queueing Systems*, 27(3–4):227–250, 1997.

[21] S. Wang. *Intelligent Buildings and Building Automation*. Taylor & Francis, New York, 2009.

[22] E. M. M. Winands, I. J. B. F. Adan, and G.-J. van Houtum. Mean value analysis for polling systems. *Queueing Systems*, 54:35–44, 2006.