Polling Systems with Periodic Server Routing in Heavy-Traffic: Distribution of the Delay

Tava Lennon Olsen[a] and R.D. van der Mei[b,c]

[a]John M. Olin School of Business, Washington University, St. Louis, MO, USA

[b]KPN Research, QoS Control, 2260 AK Leidschendam, Netherlands

[c]Vrije Universiteit, Mathematics and Computer Science, 1081HV Amsterdam, Netherlands

We consider polling systems with mixtures of exhaustive and gated service in which the server visits the queues periodically according to a general polling table. We derive exact expressions for the steady-state delay incurred at each of the queues, under standard heavy-traffic scalings. The expressions require the solution of a set of only $M - N$ linear equations, where M is the length of the polling table and N is the number of queues, but are otherwise explicit. The equations can even be expressed in closed form for several routing schemes commonly used in industry, such as the star and elevator visit order, in a general parameter setting. We also provide a hypothesis for extending our results to non-Poisson arrivals. The results are strikingly simple and reveal a number of asymptotic properties providing new and fundamental insights in the behavior of polling systems. In addition, the results lead to simple and fast approximations for the distributions and the moments of the delay in stable polling systems with periodic server routing. Numerical results demonstrate that the approximations are highly accurate for medium and heavily loaded systems.

Keywords: polling system, polling table, periodic routing, delay, waiting time, heavy-traffic, distributional approximations, switch-over times, setup times

1 Introduction

A typical polling model consists of a number of customer classes sharing a single server that incurs a switch-over time each time it changes class. Polling models are applicable to a wide range of computing, telecommunications, and manufacturing environments and have been studied extensively. Surveys of polling model research and its applications may be found in Levy and Sidi [23], Olsen [29], and Takagi [34, 35, 36, 37]. The server may move between the customer classes in a variety of fashions; here, we study models with periodic server routing where the order in which the server visits the classes is prescribed by a so-called polling table of finite length. Once service has begun at a class it continues until either all customers in the current class are served (*exhaustive service*) or until all customers that were in the current class when the server completed switching over to that class are served (*gated service*).

In this paper we study the distribution of delay in periodic polling models under heavy-traffic assumptions. Adopting the notion of pseudo-queues, from [3], we show that the distribution of the queue-length at polling instants at a pseudo-queue converges to a gamma-distribution with known parameters, requiring only the solution of a set of $M - N$ linear equations, where M is the length of the polling table and N is the number of queues. The equations can be solved explicitly under polling schemes commonly used in industry, like the star and elevator (scan) polling schemes. Further, an explicit approximation is available for a variety of "well designed" routing schemes. The results show that the scaled delay distributions depend on the system parameters only through a few aggregated system parameters, providing new and fundamental insight in the behavior of a general class of polling systems. They also allow us to propose a hypothesis for the steady-state distribution of scaled delay under general renewal arrivals and for the transient behavior of the system under heavy-traffic scaling. Finally, the results suggest simple and fast-to-evaluate approximations for the delay distribution at each of the queues in stable polling systems. Numerical experiments show that the approximations are highly accurate in practical heavy-traffic scenarios.

There has been relatively little work devoted to polling systems in heavy traffic. The work that has been done falls into three main classes: limit theorems for steady-state systems, limit theorems for transient systems, and dynamic scheduling papers that utilize heavy-traffic results. The first class is what we study here and other research is described further below. Polling model work in the second class includes Coffman et al. [9, 10], Jennings [19], and Reiman and Wein [31]. Work in the last class, namely the harnessing of heavy-traffic polling work to perform stochastic scheduling, includes Markowitz et al. [24], Markowitz and Wein [25], Reiman et al. [32], and Reiman and Wein [30].

Here we consider heavy-traffic limit theorems for systems in steady-state. Note that the transient limit theorems mentioned above can be used to produce steady-state results. Of most relevance to the work here is Coffman et al. [10] which will be discussed in more detail in Section 6.3. However, in many cases it is more straightforward to consider only steady-state results. In this case, non-heavy-traffic results can be harnessed and significant simplifications shown to result as utilization is taken to 100%. For example, Van der Mei [41, 42, 43] and Van der Mei and Levy [45, 46] use the non-heavy-traffic concept of descendant sets to obtain expressions for the moments of the delay in heavy traffic for a variety of polling models. Van der Mei and Borst [44] study cyclic polling systems with multiple servers by means of the so-called power-series algorithm [4]. Numerical results indicate that when the servers visit the queues in the same cyclic order they tend to cluster, which leads to an increase of the expected delay at the queues. Moreover, it is shown that this clustering effect is alleviated when the visit order of the servers are different. Kroese [21] analyzes the heavy-traffic behavior of continuous polling systems, and shows that the total number of customers has approximately a gamma distribution. Van der Mei [40] and [39] find the distribution of delay for cyclic server systems (where the polling table lists each queue exactly once) under heavy-traffic and heavy-switch-over scalings, respectively. Olsen [28] finds the distribution of delay for periodic service systems under heavy-switch-over scalings. However, as the scaling is based on strong-law type results, rather than the central-limit type scaling considered here, the distributional forms are simpler and therefore form a less accurate approximation than that given here for moderately loaded systems.

The results presented in this paper represent a significant advancement on previous work. In particular, they generalize the results in [40, 41, 42, 43]. Closed-form expressions for the first moments of the delay were derived in [42] for the polling model with non-cyclic periodic server routing considered in the present paper (see Section 2 for more details). Whereas, [40, 41, 43] derived closed-form expressions for LST of the delay distribution for the special case of cyclic server routing. The methodology developed in those papers, based on exploring the so-called Descendant Set Approach (DSA), has been shown to be highly effective for deriving heavy-traffic results. Therefore, in this paper we adopt the same methodological steps to derive heavy-traffic results for the model under consideration. We emphasize that the contribution of the present paper is fairly limited from a purely methodological point of view, but that the results derived in this paper contain a significant number of new elements. First, we obtain expressions for the complete distribution of the delay at each of the queues in heavy traffic for general non-cyclic polling tables, requiring the solution of a set of at most $M-N$ linear equations. As a by-product, the results are shown to lead to closed-form expressions for the distribution and the moments of the delay under the elevator polling scheme and the star polling scheme commonly used in industry. In this context we emphasize that the inclusion of non-cyclic server routing schemes significant enhances the applicability of the model, because in many applications (e.g., in computer-communication networks and in production and manufacturing systems) the server routing is clearly non-cyclic. The results show explicitly how the delay figures depend on the system parameters and as such provide new and fundamental insights in the behavior of a general class of polling models, making the results particularly interesting from a queueing-theoretical perspective. Second, the results lead to simple and fast approximations for the tail probabilities and the higher moments of the delay in stable polling systems, explicitly capturing the impact visit order. The approximations are particularly accurate when the system is heavily loaded, whereas the efficiency of the existing numerical techniques to obtain the higher moments and the tail probabilities of the delay tends to degrade significantly when the system is heavily loaded. In this way, the approximations proposed in the present paper *complement* the applicability of the existing numerical techniques. These observations make the added value of the present paper is evident, both from queueing-theoretical perspective and from an application point of view.

Exact results on the complete probability distribution of the delay are mainly restricted to two-queue models (cf., e.g., [5, 11, 12, 47]), but even in those cases non-trivial numerical techniques need to be used to obtain the tail probabilities of the delay. In the absence of exact distributional results, numerical techniques have been proposed to calculate the delay and queue-length distributions, like Blanc's power-series algorithm [4], Leung's technique based on discrete Fourier Transforms [22] or the numerical transform inversion technique in Choudhury and Whitt [7]. Federgruen and Katalan [14] propose a method to approximate the queue-length and delay distributions in a class of polling models similar to those we consider here. The relationship of their work to ours will be discussed in Section 5. However, like all the work above, their results require significant computational effort. Our (asymptotic) distributional results require only the solution to a set of $M-N$ linear equations and in many cases can be made explicit.

This paper is organized as follows. In Section 2 the model is described and some notation is introduced. In Section 3 we present a descendant set analysis for some key system parameters. In Section 4 we analyze the number of customers present at a polling instant. In Section 5 we obtain exact expressions for the delay distribution under heavy-traffic scalings. In Section 6 implications and extensions of the results are discussed. In Section 7 we propose and test approximations of the delay distribution in stable polling systems. Section 8 contains some concluding remarks and addresses a number of topics for further research.

2 Model and Notation

Consider a system consisting of N infinite-buffer queues, $Q_1, \ldots, Q_N$. Customers arrive at Q_j according to a Poisson arrival process with rate λ_j, $j = 1, \ldots, N$. The total arrival rate is denoted by $\Lambda = \sum_{j=1}^{N} \lambda_j$. The service time of a customer at queue j is a random variable B_j, with Laplace-Stieltjes Transform (LST) $\mathcal{B}_j(\cdot)$ and with finite k-th moment $b_j^{(k)}$, $j = 1, \ldots, N$, $k = 1, 2, \ldots$ The k-th moment of the service time of an arbitrary customer is denoted by $b^{(k)} = \sum_{j=1}^{N} \lambda_j b_j^{(k)} / \Lambda$, $k = 1, 2, \ldots$ The load offered to Q_j is $\rho_j = \lambda_j b_j^{(1)}$, and the total offered load is equal to $\rho = \sum_{j=1}^{N} \rho_j$. All interarrival times, service times and switch-over times (defined below) are assumed to be mutually independent and independent of the state of the system. A necessary and sufficient condition for the stability of the system is $\rho < 1$ (see, e.g., Fricker and Jaïbi [17]). Throughout, we assume that this condition is satisfied, and that the system is in steady state, unless otherwise indicated.

A single server inspects the queues periodically according to a general polling table of length M, described by a mapping $T(\cdot)$, which is used such that the server visits the queues periodically in the order $T(1), T(2), \ldots, T(M), T(1), T(2), \ldots$. Following the approach in Baker and Rubin [3], a unique pseudo-queue is associated with each entry in the polling table. Denote by PQ_i the pseudo-queue associated with the i-th entry in the polling table; its corresponding queue has index $T(i)$. Customers who arrive at $Q_{T(i)}$ and are served at PQ_i are referred to as type-i customers. The service at each pseudo-queue is either according to the gated policy or the exhaustive policy. For ease of exposition, we assume that pseudo-queues corresponding to the same queue have the same service strategy. Define $E := \{j : Q_j \text{ is served exhaustively}\}$ and $G := \{j : Q_j \text{ is served according to the gated policy}\}$. At each queue the customers are served on a First-In-First-Out (FIFO) basis.

After completing service at PQ_i the server proceeds to PQ_{i+1}, incurring a switch-over period with distribution equal to that of a random variable R_i with LST $\mathcal{R}_i$, with finite moments and with mean r_i and with higher moments $r_i^{(l)}$ $(l > 1)$, $i = 1, \ldots, M$. Denote by $r = \sum_{i=1}^{M} r_i$ the expected total switch-over time per cycle. Throughout it is assumed that $r > 0$. Let τ_{ij} be the entry in the polling table corresponding to the next visit to Q_j after a departure from PQ_i, and let σ_{ij} be the entry corresponding to the last visit to Q_j prior to an arrival of the server at PQ_i ($i = 1, \ldots, M,\ j = 1, \ldots, N$). I_E stands for the indicator function on the event E. Indices i corresponding to queues and pseudo-queues should be read as $[(i-1)\ mod\ N] + 1$ and $[(i-1)\ mod\ M] + 1$, respectively.

The moments at which the server arrives at PQ_i are referred to as the polling instants at PQ_i. Denote by V_i the visit time at PQ_i, i.e., the time between a polling instant at PQ_i and the server's successive departure from PQ_i. Denote by $I_i^{(PQ)}$ the intervisit time of PQ_i, i.e., the duration of the time between a departure of the server from $PQ_{\sigma_{iT(i)}}$ and the successive polling instant of PQ_i; denote the corresponding LST by $\mathcal{I}_i^{(PQ)}(\cdot)$. Define the sub-cycle time $C_i^{(PQ)}$ at PQ_i to be the intervisit time for PQ_i plus the previous service period at queue $T(i)$ (thus $C_i^{(PQ)} = I_i^{(PQ)} + V_{\sigma_{iT(i)}}$) and denote the corresponding LST by $\mathcal{C}_i^{(PQ)}(\cdot)$, for $i = 1, \ldots, M$.

