Polling Systems with Periodic Server Routing in Heavy Traffic: Renewal Arrivals

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
Email: olsen@olin.wustl.edu (corresponding author)
[b]CWI, Advanced Communication Networks, Amsterdam, Netherlands
[c]Vrije Universiteit, Faculty of Exact Sciences, Amsterdam, Netherlands
Email: mei@cwi.nl

This paper considers heavy-traffic limit theorems for polling models with periodic server routing with general renewal arrivals and mixtures of gated and exhaustive service policies. We provide a strong conjecture for the limiting waiting-time distribution in a general parameter setting when the load tends to 1, under proper heavy-traffic scalings.

Keywords: polling table, delay, heavy-traffic, distributional approximations, switch-over times, setup times

*The work of this author was work was supported in part by NSF CAREER grant DMI-9875202.

1 Introduction

This paper considers heavy-traffic limit theorems for polling models with periodic server routing under renewal arrivals. As such, we generalize results in Olsen and Van der Mei [14], which were proven for these systems under Poisson arrivals in steady-state. In such systems 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 (referred to as exhaustive service) or until all customers that were in the current class when the server completed switching over to that class are served (referred to as gated service). Such polling models have been widely studied and are applicable to a wide range of computing, telecommunications, and manufacturing environments. We refer to [10, 19] for overviews on the applicability of polling models.

The approach taken in this paper is fundamentally different to that in [14]. In that paper, non-heavy-traffic steady-state results were harnessed and significant simplifications shown to result as utilization was taken to 100%. Limit theorems were rigorously proven for the case with Poisson arrivals. However, such techniques rely on the existence of non-heavy-traffic results, which, in turn, rely heavily on the assumption of Poisson arrivals. In this paper, instead, we follow the approach of [6] (for the N queue case), [16], [9], [11], [17], and [15]. In this line of work, heavy traffic results are not rigorously proven, instead they are conjectured to hold based on results from the rigorously analyzed two-queue system in [6]. In particular, we extend the N queue conjecture of Coffman et al. [6] (CPR) to the periodic gated and exhaustive service models of this paper. Like these authors, we provide a strong conjecture for the system's behavior under heavy-traffic. As such, we present what we believe to be the first heavy-traffic performance approximation for polling models with periodic server routing under non-Poisson arrivals.

Our results will be validated in a number of ways. First, by utilizing the techniques of the aforementioned papers we are following a well accepted path. Second, our results will correspond exactly to the rigorously proven cases of two queues under exhaustive service and the general steady-state system under Poisson arrivals. Last, we perform numerical experiments to test the steady-state

approximations yielded by our limit theorems.

This paper is organized as follows. In Section 2 the model is described, notation introduced, and prior results for systems with Poisson arrivals outlined. In Section 3 we present the main results of the paper, extending the work of CPR to general polling systems. In Section 4 we present the results of the numerical experiments with simulations performed to validate our results and to test the accuracy of the approximations for waiting-time distributions in stable polling systems, suggested by our results. Finally, in Section 5 we address a number of topics for further research.

2 Model and Prior Steady-State Analysis

This section outlines our model and the results obtained previously in [14]. We consider a system consisting of N infinite-buffer queues, $Q_1, \ldots, Q_N$. Customers arrive at Q_j according to a general renewal 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 k-th moment $b_j^{(k)}$, $j = 1, \ldots, N$, $k = 1, 2, \ldots$ All moments are assumed to be finite in [14]; for our conjectures we only require finite first and second moments. 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., [7]). For the rest of this section, we assume that this condition is satisfied and that the system may be considered in "steady-state".

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 [2] and [14], 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 $T(i)$. 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 finite second moment and mean, r_i, $i = 1, \ldots, M$. Note that [14] assumes that all moments of R_i are finite. 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 σ_{ij} be the polling table 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. 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 $C_i^{(PQ)} = I_i^{(PQ)} + V_{\sigma_{iT(i)}}$), for $i = 1, \ldots, M$.

Denote by W_j the delay incurred by an arbitrary customer at Q_j, $j = 1, \ldots, N$. 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. Throughout, W_j and $W_i^{(PQ)}$ 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. 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 $X_n \rightarrow_d X$, if there exists a dense subset A of $\mathcal{R}$ (i.e., of real numbers) such that $\lim_{n\to\infty}$

$P(X_n < a) = P(X < a)$, for all $a \in A$.

