Performance Bounds and Pathwise Stability for Generalized Vacation and Polling Systems

Eitan Altman; Serguei Foss; Eric Riehl; Shaler Stidham, Jr.

Operations Research, Vol. 46, No. 1. (Jan. - Feb., 1998), pp. 137-148.

PERFORMANCE BOUNDS AND PATHWISE STABILITY FOR GENERALIZED VACATION AND POLLING SYSTEMS

EITAN ALTMAN

INRIA, Centre Sophia Antipolis, Sophia Antipolis Cedex, France

SERGUEI FOSS

Institute of Mathematics, Novosibirsk, Russia

ERIC RIEHL and SHALER STIDHAM, JR.

University of North Carolina, Chapel Hill, North Carolina

(Received July 1993; revision received March 1994; accepted June 1996)

We consider a generalized vacation or polling system, modeled as an input-output process operating over successive "cycles," in which the service mechanism can be in an "up" mode (processing) or "down" mode (e.g., vacation, walking). Our primary motivation is polling systems, in which there are several queues and the server moves cyclically between them providing some service in each. Our basic assumption is that the amount of work that leaves the system in a "cycle" is no less than the amount present at the beginning of the cycle. This includes the standard gated and exhaustive policies for polling systems in which a cycle begins whenever the server arrives at some prespecified queue. The input and output processes satisfy model-dependent conditions: pathwise bounds on the average rate and the burstiness (Cruz bounds); existence of long-run average rates; a pathwise generalized Law of the Iterated Logarithm; or exponentially or polynomially bounded tail probabilities of burstiness. In each model we show that these properties are inherited by performance measures such as the workload and output processes, and that the system is stable (in a model-dependent sense) if the input rate is smaller than the up-mode processing rate.

We consider a general model for an input-output process, which includes many vacation and polling systems and other queueing systems as special cases. At any point in time, the system is either in the "up" mode, with both the arrival (input) and departure (output) process operating, or the "down" mode, in which the arrival process continues to operate, but the departure process may not be operating. Downtime may correspond to vacation time or switchover ("walking") time in a vacation or polling system. The time that the system is empty will also typically be understood to be part of the system downtime. The downtimes may even correspond to operation of the server at some reduced capacity, instead of complete idleness. In that case, the sufficient conditions for stability and the performance bounds that we obtain may turn out to be conservative.

A distinguishing feature of our model is that the time axis is divided into a sequence of disjoint time intervals, or "cycles," and we make certain assumptions about the behavior of the arrival and departure processes and the downtime during each cycle. One example of this type of setup is a cyclic polling system in which a cycle is defined in the "natural" way. That is, a cycle begins each time the server begins serving the first queue. Another is a vacation model in which a cycle begins each time a vacation ends. In addition to vacation and polling systems, our model has other potential applications, including communication systems with time-division multiplexing and flexible manufacturing processes with setups and/or breakdowns.

In general, our model allows considerable flexibility in the definition of cycles. Our basic assumption throughout the paper is that the total output during each cycle is at least as great as the quantity in the system at the beginning of the cycle. Examples satisfying this Output Assumption are polling systems with exhaustive, globally gated, or locally gated service disciplines, with cycles defined as above in the "natural" way. In contrast to most of the literature on polling systems, we allow for the discipline to change over time. Another example is a vacation system in which the server goes on vacation every time the queue becomes empty. In each of our models we also assume that the downtime per cycle satisfies certain constraints or growth conditions, and that the traffic intensity (suitably defined in each model) is less than one. It is the interaction of these conditions with the Output Assumption that essentially constrains the definition of cycles and motivates our derivations of performance bounds and stability.

We examine various types of performance bounds and stability for three special cases of our general model, based on different pathwise characterizations of their arrival and (potential) departure processes. We present the basic model and some general properties in Section 1. The three special cases are discussed in the next three sections.

Subject classifications: Queues: polling and vacation systems, pathwise analysis, performance bounds. Communications: token-ring local area networks.
Area of review: Stochastic Models.