Denote by W_j the delay incurred by an arbitrary customer at Q_j, and the corresponding LST by $\mathcal{W}_j(s)$, $j = 1, \ldots, N$. Throughout, W_j will be considered as a function of ρ, where the arrival rates are variable, while the service-time distributions and the ratios of the arrival rates are kept fixed. For each variable x that is a function of ρ, $\hat{x}$ denotes its value evaluated at $\rho = 1$. It is known that when $\rho \uparrow 1$, all queues become unstable. Therefore, we focus on the random variable $(1-\rho)W_j$ (referred to as the *scaled* delay at Q_j), and derive its limiting distribution when ρ tends to unity. Thus, the analysis is focused on the distribution of the random variables

$$\widetilde{W}_j := \lim_{\rho \uparrow 1}\ (1-\rho)W_j \quad (j = 1, \ldots, N), \tag{1}$$

where the limit is defined below. The main result of the paper is the derivation of the LST of the distribution of $\widetilde{W}_j$, denoted by $\widetilde{\mathcal{W}}_j(s) := \lim_{\rho \uparrow 1}\ \mathcal{W}_j(s(1-\rho))$. For a customer served at PQ_i, we define the delay at PQ_i, $W_i^{(PQ)}$, to be the time between its arrival into the system and the moment at which the customer starts service at PQ_i. We define

$$\widetilde{W}_i^{(PQ)} := \lim_{\rho \uparrow 1}\ (1-\rho)W_i^{(PQ)} \quad (i = 1, \ldots, M). \tag{2}$$

The LST of $\widetilde{W}_i^{(PQ)}$ is denoted by $\widetilde{\mathcal{W}}_i^{(PQ)}(s)$.

Finally, a sequence of real-valued random variables $\{X_n, n = 1, 2, \ldots\}$ is said to converge in distribution to a random variable X, denoted by $X_n \to_d X$, if there exists a dense subset A of $\mathcal{R}$ (i.e., of real numbers) such that $\lim_{n\to\infty} P(X_n < a) \to P(X < a)$, for all $a \in A$.

3 Preliminaries

3.1 Relation between delays and queue lengths at polling instants

Each customer arriving at Q_j is served at exactly one pseudo-queue, say PQ_i with $T(i) = j$. The specific pseudo-queue at which a given customer C arriving at Q_i is served depends on the position of the server when C arrives at the system. Let $\pi_i > 0$ be the fraction of customers at $Q_{T(i)}$ that is served at PQ_i, $i = 1, \dots, M$. Then the LST of the delay of an arbitrary customer arriving at Q_j can be expressed as follows: For $j = 1, \dots, N$, $Re\ s > 0$,

$$\mathcal{W}_j(s) = \sum_{i:T(i)=j} \pi_i\ \mathcal{W}_i^{(PQ)}(s). \tag{3}$$

The distribution of the delays at PQ_i, in turn, can be expressed in terms of the distribution of X_i, denoting the number of customers at PQ_i at a polling instant of PQ_i, with corresponding PGF $\mathcal{X}_i(\cdot)$, as follows: For $T(i) \in G$, $Re\ s > 0$,

$$\mathcal{W}_i^{(PQ)}(s) = \frac{(1-\rho)(\mathcal{X}_i(\mathcal{B}_{T(i)}(s)) - \mathcal{X}_i(1 - s/\lambda_{T(i)}))}{\pi_i r(s - \lambda_{T(i)}(1 - \mathcal{B}_{T(i)}(s)))}, \tag{4}$$

and for $T(i) \in E$, $Re\ s > 0$,

$$\mathcal{W}_i^{(PQ)}(s) = \frac{(1-\rho)(1 - \mathcal{X}_i(1 - s/\lambda_{T(i)}))}{\pi_i r(s - \lambda_{T(i)}(1 - \mathcal{B}_{T(i)}(s)))}. \tag{5}$$

From equations (3)-(5) it follows that the distribution of W_i can be expressed in terms of π_i and $\mathcal{X}_i(\cdot)$. The derivation of these quantities is discussed below. In Section 3.2 we discuss how the π_i can be obtained by solving a set of at most $M - N$ linear equations. In Section 3.3 it is discussed how the $\mathcal{X}_i(\cdot)$ can be determined by means of the so-called Descendant Set Approach.

3.2 Sets of linear equations for π_i

To derive a set of linear equations for π_i, it is convenient to relate $E[X_i]$ to $E[V_i]$. Recall that V_i is the duration of a service period of the server at PQ_i ($i = 1, \dots, M$). To this end, it is readily verified that, for $i = 1, \dots, M$,

$$E[V_i] = \varphi_{T(i)} E[X_i], \quad \text{where } \varphi_j := b_j^{(1)} \quad (j \in G), \quad \text{and} \quad \varphi_j := \frac{b_j^{(1)}}{1-\rho_j} \quad (j \in E). \tag{6}$$

The variables $E[V_i]$ can be obtained by solving the following set of linear equations: For $i = 1, \dots, M$,

$$E[V_i] = \lambda_{T(i)} \varphi_{T(i)} \left[\sum_{k=\sigma_{iT(i)}+1}^{i-1} (r_k + E[V_k]) + r_{\sigma_{iT(i)}} + E[V_{\sigma_{iT(i)}}] I_{\{T(i)\in G\}} \right], \tag{7}$$

supplemented with the balance equations

$$\sum_{i:T(i)=j} E[V_i] = \rho_j \frac{r}{1-\rho} \quad (j = 1, \ldots, N). \tag{8}$$

It is readily verified that, for $i = 1, \ldots, M$,

$$\pi_i = \frac{E[V_i]}{\sum_{k:T(k)=i} E[V_k]} = \frac{E[V_i]}{\rho_{T(i)} r/(1-\rho)}. \tag{9}$$

Notice that both $E[V_i]$ and $E[X_i]$ have a first-order pole at $\rho = 1$, and that π_i converges to a constant for $\rho \uparrow 1$. Define the corresponding heavy-traffic residues by, for $i = 1, \ldots, M$,

$$v_i := \lim_{\rho \uparrow 1} \, (1-\rho) E[V_i], \quad \hat{\pi}_i = \lim_{\rho \uparrow 1} \, \pi_i. \tag{10}$$

Using equations (7)-(10), it is easily verified that v_i $(i = 1, \ldots, M)$, are uniquely determined by solving the following set of linear equations: For $i = 1, \ldots, M$, $j = 1, \ldots, N$,

$$v_i = \hat{\lambda}_{T(i)} \hat{\varphi}_{T(i)} \left[\sum_{k=\sigma_{iT(i)}+1}^{i-1} v_k + v_{\sigma_{iT(i)}} I_{\{T(i) \in G\}} \right], \qquad \sum_{i:T(i)=j} v_i = \hat{\rho}_j r, \tag{11}$$

and for $i = 1, \ldots, M$,

$$\hat{\pi}_i = \frac{v_i}{\hat{\rho}_{T(i)} r}. \tag{12}$$

It is readily seen that that the values v_i and $\hat{\pi}_i$ can be effectively determined, by eliminating redundant equations, by solving a set of $M - N$ linear equations. In the sequel, it will be shown that the (scaled) delay distribution when the load tends to unity can be expressed explicitly in terms $\hat{\pi}_i$, $i = 1, \ldots, M$.

3.3 Descendant Set Approach to obtain queue lengths at polling instants

Relations (3)-(5) imply that the distribution of W_j is completely characterized by the distributions of X_i for all i with $T(i) = j$. The distribution of X_i, in turn, can be characterized by means of the Descendant Set Approach (DSA), exploring the branching structure of the evolution of the system (cf. [20]). The DSA is focused on the determination of the distribution of X_1 (without loss of generality), the number of customers at PQ_1 present at an arbitrary polling instant P^* at Q_1, referred to as the reference point. The key observation is that we can evaluate $X_1(P^*)$ by considering, recursively, contributions to X_1 from waiting customers at all queues of the past polling epochs, working backward from the reference point. Let $T_{i,c}$ be a customer served at PQ_i during the c-th cycle. Define the children set of $T_{i,c}$ to be the set of customers arriving during the service of $T_{i,c}$; the descendant set of $T_{i,c}$ is recursively defined to consist of $T_{i,c}$, its children, and the descendants of its children. Let $A_{i,c}$ be the number of customers at $Q_{T(1)}$ at the reference point (at PQ_1) that are descendants of $T_{i,c}$, and let $\mathcal{A}_{i,c}(\cdot)$ denote its Probability Generating Function (PGF). In this way, $\mathcal{A}_{i,c}$ can be viewed as the *contribution* of $T_{i,c}$ to $X_1(P^*)$. Denote by $R_{i,c}$ the switch-over time (from PQ_i to the next) immediately after the visit starting at $P_{i,c}$. Let $S_{i,c}$ be the total contribution to $\mathcal{X}_1(P^*)$ of all (original) customers that arrive in the system during $R_{i,c}$, and denote the corresponding PGF by $\mathcal{S}_{i,c}(\cdot)$. Then X_1 can be expressed as the independent sum $X_1 = \sum_{i=1}^{M} \sum_{c=0}^{\infty} S_{i,c}$, or equivalently, for $|z| \leq 1$,

$$\mathcal{X}_1(z) = \prod_{i=1}^{M} \prod_{c=0}^{\infty} \mathcal{S}_{i,c}(z), \tag{13}$$

where for $i = 1, \ldots, M$, $c = 0, 1, \ldots$, $|z| \leq 1$,

$$\mathcal{S}_{i,c}(z) = \mathcal{R}_i \left(\sum_{j:\tau_{ij} > i} \lambda_j (1 - \mathcal{A}_{\tau_{ij},c}(z)) + \sum_{j:\tau_{ij} \leq i} \lambda_j (1 - \mathcal{A}_{\tau_{ij},c-1}(z)) \right). \tag{14}$$

The Descendant Set (DS) variables satisfy the following set of relations, based on the observation that the contribution to $X_1(P^*)$ of a tagged customer $T_{i,c}$ is equal to the total contribution to $X_1(P^*)$ of the children of $T_{i,c}$: For $c = 0, 1, \ldots$, $|z| \leq 1$, $T(i) \in G$,

$$\mathcal{A}_{i,c}(z) = \mathcal{B}_{T(i)} \left(\sum_{j:\tau_{ij} > i} \lambda_j (1 - \mathcal{A}_{\tau_{ij},c}(z)) + \sum_{j:\tau_{ij} \leq i} \lambda_j (1 - \mathcal{A}_{\tau_{ij},c-1}(z)) \right), \tag{15}$$

and for $T(i) \in E$,

$$\mathcal{A}_{i,c}(z) = \Theta_{T(i)} \left(\sum_{j:\tau_{ij} > i} \lambda_j (1 - \mathcal{A}_{\tau_{ij},c}(z)) + \sum_{j:\tau_{ij} < i} \lambda_j (1 - \mathcal{A}_{\tau_{ij},c-1}(z)) \right), \tag{16}$$

where $\Theta_{T(i)}(\cdot)$ stands for the LST of the busy period in an M/G/1 system with arrival rate $\lambda_{T(i)}$ and service time distributions with LST $\mathcal{B}_{T(i)}(\cdot)$. Since we focus on the number of customers at Q_1 at the reference point, the initial conditions are, for $|z| \leq 1$,

$$\mathcal{A}_{1,-1}(z) = z, \quad \mathcal{A}_{i,-1}(z) = 1, \quad i = 2, \ldots, M, \tag{17}$$

Notice that relations (13)-(17) give a complete, but not explicit, characterization of the probability distribution of X_1. This characterization is useful to obtain properties of the limiting distribution of X_1. Using equations (3)-(5) these properties, in turn, will be used to derive properties of the distribution of $\widetilde{W}_i^{(PQ)}$, $i = 1, \ldots, M$, and hence of $\widetilde{W}_j, j = 1, \ldots, N$.