Let $\pi_i > 0$ be the fraction of customers at $Q_{T(i)}$ that is served at PQ_i, $i = 1, \ldots, M$. Define the heavy-traffic residues by, for $i = 1, \ldots, M$,

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

The v_i $(i = 1, \ldots, M)$, are uniquely determined by solving the following set of linear equations (see, e.g., [14]): For $i = 1, \ldots, M$, $j = 1, \ldots, N$,

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

where

$$\varphi_j := b_j^{(1)} \quad (j \in G), \quad \text{and} \quad \varphi_j := \frac{b_j^{(1)}}{1-\rho_j} \quad (j \in E)$$

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

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

The (scaled) delay distribution when the load tends to unity can be expressed explicitly in terms of $\hat{\pi}_i$ $(i = 1, \ldots, M)$. In this context, it is convenient to define:

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

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

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

We will now formulate three properties of the heavy traffic behavior of the system that were shown to hold for the case of Poisson arrivals (see Theorems 5 and 6, and Corollary 7.1 in [14], respectively). In the next section these properties will be extended to the case of general renewal arrival processes.

Property 1

If $T(i) \in E$ and arrivals are Poisson, then

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

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

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

Property 2

If $T(i) \in G$ *and arrivals are Poisson, then*

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

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

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

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, e.g., [1]) as a random variable with p.d.f. $f_{\mathbf{X}}(x) = xf_X(x)/EX$. See [1] 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 Properties 1 and 2 for the exhaustive and gated cases, respectively. These properties yield the following elegant form for the delay distribution.

Property 3

Assume arrivals follow a Poisson process. For $T(i) \in G$,

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

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

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

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

3 Analysis

In this section we analyze the transient system, no longer assuming that the system is in steady-state. Further arrivals are no longer assumed to be Poisson but instead follow a renewal process. In fact, like other heavy-traffic work, it is not actually necessary that arrivals follow a renewal process, rather that they obey a functional central limit theorem that satisfies some technical conditions (see, e.g., CPR [6]). However, renewal processes appear to provide sufficient generality for most applications.

The results of CPR differ from traditional heavy-traffic work (see, e.g., [8]) 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 state-space collapse. 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 the polling model in [5] 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

$$d(x) = a + r(x) \tag{1}$$

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

$$\frac{E[\text{time spent in switch-overs}]}{E[\text{duration of a cycle}]} \tag{2}$$

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, state-space collapse implies that 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 (switch-overs become negligible under this scaling).

We will 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 (1) 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). As mentioned in the introduction, a similar conjecture was made for other general polling models in [16], [9], [11], [17], and [15]. It therefore remains to derive the appropriate parameters for the model.

Conjecture 1

Define $U^n(t)$ as unfinished work in the n^{th} system (see, e.g., CPR). Define $X^n(t) = n^{-1/2}U^n(nt)$. Then

$$X^n \rightarrow_d X \quad (n \rightarrow \infty),$$

where X is a diffusion on $[0,\infty)$ with drift $d(x) = a + r\delta/x$, variance σ^2, and with instantaneous reflection at the origin.

Argument: By the averaging principle, total workload in the system may be regarded as un-

changed 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 (1) and (2) we conjecture that $d(x) = a + r\delta/x$. □

Note that when service is cyclic and exhaustive Conjecture 1 corresponds to the limit found by CPR (where δ in our notation corresponds to ϱ in their's). For $a < 0$ this process has a stationary distribution corresponding to a gamma random variable with $\alpha = 2r\delta/\sigma^2 + 1$ and $\mu = 2|a|/\sigma^2$.

We can use the above analysis to find expressions for the cycle and intervisit times in this system.

Conjecture 2

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

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

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

$$\alpha := 2r\delta/\sigma^2 + 1 \text{ and } \mu_i := 2\delta/((1-\hat{\rho}_i)\hat{\pi}_i\sigma^2).$$

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

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

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

$$\alpha := 2r\delta/\sigma^2 + 1 \text{ and } \mu_i := \mu = 2\delta/(\hat{\pi}_i\sigma^2).$$

Argument: As before, when there is a scaled amount x work in the scaled system cycle time equals x/δ. Therefore, applying the averaging principle and the same reasoning as above, 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 and taking the system to steady-state (an interchange of limits that would require justification in any rigorous proof). □

When arrivals are Poisson the limits in Conjecture 2 correspond *exactly* to those in Properties 1 and 2. It therefore remains to find the delay for the individual queues.