0030-364X/98/4601-0137 $05.00

The first system (Section 2) has bounded downtimes in each cycle, and arrival and (potential) departure processes that satisfy burstiness constraints similar to those of Cruz (1991, 1992). We obtain uniform upper bounds for several performance measures, such as the amount of work in the system and the cycle duration. We also show that the output process satisfies a Cruz-type bound. The second system (Section 3) has arrival and potential departure processes with limiting average rates. We show that the system is rate stable (i.e., the quantity in the system is $o(t)$ as t approaches infinity), and we give explicit expressions for the limiting average cycle length and fraction of time the system is down. In Section 4 we look at systems whose arrival processes satisfy a pathwise analogue of the stochastic law of the iterated logarithm. We show how all three cases can be analyzed by a unified sample-path approach.

Finally (Section 5) we show how our pathwise results can be applied to stochastic models. In particular, we show that the workload and output processes inherit the property of exponentially or polynomially bounded burstiness from the input process. This implies that upper bounds, which are uniform in time, can be computed for the total workload in the system at any moment, for a wide class of stochastic arrival processes, not necessarily stationary nor ergodic. This generalizes many previous stability results for similar stochastic models.

In polling models that have been studied in the literature, it has usually been assumed that the queues are fed by independent Poisson arrival processes. Conditions for ergodicity as a measure of stability were obtained in Altman et al. (1992), Altman and Spieksma (1995), Borovkov and Schassberger (1994), Fricker and Jaibi (1994), Georgiadis and Szpankowski (1992), Resing (1993), and Zhdanov and Saksonov (1979). Kroese and Schmidt (1992, 1994) studied the stability of polling on a graph (i.e., a "continuous" polling model), and Altman and Levy (1994) studied the stability of noncyclic polling in two-dimensional and higher-dimensional planes. Altman et al. (1992) and Altman and Spieksma further present sufficient and necessary conditions for stronger notions of stability, namely geometric ergodicity and geometric rate of convergence of the moments of several performance measures (embedded at polling instants). Sufficient conditions for Central Limit Theorems and the Law of Iterated Logarithm are given in Altman and Spieksma. Altman and Liu (1994) analyse the stability of the *FDDI* protocol. Conditions for the stability of token rings with spatial reuse were obtained by Georgiadis et al. (1993).

All the above references assumed Poisson arrivals and independent walking and service times. This assumption is unrealistic, however, when dealing with many applications, e.g., Local Area Networks using token-ring protocols. The arrival processes there may be quite irregular, highly bursty, and correlated. Recently, Altman and Foss (1992) obtained sufficient stability conditions for polling systems with a general renewal arrival processes, and Massoulié (1993) reported some results on the construction of a stationary regime for general stationary ergodic arrival and service processes.

The analysis in this paper, which is based on pathwise bounds and limits, allows for more general arrival processes and makes it possible either to obtain strict upper bounds on several performance measures, such as waiting times, queue lengths, and workloads, or to derive pathwise stability conditions using a unified approach (based on the lemmas in Section 1). In Altman et al. (1994), we discuss some special cases of the models presented in this paper, with a focus on their applications to token-ring communication networks.

1. PRELIMINARY LEMMAS

Following Stidham and El-Taha (1993) (see also El-Taha and Stidham 1993 and Borovkov 1984, Chapter 2) we consider a nonnegative, real-valued deterministic process, $Z = \{Z(t), t \geqslant 0\}$—an *input-output* process—in which $Z(t) \geqslant 0$ represents quantity in a system. Specifically, we assume that the state space of Z is $S = R^+$, that $\{Z(t), t \geqslant 0\}$ is right continuous with left-hand limits, and that

$$Z(t) = Z(0) + A(t) - D(t), \quad t \geqslant 0, \tag{1}$$

where $A(t)$ $(D(t))$ is the cumulative input to (output from) the system in $[0, t]$, and both $\{A(t), t \geqslant 0\}$ and $\{D(t), t \geqslant 0\}$ are nondecreasing, right-continuous processes. Thus, $Z(t)$ has bounded variation on finite t-intervals. Note that $D(t) \leqslant Z(0) + A(t)$, since $Z(t) \geqslant 0$. We shall refer to $Z(t)$ generically as the *work* in the system at time t, with the understanding that it could be some other measure of quantity (e.g., the number of customers in a queue).