4 Analysis

In this section we derive expressions for the delay distributions at each of the queues. The expressions are in closed form, except that they require the determination of the variables $\hat{\pi}_i$, defined in (10). The derivation of the results proceeds along a number of steps, according to the DSA-based methodology developed in [40, 41, 43] for the case of cyclic server routing. Define the k-th factorial moment of the DSA-variables $A_{i,c}$ as follows: For $i = 1, \ldots, M$, $c = 0, 1, \ldots$, $k = 1, 2, \ldots$,

$$\alpha_{i,c}^{(k)} := E[A_{i,c}(A_{i,c} - 1) \cdots (A_{i,c} - k + 1)], \tag{18}$$

supplemented with the initial conditions $\alpha_{1,-1}^{(k)} := 1$ and $\alpha_{i,-1}^{(k)} = 0$ $(i = 2, \ldots, M), k = 1, 2, \ldots$. To start the derivation of the results, we decompose the sequences of DSA-variables $\{\alpha_{i,c}^{(k)}, c = 0, 1, \ldots\}$ into a dominant and a recessive part, in a sense to be specified (Theorem 1). The decomposition is then used to obtain an expression for the quantities $\tilde{h}_1^{(i,k)}$, to be defined in (41), see Theorem 2. This result, in turn, is used to obtain explicit expressions for the limiting moments $\tilde{x}_i^{(k)} := \lim_{\rho \uparrow 1}\ (1 - \rho)^k E[X_i^k]$, which are shown to match the moments of a gamma distribution with known parameters (Theorem 3). Then, the so-called Method of Moments is applied to show that $(1-\rho)X_i, (1-\rho)I_i^{(PQ)}$, and $(1-\rho)C_i^{(PQ)}$ converge to a gamma-distribution with known parameters in terms in the variables $\hat{\pi}_i$ (Theorems 4, 5, and 6). This, in turn, is shown to lead to exact expressions for the LST of $\widetilde{W}_i^{*(PQ)}(s)$, and via (3) to an expression for $\tilde{W}_j^*(s)$, which is the main result of the paper (Theorem 7). That the LSTs can be written explicitly in terms of known random variables (Corollary 7.1) further extends their possible applications.

4.1 Analysis of the DSA variables with $k = 1$

It is convenient to express relations for the $\alpha_{i,c}^{(1)}$ in matrix notation. Let $\underline{\alpha}_{\cdot,c}$ be the vector whose i-th element is $\alpha_{i,c}$, for $i = 1, \ldots, M$, $c = 1, 2, \ldots$ Moreover, let $\mathbf{P}_i$ be the M by M matrix whose (j,k)-th element equals $I_{\{j=k\}}$ for $j \neq i$, while the (i, τ_{ij})-th element equals $b_{T(i)}^{(1)} \lambda_j$ if $T(i) \in G$, and $I_{\{j \neq T(i)\}} \lambda_j^{(1)} b_{T(i)}/(1 - \rho_{T(i)})$ if $T(i) \in E$, and all other components of the i-th row are 0, $i = 1, \ldots, M$. Define $\mathbf{M} := \mathbf{P}_1 \cdots \mathbf{P}_M$, and let $\underline{q}$ be the vector whose i-th element is $\alpha_{i,0}$, for $i = 1, \ldots, M$. Then it is readily seen that

$$\underline{\alpha}_{\cdot,-1} = \underline{e}_1, \quad \underline{\alpha}_{\cdot,c} = \mathbf{M}\underline{\alpha}_{\cdot,c-1} = \mathbf{M}^{c+1}\underline{e}_1, \quad c = 0, 1, \ldots \tag{19}$$

The variables $\alpha_{i,c}$ are therefore determined by a set of first-order homogeneous difference equations. The theory of difference equations shows that the variables $\alpha_{i,c}^{(1)}$ can be solved explicitly if the eigenvalues and eigenvectors of $\mathbf{M}$ are known. In general, however, the eigenvalues and eigenvectors of $\mathbf{M}$ are unknown for $\rho < 1$. Nevertheless, to analyze the system behavior in the limiting case $\rho \uparrow 1$, there is no need to obtain all eigenvalues and eigenvectors of $\mathbf{M}$ and then let $\rho \uparrow 1$. Instead, one can separate $\mathbf{M}$ into dominant and non-dominant components as will be shown in Lemma 1.

In the analysis a key role is played by the vector $\underline{y} = (y_1, \ldots, y_M)$, where y_i is defined as follows:

$$y_i := \lambda_j \sum_{k=\sigma_{1j}}^{M} \pi_k \rho_{T(k)} \quad \text{if } i = \tau_{Mj} \text{ for some } j \in G, \tag{20}$$

$$y_i := \lambda_j \sum_{k=\sigma_{1j}+1}^{M} \pi_k \rho_{T(k)} \quad \text{if } i = \tau_{Mj} \text{ for some } j \in E, \tag{21}$$

and $y_i := 0$ otherwise.

Lemma 1 *The matrix* $\mathbf{M}$ *has maximal eigenvalue* ξ *which is real-valued, positive, has multiplicity 1, and has associated right and left eigenvectors* $\underline{u} = (u_1, \ldots, u_M)$ *and* $\underline{w} = (w_1, \ldots, w_M)$*. If these are normalized so that* $\sum_{i=1}^{M} u_i w_i = 1$ *then*

$$\mathbf{M}^c = \xi^c \underline{u}\, \underline{w}^\top + \mathbf{R}^c, \tag{22}$$

where there exist $K < \infty$ *and* $\bar{\xi}$ $(0 < \bar{\xi} < \xi)$*, such that all entries of* $\mathbf{R}^c$ *are strictly smaller than* $K\bar{\xi}^c$*. Further, for* $i = 1, \ldots, M$*,*

$$\xi < 1 \textit{ if and only if } \rho < 1; \quad \xi = 1 \textit{ if and only if } \rho = 1, \tag{23}$$

$$\lim_{\rho \uparrow 1} \; \xi = 1, \tag{24}$$

$$\lim_{\rho \uparrow 1} \; u_i = \hat{u}_i = b_{T(i)}^{(1)}, \tag{25}$$

$$\lim_{\rho \uparrow 1} \; w_i = \hat{w}_i = \frac{\hat{y}_i}{\sum_{i=1}^{M} b_{T(i)} \hat{y}_i}, \tag{26}$$

$$\lim_{\rho \uparrow 1} \; \frac{1-\xi}{1-\rho} = 1/\delta, \;\textit{ where }\; \delta := \sum_{i=1}^{M} b_{T(i)} \hat{y}_i = \sum_{j=1}^{N} \hat{\rho}_j \left(\sum_{k=\sigma_{1j}+1}^{M} \hat{\pi}_k \hat{\rho}_{T(k)} + \hat{\pi}_{\sigma_{1j}} \hat{\rho}_j I_{\{j \in G\}} \right). \tag{27}$$

Proof: Relations (23)-(24) follow directly from the Frobenius Theorem (cf., e.g., [1]), the stability assumption and the continuity of the eigenvectors as a function of the entries of $\mathbf{M}$ which, in turn, are continuous in ρ. The validity of equations (25)-(26) is verified by showing that $\underline{\hat{u}}$ and $\underline{\hat{w}}$ are right and left eigenvectors for $\mathbf{M}$ at $\rho = 1$ with eigenvalue 1, and the continuity of the eigenvalues and eigenvectors. To prove (27), we define $h_1^{(i,1)} := \sum_{c=0}^{\infty} \left(\sum_{j:\tau_{ij} > i} \lambda_j \alpha_{\tau_{ij},c}^{(1)} + \sum_{j:\tau_{ij} \leq i} \lambda_j \alpha_{\tau_{ij},c-1}^{(1)} \right)$ and $\tilde{h}_1^{(i,1)} := \lim_{\rho \uparrow 1} \; (1-\rho) h_1^{(i,1)}$. Then based on equations (19) it is readily verified that

$$\tilde{h}_1^{(i,1)} = \hat{\lambda}_{T(1)} \hat{\pi}_1 \quad (T(1) \in G), \qquad \text{and} \;\; \tilde{h}_1^{(i,1)} = \hat{\lambda}_{T(1)} \hat{\pi}_1 (1 - \hat{\rho}_{T(1)}) \quad (T(1) \in E), \tag{28}$$

independent of i. Alternatively, equations (19) and (22) imply that $\alpha_{i,c}^{(1)}$ can be expressed as

$$\alpha_{i,c}^{(1)} = \mathbf{e}_i^\top \mathbf{M}^{c+1} \mathbf{e}_1 = \xi^{c+1} \hat{u}_i \hat{w}_1 + \epsilon_{i,c}^{(1)}, \tag{29}$$

with $\epsilon_{i,c}^{(1)} < K\bar{\xi}^c$ for some $K < \infty$ and $0 < \bar{\xi} < \xi$. Using (11)-(12) and (20)-(21), it is readly verifed that

$$\hat{y}_1 = \hat{\lambda}_{T(1)} \left(\sum_{k=\sigma_{1T(1)}+1}^{M} \hat{\pi}_k \hat{\rho}_{T(k)} + \hat{\pi}_{\sigma_{1T(1)}} \hat{\rho}_{T(1)} I_{\{T(1)\in G\}} \right) = \hat{\rho}_{T(1)} \hat{\pi}_{T(1)} / \hat{\varphi}_{T(1)}. \tag{30}$$

The result then follows directly by combining (28)-(30) and several straightforward manipulations. □

4.2 Analysis of the DSA variables with $k \geq 1$

Lemma 1 and relation (29) decompose the sequences $\{\alpha_{i,c}^{(1)}, c = 0, 1, \ldots\}$ into two parts: a dominant and a recessive part. The dominant part, $f_{i,c}^{(1)} := \xi^{c+1}\hat{u}_i\hat{w}_1$, dominates the recessive part, $g_{i,c}^{(1)} := \epsilon_{i,c}^{(1)}$, in the sense that $\lim_{\rho\uparrow 1}\ (1-\rho)\sum_{c=0}^{\infty} f_{i,c}^{(1)} > 0$, whereas $\lim_{\rho\uparrow 1}\ (1-\rho)\sum_{c=0}^{\infty} g_{i,c}^{(1)} = 0$. The next result generalizes this concept of dominance to the case $k \geq 1$.

Theorem 1 *For $i = 1, 2, \ldots, M$, $c = 0, 1, \ldots$, $k = 1, 2, \ldots$,*

$$\alpha_{i,c}^{(k)} = f_{i,c}^{(k)} + \epsilon_{i,c}^{(k)}, \text{ with } f_{i,c}^{(k)} = \xi^c \sum_{j=0}^{k-1} \pi_{i,j}^{(k)} \xi^{jc}, \tag{31}$$

where for $j = 0, 1, \ldots, m-1$,

$$\text{(a)} \quad \lim_{\rho\uparrow 1} (1-\rho)^{k-1}\pi_{i,j}^{(k)} = (-1)^j \begin{pmatrix} k-1 \\ j \end{pmatrix} \frac{k!}{2^{k-1}} \frac{\hat{\lambda}_{T(1)}^k \hat{\pi}_1^k (1-\hat{\rho}_{T(1)} I_{\{T(1)\in E\}})^k}{\delta^k} \left(\frac{b^{(2)}}{b^{(1)}}\right)^{k-1} b_{T(i)}^{(1)}. \tag{32}$$

$$\text{(b)} \quad \sum_{c=0}^{\infty} \epsilon_{i,c}^{(k)} = O\left((1-\rho)^{-k-1}\right) \quad (\rho \uparrow 1). \tag{33}$$

Proof: Consider the case $T(i) \in G$. For notational convenience, define for $i = 1, \ldots, M, c = 0, 1, \ldots$, $k = 1, 2, \ldots$,

$$\gamma_{i,c}^{(k)} := \sum_{j:\tau_{ij}>i} \lambda_j \alpha_{\tau_{ij},c}^{(k)} + \sum_{j:\tau_{ij}\leq i} \lambda_j \alpha_{\tau_{ij},c-1}^{(k)}. \tag{34}$$

Then by repeatedly differentiating (15) we obtain the following recursive relations for the DS variables $\alpha_{i,c}^{(k)}$ (defined in (18)): For $c = 0, 1, \ldots$,

$$\alpha_{i,c}^{(1)} = b_{T(i)}^{(1)}\gamma_{i,c}^{(1)}, \quad \alpha_{i,c}^{(2)} = b_{T(i)}^{(1)}\gamma_{i,c}^{(2)} + b_{T(i)}^{(2)}\left(\gamma_{i,c}^{(1)}\right)^2, \tag{35}$$

and for $k = 3, 4, \ldots$ and $c = 0, 1, \ldots$,

$$\alpha_{i,c}^{(k)} = b_{T(i)}^{(1)}\gamma_{i,c}^{(k)} + b_{T(i)}^{(2)}\Gamma_{i,c}^{(k)} + \Delta_{i,c}^{(k)}, \tag{36}$$

with

$$\Gamma_{i,c}^{(k)} := \sum_{l=1}^{k-2} \zeta_{l,k}\gamma_{i,c}^{(l)}\gamma_{i,c}^{(k-l)}, \tag{37}$$

where the coefficients $\zeta_{l,k}$ in (37) are defined by the following recursive relations: For $k = 3, 4, \dots$ and $l = 0, 1, \dots, k-3$,

$$\zeta_{0,k} := 1, \quad \zeta_{k-2,k} := 3, \quad \zeta_{j,k} := \zeta_{j,k-1} + \zeta_{j-1,k-1}. \tag{38}$$