Conjecture 3

For $i = 1, \ldots, M$,

$$(1-\rho)W_i^{(PQ)} \rightarrow_d \tilde{W}_i^{(PQ)} \quad (\rho \uparrow 1),$$

where for $T(i) \in G$,

$$\tilde{W}_i^{(PQ)} =_d (\hat{\rho}_{T(i)} + (1-\hat{\rho}_{T(i)})U)\tilde{\mathbf{C}}_i^{(PQ)}, \tag{3}$$

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

$$\tilde{W}_i^{(PQ)} =_d U\tilde{\mathbf{I}}_i^{(PQ)}, \tag{4}$$

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

Argument: 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 variables.) 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 ($0 \leq u_1 \leq 1$) 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 $0 \leq u_2 \leq 1$, 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

$$\begin{aligned} P(W_i^{(PQ)} \leq t) &= P(1-(1-\hat{\rho}_{T(i)})U_1 \leq t/\mathbf{I}_i^{(PQ)})(1-\hat{\rho}_{T(i)}) \\ &\quad + P(\hat{\rho}_{T(i)}(1-U_2) \leq t/\mathbf{I}_i^{(PQ)})\hat{\rho}_{T(i)} \end{aligned}$$

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

$$\begin{aligned} P(W_i^{(PQ)} \leq \mathbf{I}_i^{(PQ)}x) &= P(1-(1-\hat{\rho}_{T(i)})U_1 \leq x)(1-\hat{\rho}_{T(i)}) \\ &\quad + P(\hat{\rho}_{T(i)}(1-U_2) \leq x)\hat{\rho}_{T(i)} \\ &= \begin{cases} (1-\hat{\rho}_{T(i)}) \times 0 + \hat{\rho}_{T(i)}x/\hat{\rho}_{T(i)} & 0 \leq x \leq \hat{\rho}_{T(i)} \\ (1-\hat{\rho}_{T(i)})\left(1-\frac{1-x}{1-\hat{\rho}_{T(i)}}\right) + \hat{\rho}_{T(i)} \times 1 & \hat{\rho}_{T(i)} < x \leq 1 \end{cases} \\ &= x \text{ for } 0 \leq x \leq 1. \end{aligned}$$

Therefore $W_i^{(PQ)} \stackrel{\mathrm{d}}{=} \mathbf{I}_i^{(PQ)}U$ where U is a uniform random variable on $[0,1]$.

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 ($0 \leq u_1 \leq 1$) 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 $0 \leq u_2 \leq 1$, 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, $W_i^{(PQ)} \stackrel{\mathrm{d}}{=} (\hat{\rho}_{T(i)} + (1-\hat{\rho}_{T(i)})U)\mathbf{C}_i^{(PQ)}$. □

We are now ready to present the main result.

Conjecture 4

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

$$(1-\rho)W_j \rightarrow_d \sum_{i:T(i)=j} \hat{\pi}_i \tilde{W}_i^{(PQ)} \quad (\rho \uparrow 1),$$

where the distribution of $\tilde{W}_i^{(PQ)}$ *is defined in* (3) *and* (4).

Argument: The result follows directly from Conjecture 3, by conditioning on the pseudo-queue at which a customer arriving at queue j is served. □

For Poisson arrivals Conjecture 3 gives limits for steady-state delay (at pseudo-queues) that are the same as those given in Property 3. In summary, a strong conjecture for the heavy-traffic behavior of a system with renewal arrivals is that Properties 1, 2, and 3 will continue to hold but with σ^2 replacing $b^{(2)}/b^{(1)}$, finally leading to Conjecture 4. Further, transient results should be as in Conjecture 1. However, given the intensive effort involved in order for CPR to prove the two-queue cyclic exhaustive case, we, like others before us, leave a rigorous proof of this result as an open problem.

4 Numerical Results

In order to test the hypothesis for systems with renewal arrivals, we have performed a variety of simulation experiments, based on the simulation code described in [13]. The results are outlined below.

We first consider a system with the following parameters: $N = 3$; $M = 4$; $T = (1, 2, 1, 3)$; $G = \{1, 2, 3\}$; the ratio between the arrival rates is $1 : 1 : 5$; the service times are exponential with means 3, 1, and 1 at queues 1, 2, and 3, respectively. The switch-over times are also exponentially distributed with means 1, 0.5, 2, and 0.5, from PQ_1 to PQ_2, from PQ_2 to PQ_3, from PQ_3 to PQ_4, and from PQ_4 to PQ_1, respectively. Using the distribution of delay derived in the previous section as an approximation, we test the accuracy of this approximation versus the simulated value of the delay at queue 1. We define the pth percentile as the number x such that $\text{Prob}\{W_1 < x\} = p$. For the heavy-traffic approximation this value was found using a goal seeking

algorithm on a spreadsheet.