Let $u(t)$ $(v(t))$ be the indicator function for "up" ("down") time. That is, $u(t) = 1$ if the system is in the "up" mode, and $u(t) = 0$ if it is "down" at time t; $v(t) = 1 - u(t)$. We assume that $u(\cdot)$ (and hence also $v(\cdot)$) is integrable over finite t-intervals. In a cyclic polling system, for example, if downtime corresponds to the time spent by the server "walking" between queues, then $\int_0^t v(s)\, ds$ equals the total walking time in $[0, t]$, and $\int_0^t u(s)\, ds$ equals the total time in $[0, t)$ that the server processes work in the queues.

In general we shall use α and δ to denote the arrival rate and the departure rate, respectively, in some sense, the exact meaning of which will depend on the context. Define $\rho = \alpha/\delta$. We define the "burstiness" of each process during a time interval, $[s, t)$ by:

$$B^A_{s,t} := A(t) - A(s) - \alpha(t - s), \tag{2}$$

$$B^D_{s,t} := D(t) - D(s) - \delta \int_s^t u(\tau)\, d\tau. \tag{3}$$

Thus, $B^A_{s,t}$ is the difference between the actual input and the input that would have occurred if the input process had operated at its "average" rate throughout the interval. Similarly, $B^D_{s,t}$ is the difference between the actual output and the output that would have occurred if the output process had operated at its "average" rate throughout the uptime in the interval (and at rate 0 during the downtime).

We assume that there is a sequence of time points, $\{t_n, n \geq 0\}$, that defines *cycles* for the system, with $0 = t_0 \leq t_1 \leq t_2 \leq \cdots$. We interpret t_n to be the time point at which the nth cycle ends for $n \geq 1$. We define quantities corresponding to each cycle. Let $T_n := t_n - t_{n-1}$, the duration of the nth cycle. Let $U_n := \int_{t_{n-1}}^{t_n} u(s)\,ds$, and let $V_n := \int_{t_{n-1}}^{t_n} v(s)\,ds$, be the total uptime and downtime, respectively, during the nth cycle. Define $W_n := Z(t_n)$, the work in the system at the end of the nth cycle. Let $B_n^A := B_{t_{n-1},t_n}^A$ be the burstiness of the arrival process during the nth cycle, and let $B_n^D := B_{t_{n-1},t_n}^D$ be the burstiness of the departure process during the nth cycle. To avoid technical difficulties, we shall assume that $t_n \to \infty$ as $n \to \infty$, which is the case, for example, when the downtimes, V_n, $n \geq 1$, are bounded below by a positive constant.

Motivated by gated and exhaustive polling systems, we shall make the following assumption throughout this paper:

Output Assumption. *For all $n \geq 1$, $D(t_n) - D(t_{n-1}) \geq W_{n-1}$. That is, the output in each cycle is at least as great as the work in the system at the beginning of the cycle.*

Remark 1. Until now, our definitions of cycles and downtimes have been very vague. In the following sections, other assumptions about the downtimes during a cycle (bounds on their duration or on their average duration) will make these notions more precise and more related to what we understand by "cycles" and downtime in applications. Note that without any further restrictions, the Output Assumption is relatively innocuous. In fact, in the general setting that we have assumed up until this point (simply an input-output process, $\{Z(t), t \geq 0\}$, with an imbedded nondecreasing sequence of time points, $\{t_n, n \geq 0\}$), we are free to define the cycles so that the Output Assumption is trivially satisfied, so long as $D(t) \to \infty$ as $t \to \infty$. (For example, define the sequence, $\{t_n, n \geq 0\}$, recursively, by $t_n := \inf\{t : D(t) - D(t_{n-1}) \geq W_{n-1}\}$.) As we have indicated, however, in each of our three applications the definition of the cycles will be implicitly constrained by the bounds or growth conditions that we impose on the downtime, V_n, in the nth cycle. These constraints are inspired by our motivating examples of vacation and polling systems, in which a cycle begins when the server begins serving a particular customer class. For additional observations and a counterexample when the constraints are not satisfied, see Remark 6 at the end of Section 3.

The following example illustrates these issues and motivates our first result (Lemma 1 below).