$\Delta_{i,c}^{(k)}$ in (36) is a linear combination of terms of the form $\prod_{l=1}^{k} \left(\gamma_{i,c}^{(p_l)}\right)^{q_l}$. The variables p_l and q_l are non-negative integers satisfying $\sum_{l=1}^{k} p_l q_l = k$ and $\sum_{l=1}^{k} q_l \geq 2$. We make the following two observations. First, one may prove via induction in k that $\sum_{c=0}^{\infty} \Gamma_{i,c}^{(k)}$ has a pole of the order k at $\rho = 1$. Second, one may show by induction in k that all terms in $\Delta_{i,c}^{(k)}$ are lower order in the sense that $\sum_{c=0}^{\infty} \Delta_{i,c}^{(k)}$ has a pole of order $\sum_{l=1}^{k}(p_l - 1)q_l + 1 < k$ at $\rho = 1$, and as such are of lower order in the limiting case. With these observations, the derivation of the result can be obtained via induction in k, along the lines as discussed in [40]. The details are omitted here for compactness of the paper. The results for $T(i) \in E$ can be obtained along similar lines. □

4.3 Analysis of the number of customers at a polling instant

In this section we use the results in Sections 4.1 and 4.2 to derive an expression for the limiting distribution of X_i, the number of customers at PQ_i at a polling instant of PQ_i. To this end, we use the results presented in Section 3.3 to obtain expressions for the limiting scaled moments of X_i. We then used these moments to derived limiting distributional results for appropriately scaled versions of X_i, $I_i^{(PQ)}$, and $C_i^{(PQ)}$, $i = 1, \dots, M$.

The following notation is useful. Defining $x_i^{(k)} := E[X_i^k]$, we will show (see Theorem 3) that $x_i^{(k)}$ possesses a k-th order pole at $\rho = 1$. Therefore, it is convenient to define, for $i = 1, \dots, M$, $k = 1, 2, \dots$,

$$\tilde{x}_i^{(k)} := \lim_{\rho \uparrow 1} \ (1-\rho)^k E[X_i^k], \tag{39}$$

referred to as the heavy-traffic residue of $E[X_i^k]$. Notice that $\tilde{x}_i^{(k)}$ is also equal to the heavy-traffic residue of the k-th factorial moment of X_i, i.e., $\tilde{x}_i^{(k)} = \lim_{\rho\uparrow 1} \ (1-\rho)^k E[X_i(X_i - 1)\cdots(X_i - k + 1)]$, for $i = 1, \dots, M$, $k = 1, 2, \dots$. To express the moments of X_i in terms of the moments of the descendent set variables $A_{i,c}$, we define for $i = 1, \dots, M$, $|z| \leq 1$,

$$\mathcal{H}_1^{(i)}(z) := \sum_{c=0}^{\infty} \left(\sum_{j:\tau_{ij} > i} \lambda_j (1 - \mathcal{A}_{\tau_{ij},c}(z)) + \sum_{j:\tau_{ij} \leq i} \lambda_j (1 - \mathcal{A}_{\tau_{ij},c-1}(z)) \right), \tag{40}$$

and the k-th order derivatives, for $k = 1, 2 \dots$,

$$h_1^{(i,k)} := \sum_{c=0}^{\infty} \left(\sum_{j:\tau_{ij} > i} \lambda_j \alpha_{\tau_{ij},c}^{(k)} + \sum_{j:\tau_{ij} \leq i} \lambda_j \alpha_{\tau_{ij},c-1}^{(k)} \right) \quad \text{and} \quad \tilde{h}_1^{(i,k)} := \lim_{\rho\uparrow 1} \ (1-\rho)^k h_1^{(i,k)}. \tag{41}$$

Theorem 2 *For $m = 1, \ldots, M, \ k = 1, 2, \ldots,$*

$$\tilde{h}_1^{(m,k)} = \frac{(k-1)!}{2^{k-1}} \frac{\hat{\lambda}_{T(1)}^k \hat{\pi}_1^k (1 - \hat{\rho}_{T(1)} I_{\{T(1) \in E\}})^k}{\delta^{k-1}} \left(\frac{b^{(2)}}{b^{(1)}} \right)^{k-1}, \ \textit{independent of } m. \tag{42}$$

Proof: The proof can be obtained from Lemma 1 and Theorem 1 and several straightforward manipulations. □

Theorem 3 *For $k = 1, 2, \ldots,$*

$$\tilde{x}_1^{(k)} = \hat{\lambda}_{T(1)}^k \hat{\pi}_1^k (1 - \hat{\rho}_{T(1)} I_{\{T(1) \in E\}})^k \prod_{m=0}^{k-1} \left(r + \frac{m}{2\delta} \frac{b^{(2)}}{b^{(1)}} \right). \tag{43}$$

Proof: For $k = 1$, it follows directly from (13)-(14), (28), and by taking the proper limits that

$$\tilde{x}_1^{(1)} = \sum_{i=1}^{M} r_i \tilde{h}_1^{(i,1)} = r \hat{\lambda}_{T(1)} \hat{\pi}_1 (1 - \hat{\rho}_{T(1)} I_{\{T(1) \in E\}}), \tag{44}$$

which shows the validity of (43) for $k = 1$. To show the result for $k > 1$, let us first assume that the switch-over times are deterministic, so that $\mathcal{R}_i(s) = \exp\{-s r_i\}$. Then it is readily seen that, for $|z| \leq 1$,

$$\mathcal{X}_1(z) = \exp \left\{ - \sum_{i=1}^{M} r_i \mathcal{H}_1^{(i)}(z) \right\}, \tag{45}$$

which is a composite function of the form $f(g(z))$ with $f(z) := e^{-z}$ and $g(z) := \sum_{i=1}^{M} r_i \mathcal{H}_1^{(i)}(z)$. Then by repeatedly applying standard rules for differentiating composite functions, and taking the proper limits, we obtain the following result: For $k = 1, 2, \ldots,$

$$\tilde{x}_1^{(k)} = \sum_{\underline{m}^{(k)} \in S_k} c_k(\underline{m}^{(k)}) \prod_{n=1}^{k} \left(\sum_{i=1}^{M} r_i \tilde{h}_1^{(i,n)} \right)^{m_n} = \sum_{\underline{m}^{(k)} \in S_k} c_k(\underline{m}^{(k)}) \ r^{l_k} \prod_{n=1}^{k} \left(\tilde{h}_1^{(1,n)} \right)^{m_n}, \tag{46}$$

where $l_k := \sum_{j=1}^{k} m_j$, and where S_k is the set of vectors $\underline{m}^{(k)} = (m_1, \ldots, m_k)$ with integers $m_j \geq 0$ satisfying $\sum_{j=1}^{k} j m_j = k$, and where the coefficients $c_k(\underline{m}^{(k)})$ are defined by the following recursive relations (cf. [43]): $c_1(1) := 1$, and for $k = 2, 3, \ldots,$

$$c_k(\underline{m}^{(k)}) = c_{k-1}(m_1 - 1, m_2, \ldots, m_{k-1}) I_{\{m_1 > 0\}} \tag{47}$$

$$+ \sum_{j=1}^{k-1} (m_j + 1) \, c_{k-1}(m_1, \ldots, m_{j-1}, m_j + 1, m_{j+1} - 1, m_{j+2}, \ldots, m_{k-1}) I_{\{m_{j+1} > 0\}}. \tag{48}$$

Note that the second equality in (46) follows directly from Theorem 2. The proof for the case of deterministic switch-over times is then readily obtained via induction in k by exploring the recursive relations (46)-(48) and using Theorem 2, following the lines of the proof for the case of cyclic server routing in [43].

To prove for the case of non-determinstic switch-over times, with higher moments $r_i^{(l)}$ $(l > 1)$, it suffices to show that $\tilde{x}_1^{(k)}$ is independent of $r_i^{(l)}$ for $l > 1$. To this end, define $s_{i,c}^{(k)}$ to be the k-th factorial moment of $S_{i,c}$ (see in Section 3.3). Then $s_{i,c}^{(k)}$ can be expressed in terms of the DSA-variables $\alpha_{i,c}^{(k)}$ (defined in (18)) as follows: For $i = 1, \ldots, M, c = 0, 1, \ldots, k = 1, 2, \ldots,$

$$s_{i,c}^{(k)} = r_i^{(1)} \left(\sum_{j:\tau_{ij}>i} \lambda_j \alpha_{\tau_{ij},c}^{(k)} + \sum_{j:\tau_{ij}\leq i} \lambda_j \alpha_{\tau_{ij},c-1}^{(k)} \right) + \zeta_{i,c}^{(k)} = r_i^{(1)} \gamma_{i,c}^{(k)} + \zeta_{i,c}^{(k)}, \tag{49}$$

where $\zeta_{i,c}^{(k)}$ is a linear combination of terms of the form $r_i^{(l_k)} \prod_{l=1}^{k} \left(\gamma_{i,c}^{(k)}\right)^{m_l}$, with $\underline{m}^{(k)} = (m_1, \ldots, m_k) \in S_k$ and $l_k = \sum_{j=1}^{k} m_j > 1$. Note that the second equality in (49) follows directly from the definition of $\gamma_{i,c}^{(k)}$ in (34). Then, using (49), Lemma 1 and Theorem 2 one may verify that

$$\lim_{\rho\uparrow 1} \; (1-\rho)^k \sum_{i=1}^{M} \sum_{c=0}^{\infty} s_{i,c}^{(k)} = \sum_{i=1}^{M} r_i \tilde{h}_1^{(i,k)} + 0 = r\tilde{h}_1^{(1,k)}, \tag{50}$$

where the terms $\zeta_{i,c}^{(k)}$ are of lower order in the sense that $\sum_{i=1}^{M} \sum_{c=0}^{\infty} \zeta_{i,c}^{(k)} = O\left((1-\rho)^{-(k-1)}\right)$ for $\rho \uparrow 1$. To proceed, by exploring (13) one may verify that the k-th factorial moment of X_1 can be expressed as a linear combination of terms of the form $\prod_{p=1}^{k} \Gamma_p^{n_p}$, with $\underline{n}^{(k)} = (n_1, \ldots, n_k) \in S_k$, where Γ_p is a linear combination of terms of the form $\sum_{i=1}^{M} \sum_{c=0}^{\infty} \prod_{j=1}^{p} \left(s_{i,c}^{(j)}\right)^{m_j}$, with $\underline{m}^{(p)} \in S_p$, $p = 1, \ldots, k$. Then, by using (49) and Theorem 1 one may verify that, for $\underline{m}^{(p)} \in S_p$, $\sum_{i=1}^{M} \sum_{c=0}^{\infty} \prod_{j=1}^{p} \left(s_{i,c}^{(j)}\right)^{m_j} = O\left((1-\rho)^{-(p-l_p+1)}\right)$, $\rho \uparrow 1$. Consequently, the k-th factorial moment of X_1 can be expressed as a linear combination of terms of the form $\prod_{p=1}^{k} \Gamma_p^{n_p}$, $\underline{n}^{(k)} \in S_k$, that have a pole at $\rho = 1$ of the order $\sum_{p=1}^{k} n_p(p-l_p+1) = k - \sum_{p=1}^{k} n_p(l_p-1)$. Then, premultiplying the k-th factorial moment of X_1 by a factor $(1-\rho)^k$ and letting $\rho \uparrow 1$, all cross-terms for which there exists p $(1 \leq p \leq k)$ such that $n_p > 0$ and $l_p := \sum_{j=1}^{p} m_j > 1$, with $\underline{m}^{(p)} \in S_p$, vanish in the limiting case. The only terms that do not vanish in the limiting case are those for which $\sum_{p=1}^{k} n_p(l_p - 1) = 0$, i.e., for which for each p with $n_p > 0$ we have $l_p = 1$, or equivalently, $m_p = 1$, $m_j = 0$ for $j \leq p-1$. Consequently, $\tilde{x}_1^{(k)}$ is independent of the higher moment $r_i^{(l)}$ $(l > 1)$ of the switch-over times. This completes the proof. □

We have therefore characterized the moments $\tilde{x}_1^{(k)}, k = 1, 2, \ldots$. Below we will show that these moments correspond to those of a gamma random variable defined as follows.