		mean	50th perc.	60th perc.	70th perc.	80th perc.	90th perc.	95th perc.
	ρ=0.85	0.19	1.59	-1.19	-2.93	-3.89	-4.13	-4.16
Exp.	ρ=0.9	0.07	0.29	-1.38	-2.32	-2.69	-2.68	-2.82
	ρ=0.95	-0.17	-0.46	-1.26	-1.64	-1.58	-1.46	-1.57
	ρ=0.85	5.46	8.54	5.00	2.29	0.49	-1.08	-2.14
Erl. 2	ρ=0.9	3.57	4.70	2.56	0.98	0.06	-0.74	-1.34
	ρ=0.95	1.59	1.73	0.81	0.10	-0.20	-0.47	-0.84
	ρ=0.85	12.18	17.49	12.88	8.99	5.49	2.78	0.57
Det.	ρ=0.9	7.92	10.51	7.27	5.01	3.21	1.65	0.12
	ρ=0.95	3.40	4.08	2.72	1.74	0.87	0.32	-0.49

Table 1: Relative error in mean and percentiles for different interarrival time distributions and load values.

Table 1 shows the percentage error between the simulated values and the approximation, where percentage error is defined as

$$100\frac{E-S}{S},$$

where E stands for the estimated value and S stands for the simulated value. Errors in approximating both the mean of delay and its percentiles are reported (only point estimates are presented here, confidence intervals are omitted for compactness of the presentation). Exponential, Erlang 2, and deterministic interarrival times are tested. In all cases the approximations become more accurate as ρ increases.

We next 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 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. 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 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 inaccuracies than real differences. There are few patterns in the errors across the different arrival distributions and the different mean switch-overs.

The numerical examples described in Tables 1 and 2 provide empirical evidence for the validity of our conjecture for renewal arrivals. 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 taken in [12]). In particular, evaluating the parameters at $\rho < 1$ rather than at $\rho = 1$ provided significantly more accurate estimates for the mean delay in Table 1. However, the primary goal of this paper has been a presentation of new limit theorems, and refinement of the approximations presented in Table 1 and 2 is beyond the scope of the present paper.

5 Topics for Further Research

First, in this paper we have provided a strong hypothesis for the behavior of polling models with periodic server routing and renewal arrivals under heavy-traffic. A rigorous proof is left as the subject of future research.

Second, the exact and easy-to-evaluate expressions as presented here open possibilities for obtaining approximate solutions for solving system design problems under heavy-traffic assumptions. A typical practically relevant problem is the following "Can we construct a polling table such that $\text{Prob}\{W_j > x_j\} < \alpha_j$ $(j = 1, \ldots, N)$?", for given values of x_j and α_j $(j = 1, \ldots, N)$. The results presented in this paper may be highly useful to address this type of feasibility problem.

Third, it is assumed that the first two moments of the service times and interarrival times, and the first moment of the switch-over times are finite. A challenging area for further research is to analyze the impact of heavy-tailed (say, with infinite variance) interarrival time and service-time distributions on the distributions of the delay in heavy traffic. In this context, interesting and promising results have been obtained in [4], which studies the tail behavior of the waiting times in polling systems with so-called regularly varying service times and switch-over times, and in [3], which derives the heavy-traffic limiting distribution for the waiting times in the single-server queue with a class of heavy-tailed service-time distributions.

Finally, in the model considered here, it is assumed that the service disciplines at the queues are exhaustive or gated. However, the results should be extendable to more general service disciplines satisfying the "branching" structure studied in [18], as demonstrated for the case of cyclic polling in [20]. Extending our work to such more general systems is left as the subject of future research.

References

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

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

[3] Boxma, O.J. and Cohen, J.W. (2000). The single server queue: heavy tails and heavy traffic. In: *Self-Similar Network Traffic and Performance Evaluation*, eds. K. Park and W. Willinger (Wiley, New York), 143-169.

[4] Boxma, O.J., Deng, Q. and Resing, J.A.C. (2000). Polling systems with regularly varying service and/or switch-over times. *Adv. Perf. Anal.* **3**, 71-107.

[5] 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.

[6] 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.

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

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

[9] 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.

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

[11] 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.

[12] 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.

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

[14] T. L. Olsen and van der Mei, R. D. (2002). Polling systems with periodic server routing in heavy traffic: distribution of the delay. *J. of Appl. Prob.* **40**, 305-326.

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

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

[17] 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.

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

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

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