Example 1. *A Globally Gated Polling System.* In a globally gated polling system, a single server attends m queues, labeled $i = 1, \ldots, m$, in sequence, serving at each queue all work that was present at the beginning of the cycle and then moving on to the next queue. In this setting, a cycle begins each time the server arrives and begins service at queue 1. The server spends a certain amount of downtime or *walking* time moving from queue i to queue $i + 1$. (We identify queue $m + 1$ with queue 1.) The total walking time in cycle n is V_n. For simplicity, assume that $V_n = V$, $n \geq 1$. If the departures were at a uniform rate, δ, then the duration of the nth cycle would be $V + W_{n-1}/\delta$. If arrivals were at a uniform rate, α, then W_n would equal $\alpha V + \rho W_{n-1}$, and therefore also equal $\rho^n W_0 + \alpha V \sum_{i=1}^n \rho^{n-i}$. Now suppose arrivals and departures are at these uniform rates, except for a single arrival burst of size X in the third cycle. Then for $n > 3$, $W_n = \rho^n W_0 + \alpha V \sum_{i=1}^n \rho^{n-i} + \rho^{n-3} X$.

The following is a similar result for our more general setup.

Lemma 1. *Suppose that $\rho = \alpha/\delta < 1$, and let N be a nonnegative integer. Then*

$$W_n \leq \rho^{n-N} W_N + \alpha \sum_{i=N+1}^{n} \rho^{n-i} V_i + \sum_{i=N+1}^{n} \rho^{n-i} B_i^A - \rho \sum_{i=N+1}^{n} \rho^{n-i} B_i^D, \quad n \geq N + 1. \tag{4}$$

Proof. First, note that

$$\begin{aligned} W_n &= W_{n-1} + A(t_n) - A(t_{n-1}) - [D(t_n) - D(t_{n-1})] \\ &= W_{n-1} + \alpha T_n + B_n^A - [\delta(T_n - V_n) + B_n^D], \end{aligned}$$

and hence

$$\delta(T_n - V_n) + B_n^D - W_{n-1} = \alpha T_n + B_n^A - W_n.$$

By our assumption that the output during the cycle must be no smaller than the work at the beginning of the cycle, both sides of this equation are nonnegative. Noting $\rho < 1$, we subtract ρ times the left side from the right side and rearrange them to obtain

$$W_n \leq \rho W_{n-1} + \alpha V_n + B_n^A - \rho B_n^D. \tag{5}$$

Beginning with W_N, and iteratively substituting for W_{n-1} in (5) leads to (4). □

Next, we bound the sum of the "discounted" bursts in Equation (4) by the sum of the corresponding "undiscounted" bursts when $\rho_i < 1$. This result, which is a consequence of the following lemma, is a key tool in establishing the bounds for all the models considered in the following sections.

Lemma 2. *Suppose we have two sequences of real numbers, $\{B_i\}_{i=1}^n$ and $\{\rho_i\}_{i=1}^n$, such that $0 \leq \rho_1 \leq \cdots \leq \rho_n$. Then*:

$$\begin{aligned} &\rho_1 B_1 + \cdots + \rho_n B_n \\ &\quad \leq \rho_n \max\{B_n, B_n + B_{n-1}, \ldots, B_n + \cdots + B_1\}. \end{aligned} \tag{6}$$

Proof. With $\rho_0 := 0$, we have:

$$\begin{aligned}
&\rho_1 B_1 + \cdots + \rho_n B_n \\
&\quad = \sum_{i=1}^{n} \sum_{j=1}^{i} (\rho_j - \rho_{j-1}) B_i \\
&\quad = \sum_{j=1}^{n} (\rho_j - \rho_{j-1})(B_n + B_{n-1} + \cdots + B_j) \\
&\quad \leqslant \sum_{j=1}^{n} (\rho_j - \rho_{j-1}) \max\{B_n, B_n + B_{n-1}, \ldots, B_n \\
&\qquad\qquad + \cdots + B_1\} \\
&\quad = \rho_n \max\{B_n, B_n + B_{n-1}, \ldots, B_n + \cdots + B_1\},
\end{aligned}$$

thus establishing (6). □

Lemma 3. *Suppose that $\rho = \alpha/\delta < 1$, and let N be a nonnegative integer. Suppose also that $V_i \leqslant V$ whenever $i > N$. Then, for $n \geqslant N + 1$:*