Definition 1 *A random variable Γ with a gamma-distribution with scale parameter $\alpha > 0$ and rate parameter $\mu > 0$ has the following probability density function:*

$$f_\Gamma(t) := \frac{1}{\Gamma(\alpha)} e^{-\mu t} \mu^\alpha t^{\alpha-1}, \quad t \geq 0, \quad \text{where} \;\; \Gamma(\alpha) := \int_0^\infty e^{-t} t^{\alpha-1} dt. \tag{51}$$

It is readily verified that the LST and the moments of Γ are given by

$$\Gamma(s) = \left(\frac{\mu}{\mu+s}\right)^\alpha \quad (Re\ s > 0), \quad \text{and} \quad E[\Gamma^k] = \frac{\prod_{m=0}^{k-1}(\alpha+m)}{\mu^k} \quad (k = 1, 2, \ldots). \tag{52}$$

From (43) it can be seen that the limiting moments $\tilde{x}_1^{(k)}$ are identical to those of a gamma-distribution with parameters $\alpha = 2r\delta b^{(1)}/b^{(2)}$, and $\mu = 2\delta b^{(1)}/(\hat{\lambda}_{T(1)}\hat{\pi}_1(1-\hat{\rho}_{T(1)}I_{\{T(1)\in E\}})b^{(2)})$. The following result shows that this convergence in moments implies that $(1-\rho)X_i$ also converges *in distribution* to the gamma-distribution with the same parameters.

Theorem 4 *For $i = 1, \ldots, M$*

$$(1-\rho)X_i \rightarrow_d \tilde{X}_i \quad (\rho \uparrow 1), \tag{53}$$

where $\tilde{X}_i$ has a gamma-distribution with parameters

$$\alpha := 2r\delta\frac{b^{(1)}}{b^{(2)}} \text{ and } \mu_i := \frac{2\delta}{\hat{\lambda}_{T(i)}(1-\hat{\rho}_{T(i)}I_{\{T(i)\in E\}})\hat{\pi}_i}\frac{b^{(1)}}{b^{(2)}}. \tag{54}$$

Proof: The proof can be obtained by application of the so-called Method of Moments (cf., e.g., [8]), stating that convergence in moments, under some assumptions on the moments $\tilde{x}_1^{(k)}$, implies convergence in distribution. Based on Theorem 3 it is readily verified that these assumptions are indeed met, so that the Methods of Moments is applicable, which immediately leads to the result for $i = 1$. The result for arbitrary i follows then directly by noting that in the DSA-based approach discussed in Section 3, and also in Theorems 1 to 3, we assumed $i = 1$ without loss of generality. □

Theorem 5 *If $T(i) \in E$, then*

$$(1-\rho)I_i^{(PQ)} \rightarrow_d \tilde{I}_i^{(PQ)} \quad (\rho \uparrow 1), \tag{55}$$

where $\tilde{I}_i^{(PQ)}$ has a gamma-distribution with parameters

$$\alpha := 2r\delta\frac{b^{(1)}}{b^{(2)}} \text{ and } \mu_i := \frac{2\delta}{(1-\hat{\rho}_{T(i)})\hat{\pi}_i}\frac{b^{(1)}}{b^{(2)}}. \tag{56}$$

Proof: For the case of exhaustive service at PQ_i, the customers present at a polling instant at PQ_i are exactly those who arrived during the preceding intervisit period. This implies, for $|z| \leq 1$, $\mathcal{X}_i(z) = \mathcal{I}_i^{(PQ)}(\lambda_{T(i)}(1-z))$, or equivalently $\mathcal{I}_i^{(PQ)}(s) = \mathcal{X}_i(1 - s/\lambda_{T(i)})$, for $Re\ s > 0$. Then it is readily verified that, for $k = 1, 2, \ldots$, $\lim_{\rho\uparrow 1}\ (1-\rho)^k E\left[(I_i^{(PQ)})^k\right] = \tilde{x}_i^{(k)}/\hat{\lambda}_{T(i)}^k$. The result then follows directly by applying the Method of Moments addressed in the proof of Theorem 4. □

Theorem 6 *If $T(i) \in G$, then*

$$(1-\rho)C_i^{(PQ)} \rightarrow_d \tilde{C}_i^{(PQ)} \quad (\rho \uparrow 1), \tag{57}$$

where $\tilde{C}_i^{(PQ)}$ has a gamma-distribution with parameters

$$\alpha := 2r\delta\frac{b^{(1)}}{b^{(2)}} \text{ and } \mu_i := \frac{2\delta}{\hat{\pi}_i}\frac{b^{(1)}}{b^{(2)}}. \tag{58}$$

Proof: Similar to that of Theorem 5 except that $\mathcal{X}_i(z) = \mathcal{C}_i^{(PQ)}(\lambda_{T(i)}(1-z))$, for $|z| \leq 1$, or equivalently, $\mathcal{C}_i^{(PQ)}(s) = \mathcal{X}_i(1 - s/\lambda_{T(i)})$, $i = 1, \ldots, M$, $Re\ s > 0$. □

5 Main Result

We are now ready to present the main results of the paper.

Theorem 7 (Main result)

For $j = 1, \ldots, N$,

$$(1-\rho)\mathcal{W}_j \rightarrow_d \widetilde{W}_j \quad (\rho \uparrow 1), \tag{59}$$

where the Laplace-Stieltjes Transform of $\widetilde{W}_j$ *is given by*

$$\widetilde{\mathcal{W}}_j(s) = \sum_{i:T(i)=j} \hat{\pi}_i \widetilde{\mathcal{W}}_i^{(PQ)}(s), \tag{60}$$

where for Re $s > 0$,

$$\widetilde{\mathcal{W}}_i^{(PQ)}(s) = \frac{1}{(1-\hat{\rho}_{T(i)})\hat{\pi}_i r s} \left\{ \left(\frac{\mu_i}{\mu_i + s\hat{\rho}_{T(i)}} \right)^{\alpha} - \left(\frac{\mu_i}{\mu_i + s} \right)^{\alpha} \right\} \quad (T(i) \in G), \tag{61}$$

and

$$\widetilde{\mathcal{W}}_i^{(PQ)}(s) = \frac{1}{(1-\hat{\rho}_{T(i)})\hat{\pi}_i r s} \left\{ 1 - \left(\frac{\mu_i}{\mu_i + s} \right)^{\alpha} \right\} \quad (T(i) \in E), \tag{62}$$

with

$$\alpha := 2r\delta \frac{b^{(1)}}{b^{(2)}}, \quad \mu_i := \frac{2\delta}{\hat{\pi}_i \left(1 - \hat{\rho}_{T(i)} I_{\{T(i)\in E\}}\right)} \frac{b^{(1)}}{b^{(2)}}, \tag{63}$$

and

$$\delta := \sum_{j=1}^{N} \hat{\rho}_j \left(\sum_{k=\sigma_{1j}+1}^{M} \hat{\pi}_k \hat{\rho}_{T(k)} + \hat{\pi}_{\sigma_{1j}} \hat{\rho}_j I_{\{j\in G\}} \right). \tag{64}$$

Proof: The result is obtained directly by combining Lemma 1, Theorem 4, equations (3)-(5) and several straightforward manipulations. □

It turns out that the LSTs in Theorem 7 can be written explictly in terms of known random variables. Before doing so, we need a few definitions. In particular, if X is some random variable with probability density function (p.d.f.) $f_X(x)$ and finite expectation EX then we define the *length-biased* (or time-averaged) random variable $\mathbf{X}$ (see, for example, Baccelli and Brémaud [2]) as a random variable with p.d.f. $f_{\mathbf{X}}(x) = x f_X(x)/EX$. See [2] for a more general definition if the random variable has no p.d.f. or infinite expectation. Therefore $\tilde{\mathbf{I}}_k^{(PQ)}$ and $\tilde{\mathbf{C}}_k^{(PQ)}$ represent the length-biased versions of $\tilde{I}_k^{(PQ)}$ and $\tilde{C}_k^{(PQ)}$, respectively, $k = 1, \ldots, M$. It is straightforward to show that if a gamma random variable has parameters α and μ then its length biased version has parameters $\alpha + 1$ and μ. Therefore the parameters of $\tilde{\mathbf{I}}_k^{(PQ)}$ and $\tilde{\mathbf{C}}_k^{(PQ)}$ may be found directly from Theorems 5 and 6 for the exhaustive and gated cases, respectively. These properties yield the following elegant form for the delay distribution.

Corollary 7.1

For $T(i) \in G$,

$$(1-\rho)W_i^{(PQ)} \rightarrow_d (\hat{\rho_{T(i)}} + (1-\hat{\rho_{T(i)}})U)\tilde{\mathbf{C}}_i^{(PQ)} \quad (\rho \uparrow 1), \tag{65}$$

and for $T(i) \in E$,

$$(1-\rho)W_i^{(PQ)} \rightarrow_d U\tilde{\mathbf{I}}_i^{(PQ)} \quad (\rho \uparrow 1), \tag{66}$$

where U is a uniform[0,1] random variable and is independent of $\tilde{\mathbf{C}}_i^{(PQ)}$ and $\tilde{\mathbf{I}}_i^{(PQ)}$.

Proof: The result is obtained directly by combining the LSTs in Theorems 5, 6 and 7.

Remark 5.1

The elegant form in Corollary 7.1 is not surprising given the Fuhrmann and Cooper [16] decomposition. It can be shown that the Fuhrmann and Cooper [16] decomposition applied to periodic polling models satisfying our assumptions can be written as $W_j =_d U\mathbf{I}_j$ for $j \in E$ where W_j is queue-specific delay, $\mathbf{I}_j$ is the length-biased intervisit time for queue j, and U is a uniform[0,1] random variable. The case of gated random variable is similar but slightly more complicated. In this case the Fuhrmann and Cooper [16] decomposition implies that $W_j =_d U\mathbf{I}_j + \mathbf{V}_j$ for $j \in G$ where U, $\mathbf{I}_j$ are as previously defined and $\mathbf{V}_j$ is the length-biased visit time for queue j. If it were true that $\mathbf{I}_j =_d (1-\rho_j)\mathbf{C}_j$ and $\mathbf{V}_j =_d \rho_j\mathbf{C}_j$ (as in true in expectation for the non-length-biased random variables) then the result would follow.

Remark 5.2

The expressions in Corollary 7.1 are remarkably easy to calculate. First note that the right hand side of equation (65) can be written as $\mathcal{U}\tilde{\mathbf{C}}_i^{(PQ)}$

where $\mathcal{U}$ is uniform on $[\hat{\rho}_{T(i)}; 1]$. By direct manipulation of the distribution functions it is easy to show that, for $T(i) \in G$,

$$\begin{aligned}\text{Prob}\left\{\mathcal{U}\tilde{\mathbf{C}}_i^{(PQ)} \leq x\right\} &= \\ F_{\alpha+1,\mu_i}(x) + \frac{x\mu_i}{\alpha(1-\rho_i)}(F_{\alpha,\mu_i}(x/\rho_i) - F_{\alpha,\mu_i}(x)) &- \frac{\rho_i}{1-\rho_i}(F_{\alpha+1,\mu_i}(x/\rho_i) - F_{\alpha+1,\mu_i}(x))\end{aligned} \tag{67}$$

and, for $T(i) \in E$,

$$\text{Prob}\left\{U\tilde{\mathbf{I}}_i^{(PQ)} \leq x\right\} = F_{\alpha+1,\mu_i}(x) + \frac{x\mu_i}{\alpha}(1 - F_{\alpha,\mu_i}(x)), \tag{68}$$

where $F_{\alpha,\mu}(x)$ is the cumulative distribution function of a Gamma(α, μ) random variable and α and μ_i are defined in Theorem 7.

Corollary 7.2

Define the moments of the limiting delay distribution as

$$\omega_j^{(k)} := \lim_{\rho \uparrow 1} (1-\rho)^k E[W_j^k]. \tag{69}$$

Then, for $k = 1, 2, \ldots$,

$$\omega_j^{(k)} = \frac{1}{k+1} \sum_{i:T(i)=j} \hat{\pi}_i^{k+1} \left(1 + \hat{\rho}_{T(i)} + \cdots + \hat{\rho}_{T(i)}^k\right) \prod_{m=1}^{k} \left(r + \frac{m}{2\delta} \frac{b^{(2)}}{b^{(1)}}\right) \quad (j \in G), \tag{70}$$

$$\omega_j^{(k)} = \frac{1}{k+1} \sum_{i:T(i)=j} \hat{\pi}_i^{k+1} \left(1 - \hat{\rho}_{T(i)}\right)^k \prod_{m=1}^{k} \left(r + \frac{m}{2\delta} \frac{b^{(2)}}{b^{(1)}}\right) \quad (j \in E), \tag{71}$$

where δ is defined in (64).

Proof: Follows directly from taking expectations in Corollary 7.1 (or by repeatedly differentiating relations (60)-(62), and substituting $s = 0$).

Corollary 7.2 suggests simple and fast approximations for the moments of the delay in stable systems (i.e., with $\rho < 1$). The accuracy of the approximations is discussed in more details in Section 7.

Remark 5.3

The results in Theorem 7 generalize the results in [40], where we consider the special case cyclic server routing. In that case, it is readily verified that $M = N$, $T(i) = i$ $(i = 1, \ldots, N)$, $v_i = \hat{\rho}_i r$, $\pi_i = \hat{\pi}_i = 1$ $(i = 1, \ldots, N)$, and $\delta = [1 - \sum_{m \in E} \hat{\rho}_m^2 + \sum_{m \in G} \hat{\rho}^2]/2$. The results in (59)-(64) and (69)-(71) generalize those in [42], where we consider the first moments of the delay in polling systems with non-cyclic periodic server routing. In addition, the results in this paper are in line with a number of other known results in special cases.

Remark 5.4

Federgruen and Katalan [14] propose a numerical method to approximate the queue-length and delay distributions in polling systems similar to those considered here. The method is based on fitting the first few moments of the intervisit times and cycle times to their exact values, which can be determined numerically by means of the Descendant Set Approach (DSA) [20]. Interestingly, numerical experiments in [14] show that highly accurate approximations are obtained by fitting the *first two* moments of the intervisit times and cycle times only. Theorems 5 and 6 given insight into why such an approximation may be so accurate.

6 Implications and Extensions of the Results

6.1 Insensitivity Properties

Theorem 7 reveals a number of asymptotic properties of the queueing behavior in a general parameter setting, providing new insight in the behavior of polling models with periodic server routing schemes. In this section we discuss these properties in more detail.

Corollary 7.3

For $j = 1, \ldots, N$, the probability distribution of $\widetilde{W}_j$

1. *depends on the switch-over time distributions only through r, the total mean switch-over time per cycle;*
2. *depends on the second moments of the service-time distributions only through $b^{(2)}$, i.e., the second moment of the service time of an "arbitrary" customer, and*
3. *is independent of l-th moment of the service-time distributions at the queues for $l > 2$.*

Corollary 7.3 is generally not valid for stable systems (i.e., for $\rho < 1$), where the individual switch-over time distributions and the higher moments of the service-time distributions *do* have an impact on the distribution of the delays. Hence, Corollary 7.3 shows that the influence of these parameters on the delay distributions *vanishes* when the load tends to unity, and as such can be viewed as *lower-order* effects in heavy traffic.

Item (1) in Corollary 7.3 has some interesting implications for limiting versions of the system. First, consider the limiting delay distribution when switch-over times are reduced to zero. How the individual switch-over time distributions converge to 0 is irrelevant to this limit. This is unlike previous work in the case of zero switch-over times where the expected delay may not even be uniquely determined (cf. [6, 13]). Apparently, the differences in the expected delays caused by different limiting regimes (for $r \to 0$) vanish when the system reaches saturation. Second, consider the limiting delay distribution when switch-over times are increased without bound. It is readily verified from Theorems 5 and 6 that both $\tilde{I}_i^{(PQ)}/r$ and $\tilde{C}_i^{(PQ)}/r$ converge to constants when r tends to infinity. Similar limits occurred in Olsen [28] and Van der Mei [39] when switch-overs were taken to be deterministic. When the system reaches saturation the switch-overs can now be stochastic for similar limits to occur. This extends the results in Olsen [28] where heavy-traffic then heavy-switch-over limits were taken for cyclic exhaustive service systems.

6.2 Closed-form expressions for "well-behaved" routing schemes

In this section we present closed-form expressions for the complete asymptotic distribution of the delay for several "well-behaved" routing schemes, in a general parameters setting. In this section, we present the results for the *star* polling scheme, with visit order $T = (1, 2, 1, 3, \ldots, 1, N)$, and the *elevator* polling scheme (cf., e.g. [38]), with visit order $T = (1, 2, \ldots, N, N, N-1, \ldots, 1)$.

Corollary 7.4 (Star polling)

For the star polling configuration, the distribution of $\widetilde{W}_j$ is given by equations (60)-(64), with

(1) *if k even, then $\hat{\pi}_k = 1$;*

(2) *if k odd, then*

$$\hat{\pi}_k = \frac{\hat{\rho}_{T(k-1)}}{1-\hat{\rho}_1} \quad (1 \in E), \qquad \hat{\pi}_k = \frac{\sum_{j=1}^{N-1} \hat{\rho}_1^{j-1} \hat{\rho}_{k-j}}{1-\hat{\rho}_1^{N-1}} \quad (1 \in G). \tag{72}$$

Corollary 7.5 (Elevator polling)

For the elevator polling order, the distribution of $\widetilde{W}_j$ is given by equations (60)-(64), and

$$\hat{\pi}_k = \frac{\sum_{i=1}^{k-1} \hat{\rho}_i}{1-\hat{\rho}_k}, \quad \hat{\pi}_{2N-k+1} = \frac{\sum_{i=k+1}^{N} \hat{\rho}_i}{1-\hat{\rho}_k} \quad (k \in E), \tag{73}$$

and

$$\hat{\pi}_k = \frac{\sum_{i=1}^{k} \hat{\rho}_i}{1+\hat{\rho}_k}, \quad \hat{\pi}_{2N-k+1} = \frac{\sum_{i=k}^{N} \hat{\rho}_i}{1+\hat{\rho}_k} \quad (k \in G). \tag{74}$$

We emphasize that the results presented here do not pose any additional symmetry restriction on the parameters, making the results applicable in a general parameter setting. To the best of our knowledge, similar closed-form expressions have not been presented in the literature in this general context.

For other general polling schemes our results require the solution of $M - N$ linear equations, which can hardly be considered onerous. However, if an explicit form is absolutely necessary then one may consider approximating v_k by $\rho_{T(k)} r / m_{T(k)}$ $(k = 1, \ldots, M)$, where m_i is the number of times queue i appears on the polling table, $i = 1, \ldots, M$. As one recommended method for designing polling tables is to balance the table so that the intervisit times have similar lengths (see, e.g., Federgruen and Katalan [15]) this approximation will work particularly well in such well-designed tables.

6.3 Generalization of Transient Approximations

The results in this paper are for steady-state systems. An alternate (more complex) approach to obtain these results would be to first derive transient heavy-traffic results and then obtain the steady-state limit of the limiting transient distribution. This is the approach taken by Coffman et al. [9], [10]. Coffman et al. [9] considers a model with cyclic exhaustive service at all queues and with zero switch-over times. Coffman et al. [10] (CPR) extends the results of [9] to nonzero switch-over times. The results are proven for the two-queue model and conjectured for the N queue model. Here we extend their N queue conjecture to the periodic gated and exhaustive service models of this paper. We show how this conjecture relates to the results of the previous section.

The results of CPR differ from traditional heavy-traffic work (see, e.g., Harrison and Nguyen [18] for an overview) in two fundamental ways. First, the drift term for the limiting diffusion contains an extra term due to the presence of switch-overs. Second, in order to find an individual queue's workload, the averaging principle is substituted for the snapshot principle. Below we discuss these differences and show how they carry over to the models in this paper.

To obtain heavy-traffic limits, a sequence of systems $n = 1, 2, ...$ is considered with

$$\lim_{n\to\infty} \sqrt{n}(\rho^n - 1) = a,$$

where ρ^n is the utilization in the nth system. In traditional models, the heavy-traffic limit for the workload process is a reflected Brownian motion with instantaneous drift a and instantaneous variance

$$\sigma^2 = \sum_{i=1}^{N} \hat{\lambda}_i (Var[B_i] + \hat{\rho}_i^2 Var[\hat{A}_i]),$$

where $Var[B_i]$ and $Var[A_i]$ are the variances of the service and interarrival times, respectively. In the case of Poisson arrivals this simplifies to $\sigma^2 = \hat{\rho} b^{(2)}/b^{(1)} = b^{(2)}/b^{(1)}$. This limit holds regardless of service order so long as the system is work conserving. Indeed, it holds in the case of Coffman et al. [9] where switch-over times are zero. However, when switch-overs are introduced, the system is no longer work-conserving and CPR show that the instantaneous drift is given by

$$d(x) = a + r(x) \tag{75}$$

where $r(x)$ is the scaled limiting value of

$$\frac{E[\text{time spent in switch-overs}]}{E[\text{duration of a cycle}]} \tag{76}$$

when there is a scaled amount x of unfinished work in the system. The infinitesimal variance is the same as that for any work conserving N queue system and does not depend on the switch-over times.

In traditional heavy-traffic models the snap-shot principle holds and one may deterministically predict how much work there is in each queue given the workload level. For example, in priority queues all work is kept in the lowest priority queue. Under FIFO service, if x is the total work in the system then each queue will have $\hat{\rho}_i x$ work. However, in systems under polling the work in each queue is emptied and refilled at a rate that is $(1-\rho)^{-1}$ faster than the rate that workload is changing. CPR provide the averaging principle which implies that during the course of a cycle total scaled workload is effectively constant and the individual queues' workloads are defined by a deterministic trajectory. This trajectory may be thought of as a fluid model where work is flowing in at constant rate $\hat{\rho}_i$ to each queue i, $1 \leq i \leq N$, and flows out at rate 1 whenever the server is not performing a switch-over.

One may conjecture that similar results should hold for periodic service models. In particular, the averaging principle should hold (but with a different fluid trajectory) and scaled work in the system should converge to a diffusion with instantaneous drift having a similar form to (75) and instantaneous variance σ^2 (as variance remained unchanged going from zero switch-overs to switch-overs it should still remain unchanged when service changes from cyclic to periodic). It therefore remains to conjecture the appropriate parameters for the model.

By the averaging principle, total workload in the system may be regarded as unchanged over the course of a cycle. Further, the distribution of that workload is in proportion to a fluid model. Thus, when there is a total of x units in the system the scaled visit times at each queue will be $v_i c(x)/c$, where $c(x)$ is the cycle time (as a function of x) and c is the long-run average cycle time. Consider the scaled work at the start of service at pseudo-queue 1. As switch-over times are negligible, this is precisely $\delta c(x)$. Using this in equations (75) and (76) we conjecture that

$$d(x) = a + r\delta/x \tag{77}$$

When service is cyclic and exhaustive this corresponds to the value found by CPR (where δ in our notation corresponds to ϱ in their's). For $a < 1$ this process converges to a gamma random variable with $\alpha = 2r\delta/\sigma^2 + 1$ and $\mu = 2|a|/\sigma^2$. In traditional heavy-traffic fashion, these results imply that steady-state workload may be approximated by a gamma random variable with parameters $\alpha = 2r\delta/\sigma^2 + 1$ and $\mu = 2(1-\rho)/\sigma^2$.

We can use the above analysis to find expressions for the cycle and intervisit times in this system. As before, when there is a scaled amount x work in the system scaled cycle time equals x/δ. Therefore, applying the averaging principle and the same reasoning as in Section 3.2, sub-cycle time i has length $\hat{\pi}_i x/\delta$ and intervisit time i has length $(1-\hat{\rho}_i)\hat{\pi}_i x/\delta$. We are observing these random variables from a time-averaged or length-biased perspective. Thus the heavy-traffic approximation for $\mathbf{I}_i^{(PQ)}$ is a gamma random variable with parameters $\alpha = 2r\delta/\sigma^2+1$ and $\mu = 2(1-\rho)\delta/((1-\hat{\rho}_i)\hat{\pi}_i\sigma^2)$ and the heavy-traffic approximation for $\mathbf{C}_i^{(PQ)}$ is a gamma random variable with parameters $\alpha = 2r\delta/\sigma^2+1$ and $\mu = 2(1-\rho)\delta/(\hat{\pi}_i\sigma^2)$. When arrivals are Poisson the scaled version of these values correspond *exactly* to those found in Theorems 5 and 6.

It therefore remains to find the delay for the individual queues. By calling on the averaging principle, a fluid analysis suffices. To this end, consider the sub-cycle time for pseudo-queue i, $i = 1, \ldots, M$, of a fluid system under exhaustive service with intervisit time of $\mathbf{I}_i^{(PQ)}$ and visit time of $\mathbf{V}_i$. (This represents a slight abuse of notation as we are considering the deterministic versions of these values where they previously represented steady-state random values.) Note that by definition of the fluid system $\mathbf{V}_i = \hat{\rho}_{T(i)}\mathbf{I}_i^{(PQ)}/(1-\hat{\rho}_{T(i)})$. We wish to find the distribution of delay, $W_i^{(PQ)}$. A particle of class i work will arrive during $\mathbf{I}_i^{(PQ)}$ with probability $(1-\hat{\rho}_{T(i)})$ and during $\mathbf{V}_i$ with probability $\hat{\rho}_{T(i)}$. If $u_1\mathbf{I}_i^{(PQ)}$ time units of $\mathbf{I}_i^{(PQ)}$ have passed when it arrives $(0 \le u_1 \le 1)$ then it will find $\hat{\rho}_{T(i)}u_1\mathbf{I}_i^{(PQ)}$ work ahead of it and the delay will be $\mathbf{I}_i^{(PQ)}(1-u_1) + \hat{\rho}_{T(i)}u_1\mathbf{I}_i^{(PQ)}$. If, for $0 \le u_2 \le 1$, it arrives $u_2\hat{\rho}_{T(i)}/(1-\hat{\rho}_{T(i)})\mathbf{I}_i^{(PQ)}$ of the time into the visit time (which has length $\hat{\rho}_{T(i)}/(1-\hat{\rho}_{T(i)})\mathbf{I}_i^{(PQ)}$) then the delay will be $\hat{\rho}_{T(i)}(\mathbf{I}_i^{(PQ)} + u_2\hat{\rho}_{T(i)}/(1-\hat{\rho}_{T(i)})\mathbf{I}_i^{(PQ)}) - u_2\hat{\rho}_{T(i)}/(1-\hat{\rho}_{T(i)})\mathbf{I}_i^{(PQ)}$ (the work that arrived ahead of it minus that which has already been processed). Now arriving work is evenly distributed over each period so the delay can be written as

$$P(W_i^{(PQ)} \le t) = \begin{cases} P(1-(1-\hat{\rho}_{T(i)})U_1 \le t/\mathbf{I}_i^{(PQ)}) & \text{with probability } (1-\hat{\rho}_{T(i)}) \\ P(\hat{\rho}_{T(i)}(1-U_2) \le t/\mathbf{I}_i^{(PQ)}) & \text{with probability } \hat{\rho}_{T(i)} \end{cases} \tag{78}$$

where U_1 and U_2 are uniform random variables on $[0,1]$. Let $x = t/\mathbf{I}_i^{(PQ)}$ then

$$P(W_i^{(PQ)} \le \mathbf{I}_i^{(PQ)}x) \quad = \quad \begin{cases} P(1-(1-\hat{\rho}_{T(i)})U_1 \le x) & \text{with probability } (1-\hat{\rho}_{T(i)}) \\ P(\hat{\rho}_{T(i)}(1-U_2) \le x) & \text{with probability } \hat{\rho}_{T(i)} \end{cases} \tag{79}$$

$$= \begin{cases} (1-\hat{\rho}_{T(i)}) \times 0 + \hat{\rho}_{T(i)} x / \hat{\rho}_{T(i)} & 0 \le x \le \hat{\rho}_{T(i)} \\ (1-\hat{\rho}_{T(i)}) \left(1 - \frac{1-x}{1-\hat{\rho}_{T(i)}}\right) + \hat{\rho}_{T(i)} \times 1 & \hat{\rho}_{T(i)} < x \le 1 \end{cases} \tag{80}$$

$$= x \text{ for } 0 \le x \le 1. \tag{81}$$

Therefore $W_i^{(PQ)} \stackrel{\text{d}}{=} \mathbf{I}_i^{(PQ)} U$ where U is a uniform random variable on $[0,1]$. For Poisson arrivals this delay is the same as that given in Corollary 7.1.

Now consider the corresponding gated case. A particle of i work in such a system will arrive during $\mathbf{I}_i^{(PQ)}$ with probability $\mathbf{I}_i^{(PQ)}/\mathbf{C}_i^{(PQ)}$ and during the preceding visit time $\mathbf{V}_{\sigma_{iT(i)}}$ with probability $\mathbf{V}_{\sigma_{iT(i)}}/\mathbf{C}_i^{(PQ)}$. If $u_1 \mathbf{I}_i^{(PQ)}$ time units of $\mathbf{I}_i^{(PQ)}$ have passed when it arrives $(0 \le u_1 \le 1)$ then it will find $\hat{\rho}_{T(i)}(u_1 \mathbf{I}_i^{(PQ)} + \mathbf{V}_{\sigma_{iT(i)}}^{(PQ)})$ work ahead of it and the delay will be $\mathbf{I}_i^{(PQ)}(1-u_1) + \rho_i(u_1 \mathbf{I}_i^{(PQ)} + \mathbf{V}_{\sigma_{iT(i)}})$. If, for $0 \le u_2 \le 1$, it arrives $u_2 \mathbf{V}_{\sigma_{iT(i)}}$ of the time into the preceding visit time then the delay will be $\hat{\rho}_{T(i)} u_2 \mathbf{V}_{\sigma_{iT(i)}} + (1-u_2)\mathbf{V}_{\sigma_{iT(i)}}) + \mathbf{I}_i^{(PQ)}$. Using an analysis similar to that in the exhaustive case, $W_i^{(PQ)} \stackrel{\text{d}}{=} (\hat{\rho}_{T(i)} + (1-\hat{\rho}_{T(i)})U)\mathbf{C}_i^{(PQ)}$, again corresponding to the form found in Corollary 7.1.

Putting the above together, a strong conjecture for the heavy-traffic behavior of a system with non-Poisson arrivals is that Theorems 5, 6, and 7, and Corollary 7.1 will continue to hold but with σ^2 replacing $b^{(2)}/b^{(1)}$. Further, transient results should be as above. However, given the intensive effort involved in order for CPR to prove the two-queue cyclic exhaustive case, we leave a rigorous proof of this result as an open problem.

7 Numerical Results and Approximations

Theorem 7 suggests the following approximation for the delay distributions in stable polling systems: For $j = 1, \ldots, N$, $\rho < 1$,

$$\text{Prob}\{W_j < x\} \approx \text{Prob}\left\{\widetilde{W}_j < x(1-\rho)\right\} \quad (x > 0). \tag{82}$$

Using Remark 5.2, the distribution of $\widetilde{W}_j$ is readily calculated on a spreadsheet, or via the Numerical Transform Inversion (NTI) based technique discussed in [7]. Similarly, equations (69)-(71) also suggest the following approximations for the moments of the delay: For $j \in G$, $k = 1, 2, \ldots$,

$$E[W_j^k(app)] = \frac{1}{(1-\rho)^k} \left\{ \frac{1}{k+1} \sum_{i:T(i)=j} \hat{\pi}_i^{k+1} \left(1 + \hat{\rho}_{T(i)} + \cdots + \hat{\rho}_{T(i)}^k\right) \prod_{m=1}^{k} \left(r + \frac{m}{2\delta} \frac{b^{(2)}}{b^{(1)}}\right) \right\}, \tag{83}$$

and for $j \in E$, $k = 1, 2, \ldots$,

$$E[W_j^k(app)] = \frac{1}{(1-\rho)^k} \left\{ \frac{1}{k+1} \sum_{i:T(i)=j} \hat{\pi}_i^{k+1} \left(1 - \hat{\rho}_{T(i)})^k\right) \prod_{m=1}^{k} \left(r + \frac{m}{2\delta} \frac{b^{(2)}}{b^{(1)}}\right) \right\}, \tag{84}$$

where δ is defined in (64). Notice that Theorem 7 and Corollary 7.2 imply that the approximations (82)-(84) are asymptotically exact for $\rho \uparrow 1$. This raises the natural question "How heavy should the traffic be to make the approximation (82)-(84) work well?" To this end, we have performed numerical experiments to investigate the accuracy of the approximations for different values of the system load. The results are outlined below.

We consider a system with the following parameters: $N = 3$; $M = 4$; $T = (1, 2, 1, 3)$; $G = \{1, 3\}$; $E = \{2\}$; the ratio between the arrival rates is $1 : 1 : 5$; the service times at Q_1 are deterministic with mean 3, the service times are Q_2 are exponentially distributed with mean 1 and the service times at Q_3 are gamma-distributed with mean 1, and squared coefficient of variation 5. The switch-over times from QP_2 to PQ_3 and from PQ_4 to PQ_1 are deterministic with means 0.25 and 0.2, respectively. The switch-over times from PQ_1 to PQ_2 are gamma-distributed with mean 0.5 and squared coefficient of variation 10, and the switch-over times from PQ_2 to PQ_3 are exponentially distributed with mean 0.05.

It is readily verified that for this model that the mean total switch-over time per cycle is 1, and that the first two moments of the aggregated service times are $b^{(1)} = 9/7$ and $b^{(2)} = 41/7$. Moreover, one may verify that for the present model $\hat{\rho}_1 = 1/3$, $\hat{\rho}_2 = 1/9$ and $\hat{\rho}_3 = 5/9$, and that $\hat{\lambda}_1 = 1/9$, $\hat{\lambda}_2 = 1/9$ and $\hat{\lambda}_3 = 5/9$. Similarly, from equations (11)-(12) it follows that $v_1 = 2/9, v_2 = 1/9, v_3 = 1/9, v_4 = 5/9$ and that the $\hat{\pi}$-values are given by $\hat{\pi}_1 = 2/3, \hat{\pi}_2 = 1, \hat{\pi}_3 = 1/3$ and $\hat{\pi}_4 = 1$. Finally, it is readily verified from (20)-(21) that $\hat{y}_1 = 6/81$, $\hat{y}_2 = 6/81$, $\hat{y}_3 = 0$ and $\hat{y}_4 = 25/81$, and hence $\delta = \sum_{i=1}^{4} b_{T(i)}\hat{y}_i = 49/81$.

For this model, we have computed the exact and approximated values of the k-th moments of $(1-\rho)W_1$, i.e., the scaled delay at Q_1, for different values of the load. Here, the exact results have been obtained by means of the numerical technique discussed in Choudhury and Whitt [7], and the approximations have been calculated from (83)-(84). Table 1 below shows the relative error of the approximations, defined by, for $i = 1, \ldots, N$,

$$err_j\% = \frac{E[W_j^k(app)] - E[W_j^k]}{E[W_j^k]} \times 100\%, \tag{85}$$

as a function of the load, for $k = 1, 2, \ldots, 5$.

k	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$	$\rho = 0.99$
1	3.69	2.20	1.67	1.03	0.58	0.13
2	-15.36	-6.69	-3.37	-1.20	-0.52	-0.10
3	-30.93	-13.99	-7.62	-3.22	-1.55	-0.31
4	-43.41	-20.58	-11.76	-5.29	-2.58	-0.51
5	-53.56	-26.84	-15.96	-7.42	-3.64	-0.72

Table 1: Relative error in approximation of moments.

From Table 1 it is clear that relative error decreases as ρ increases. As the approximation is asymptotically exact this is not surprising. What is more interesting is how well the approximations perform for ρ significantly less than 1. The approximation is highly accurate for estimating the mean delay even for ρ as low as 0.5. It becomes less accurate as the degree of the moment increases but is still quite accurate at moderately high loads. Recall that the approximation depends purely on the first two moments of the service distribution and only the first moments of the switch-over time distributions so this degree of accuracy is perhaps surprising. We also observe that for $k > 1$ the approximation consistently *under*estimates the exact value, obtained via the approach in [7]. This observation can be explained by the fact that the approximation based on the asymptotic (70)-(71) neglects the impact of the terms in $\Delta_{i,c}^{(k)}$ in (36), whereas these positive terms may be significant for $\rho < 1$.

Figure 1 presents the distribution of the delay for the same data set. For varying values of ρ, $\text{Prob}\,\{W_1 \leq x(1-\rho)\}$ is plotted versus varying values of x.

<< Figure 1 here >>

As is predicted by Theorem 7, the values plotted in Figure 1 are converging as ρ increases. Further they increase to the value predicted by Theorem 7 (labeled as "heavy-traffic" on the figure).

In order to test the hypothesis given in Section 6.3 for systems with non-Poisson arrivals, simulation must be used because, to the best of our knowledge, there are no exact results available for such systems. We therefore use the simulation code described in Olsen [27]. We test a four queue system with the following parameters: $N = 4$; $M = 6$; $T = (1, 2, 1, 3, 1, 4)$; $E = \{1, 2, 3, 4\}$; the ratio between the arrival rates is $2:3:4:1$. The service times times at all queues are exponential with means 2, 1, 3, and 1.5, respectively. All visits to queue 1 have the same switch-over time distribution, which is exponential with mean 0.5. Visits to queues 2, 3, and 4 also have exponentially distributed switch-over times with means 0.75, 1, and 0.75 respectively. Table 2 shows the relative error in approximating the 95th percentile of delay at queue 1 for varying values of ρ, total switch-over r, and arrival distribution. We define the 95th percentile as the number x such that Prob $\{W_1 < x\} = 0.95$. For the heavy-traffic approximation this value was found using a goal seeking algorithm on a spreadsheet. Interarrival times are taken to have an m-Erlang distribution with m =1, 2, and 4. Of course, for $m = 1$ arrivals are Poisson. This case is included as a benchmark. The mean switch-over times are as given above ($r = 4$), divided by two ($r = 2$), or multiplied by two ($r = 8$).

		$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$
	$r = 2$	-25.50	-7.86	-1.88	1.35	1.37
Erlang 1	$r = 4$	-24.74	-8.30	-2.85	0.74	0.97
	$r = 8$	-22.52	-7.87	-2.98	0.57	0.94
	$r = 2$	-25.02	-7.69	-1.88	1.20	0.88
Erlang 2	$r = 4$	-24.37	-8.31	-2.88	-0.06	-0.06
	$r = 8$	-20.75	-7.16	-2.19	0.58	0.89
	$r = 2$	-25.95	-8.53	-2.86	0.04	0.38
Erlang 4	$r = 4$	-24.01	-8.26	-2.77	0.04	0.61
	$r = 8$	-20.81	-8.00	-3.40	-0.67	-0.11

Table 2: Relative error in 95th percentile for different interarrival time distributions and load values.

It can be seen that the approximation is again very accurate for moderate to high loads. All approximations for $\rho = 0.9$ and higher lay within the confidence interval for the simulation but none of those for $\rho = 0.8$ and below did. The difference in delay between 1-Erlang and 4-Erlang was on average 39% so the small errors are not a function of small changes in the delay. Again, as would be expected, the errors generally decrease as ρ increases. The exceptions are possibly more due to simulation errors than real differences. There are few patterns in the errors across the different arrival distributions and the different mean switch-overs.

In summary, we have demonstrated the accuracy of the moment and distributional approximation suggested in this paper. We have provided evidence for the validity of our conjecture for non-Poisson arrivals. Further, this numerical analysis provides validation of our theoretical results. Here, we have sought to give a flavor of the behavior of the approximations; clearly, a more extensive numerical study is possible. It may also be possible to refine the approximations in this paper using non-heavy-traffic results (e.g., using the approach take in Olsen [26]). However, the primary goal of this paper has been a formal presentation of new limit theorems. We leave such extentions as the subject for future work.

8 Topics for Future Research

Finally, we address a number of topics for further research. First, the exact, easy-to-evaluate, expressions presented here depend only on a limited number of system parameters. Therefore obtaining "best" system performance should be a much easier task in the limiting regime than in non-scaled systems. Optimization of system performance is left as a task for future research. An easier first step however is to consider feasibility problems of the form "Can the system be operated such that Prob$\{W_j > x_j\} < \alpha_j$ $(j = 1, \ldots, N)$?" Such questions are of much practical importance. Here, the values of x_j and α_j are typically specified by the users, and typical decision variables are the choice of the service policies and the visit order. The results in this paper open possibilities for obtaining (approximate) solutions for solving feasibility problems under heavy-traffic assumptions. Second, in the model considered here, it is assumed that the service disciplines at the queues are exhaustive or gated. However, the results should be able to be extended to more general service disciplines satisfying the "branching" structure defined Resing [33], as demonstrated for the case of cyclic polling in Van der Mei and Levy [46]. Extending our work to such more general systems is left as the subject of future research. Finally, although we have provided a convincing hypothesis for the behaviour of the system under non-Poisson arrival processes, we have provided no proof for this result. Such a proof is also left as the subject of future research.

Acknowledgment

The authors wish to thank Gagan Choudhury for his help in obtaining the numerical results presented in Section 7.

References

[1] Athreya, K.B and Ney, P.E. (1971). *Branching Processes* (Springer, Berlin).

[2] F. Baccelli and P. Brémaud. *Elements of Queueing Theory.* Springer Verlag, Berlin, 1991.

[3] Baker, J.E. and Rubin, I.R. (1987). Polling with a general-service order table. *IEEE Trans. Commun.* **35**, 283-288.

[4] Blanc, J.P.C. (1993). Performance analysis and optimization with the power-series algorithm. In: *Performance Evaluation of Computer and Communication Systems*, eds. L. Donatiello and R. Nelson (Springer-Verlag, Berlin), 53-80.

[5] Boxma, O.J., Groenendijk, W.P. and Weststrate, J.A. (1990). A pseudoconservation law for service systems with a polling table. *IEEE Trans. Commun.* **38**, 1865-1870.

[6] Choudhury, G.L. (1990) Polling with a general service order table: gated service. In: *Proc. INFOCOM '90*, 268-276.

[7] Choudhury, G. and Whitt, W. (1994). Computing transient and steady state distributions in polling models by numerical transform inversion. *Perf. Eval.* **25**, 267-292.

[8] Chung, K.L. (1974). *A Course in Probability* (Academic Press, 2nd ed.).

[9] Coffman, E.G., Puhalskii, A.A. and Reiman, M.I. (1995). Polling systems with zero switch-over times: a heavy-traffic principle. *Ann. Appl. Prob.* **5**, 681-719.

[10] Coffman, E.G., Puhalskii, A.A. and Reiman, M.I. (1998). Polling systems in heavy-traffic: a Bessel process limit. *Math. Oper. Res.* **23**, 257-304.

[11] Cohen, J.W. and Boxma, O.J. (1981). The M/G/1 queue with alternating service formulated as a Rieman-Hilbert boundary value problem. In: *Performance '81*, ed. F.J. Kylstra (North Holland, Amsterdam), 181-199.

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

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

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

[15] Federgruen, A. and Katalan, Z. (1998). Determining Production Schedules Under Base-Stock Policies in Single Facility Multi-Item Production Systems. *Oper. Res.* **46**, 883–898.

[16] Fuhrmann, S.W. and Cooper, R.D. (1985). Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Oper. Res.* **33**, 1117-1129.

[17] Fricker, C. and Jaïbi, M.R. (1994). Monotonicity and stability of periodic polling models. *Queueing Systems* **15**, 211-238.

[18] Harrison, J.M., and Nguyen, V. (1993). Brownian models of multiclass queueing networks: Current status and open problems. *Queueing Systems* **13**, 5-40.

[19] Jennings, O.B. (1999). *Generalized Round Robin Service Disciplines in Stochastic Networks with Setups: Stability Analysis and Diffusion Approximation.* Ph.D. Thesis, Georgia Institute of Technology, Atlanta GA, USA.

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

[21] Kroese, D.P. (1997). Heavy traffic analysis for continuous polling models. *J. Appl. Prob.* **34**, 720-732.

[22] Leung, K.K. (1991). Cyclic service systems with probabilistically-limited service. *IEEE J. Sel. Areas Commun.* **9**, 185-193.

[23] Levy, H. and Sidi, M. (1991). Polling models: applications, modeling and optimization. *IEEE Trans. Commun.* **38**, 1750–1760.

[24] Markowitz, D.M., Reiman, M.I., and Wein, L.M. (2000). The stochastic economic lot scheduling problem: heavy traffic analysis of dynamic cyclic policies. *Oper. Res.* **48**, 136-154.

[25] Markowitz, D.M., and Wein, L.M. (2001). Heavy traffic analysis of dynamic cyclic policies: A unified treatment of the single machine scheduling problem. *Oper. Res.* **49**, 246-270.

[26] T. L. Olsen (2001). Approximations for the waiting time distribution in polling models with and without state-dependent setups. *Oper. Res. Lett.* **28**, 113-123.

[27] T. L. Olsen (1999). A practical scheduling method for multi-class production systems with setups. *Mgmt. Sc.* **45**, 116-130.

[28] T. L. Olsen (2001). Limit Theorems for Polling Models with Increasing Setups. *Prob. Eng. Inf. Sc.* **15**, 35-55.

[29] T. L. Olsen (2001). On Multi-Item Production Systems with Setups: Review and Intuition. Working Paper, Washington University in St. Louis, MO.

[30] Reiman, M.I., and Wein, L.M. (1998). Dynamic scheduling of a two-class queue with setups. *Oper. Res.* **46**, 532-547.

[31] Reiman, M.I., and Wein, L.M. (1999). Heavy traffic analysis of polling systems in tandem. *Oper. Res.* **47**, 524-534.

[32] Reiman, M.I., Rubio, R., and Wein, L.M. (1999). Heavy traffic analysis of the dynamic stochastic inventory-routing problem. *Transp. Sc.* **33**, 361-380.

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

[34] Takagi, H. (1986). *Analysis of Polling Systems* (The MIT Press, Cambridge, MA).

[35] Takagi, H. (1990). Queueing analysis of polling models: an update. In: *Stochastic Analysis of Computer and Communication Systems*, ed. H. Takagi (North-Holland, Amsterdam), 267-318.

[36] Takagi, H. (1991). Applications of polling models to computer networks. *Comp. Netw. ISDN Syst.* **22**, 193-211.

[37] Takagi, H. (1997). Queueing analysis of polling models: progress in 1990-1994. In: *Frontiers in Queueing: Models and Applications in Science and Engineering*, Chapter 5, J. H. Dshalalow (ed.), CRC Press, New York.

[38] Takagi, H. and M. Murata (1986). Queueing analysis of scan-type TDM and polling systems. In: *Computer Networking and Performance Evaluation*, eds. T. Hasegawa, H. Takagi and Y. Takahashi (North Holland, Amsterdam), 199-211.

[39] Van der Mei, R.D. (1999). Delay in polling systems with large switch-over times. *J. Appl. Prob.* **36**, 232-243.

[40] Van der Mei, R.D. (1999). Distribution of the delay in polling systems in heavy traffic. *Perf. Eval.* **38**, 133-148.

[41] Van der Mei, R.D. (1999). Polling systems in heavy traffic: higher moments of the delay. *Queueing Systems* **31**, 265-294.

[42] Van der Mei, R.D. (1999). Polling systems with periodic server routing in heavy traffic. *Stoch. Mod.* **15**, 273-297.

[43] Van der Mei, R.D. (2000). Polling systems with switch-over times under heavy load: Moments of the delay. *Queueing Systems* **36**, 381-404.

[44] Van der Mei, R.D. and Borst, S.C. (1997). Analysis of multiple-server polling systems by means of the power-series algorithm. *Stoch. Mod.* **13**, 339-369.

[45] Van der Mei, R.D. and Levy, H. (1998). Expected delay analysis of polling systems in heavy traffic. *Adv. Appl. Prob.* **30**, 586-602.

[46] Van der Mei, R.D. and Levy, H. (1997). Polling systems in heavy traffic: Exhaustiveness of service policies. *Queueing Systems* **27**, 227-250.

[47] Weststrate, J.A. and Van der Mei, R.D. (1994). Waiting times in a two-queue model with exhaustive and Bernoulli service. *Z.O.R. Math. Meth. Oper. Res.* **40**, 289-303.

Figure 1: Cumulative distribution of $(1-\rho)W_1$ for different values of the load.