$$\begin{aligned}
W_n \leqslant{} & \rho^{n-N} W_N + \frac{\alpha(1 - \rho^{n-N})V}{1 - \rho} \\
& + \max\{0, B_n^A, B_n^A + B_{n-1}^A, \ldots, B_n^A \\
& \qquad + \cdots + B_{N+1}^A\} \\
& + \rho \max\{0, (-B_n^D), -(B_n^D + B_{n-1}^D), \ldots, \\
& \qquad -(B_n^D + \cdots + B_{N+1}^D)\};
\end{aligned} \tag{7}$$

$$\begin{aligned}
T_n \leqslant{} & \frac{\rho^{n-N} W_N}{\alpha(1 - \rho)} + \frac{(1 - \rho^{n-N})V}{(1 - \rho^2)} \\
& + \frac{1}{\delta(1 - \rho)} \max\{0, B_n^A, B_n^A + B_{n-1}^A, \ldots, B_n^A \\
& \qquad + \cdots + B_N^A\} \\
& + \frac{1}{\delta(1 - \rho)} \max\{0, -B_n^D, -(B_n^D + B_{n-1}^D), \\
& \qquad \ldots, -(B_n^D + \cdots + B_N^D\};
\end{aligned} \tag{8}$$

and for $t_n < t_{n+1}$,

$$\begin{aligned}
Z(t) \leqslant{} & \rho^{n-N} W_N + \frac{\alpha(2 - \rho - \rho^{n-N})V}{1 - \rho} \\
& + \max\{0, B_{t_n,t}^A, B_{t_n,t}^A + B_n^A, \ldots, B_{t_n,t}^A + B_n^A \\
& \qquad + \cdots + B_{N+1}^A\} \\
& + \rho \max\{0, -(B_{t_n,t}^D), -(B_{t_n,t}^D + B_n^D), \ldots, \\
& \qquad -(B_{t_n,t}^D + B_n^D + \cdots + B_{N+1}^D)\}.
\end{aligned} \tag{9}$$

Proof. For (7), we bound the terms of (4). Since the downtimes are bounded by V,

$$\alpha \sum_{i=N+1}^{n} \rho^{n-i} V_n \leqslant \alpha \sum_{i=N+1}^{n} \rho^{n-i} V = \frac{\alpha(1 - \rho^{n-N})V}{1 - \rho}. \tag{10}$$

The bounds on the other terms follow from Lemma 2.

To show (8), note that it follows from (3) that

$$\begin{aligned}
\delta T_n &= D(t_n) - D(t_{n-1}) + \delta V_n - B_n^D \\
&\leqslant W_{n-1} + \alpha T_n + B_n^A + \delta V_n - B_n^D,
\end{aligned}$$

since the output during a cycle can be no more than the work at the beginning of the cycle plus the input during the cycle. Thus,

$$\begin{aligned}
\alpha(1 - \rho)T_n &\leqslant \rho W_{n-1} + \rho B_n^A - \rho B_n^D + \alpha V_n \\
&\leqslant \rho^{n-N} W_N + \alpha \sum_{i=N+1}^{n} \rho^{n-i} V_i \\
&\quad + \rho\left(\sum_{i=N+1}^{n-1} \rho^{n-1-i} B_i^A + B_n^A \right) \\
&\quad - \rho \sum_{i=N+1}^{n} \rho^{n-i} B_i^D,
\end{aligned} \tag{11}$$

where we have used Lemma 1 to bound W_{n-1}. The result then follows using Lemma 2 and (10).

To prove (9), first note that

$$\begin{aligned}
D(t) - D(t_n) &= \delta(t - t_n) - \delta \int_{t_n}^{t} v(s)\, ds + B_{t_n,t}^D \\
&\geqslant \delta(t - t_n) - \delta V_{n+1} + B_{t_n,t}^D,
\end{aligned}$$

for $t \geqslant t_n$, which implies that

$$\begin{aligned}
Z(t) &= W_n + \alpha(t - t_n) + B_{t_n,t}^A - (D(t) - D(t_n)) \\
&\leqslant W_n + \rho(D(t) - D(t_n) + \delta V_{n+1} - B_{t_n,t}^D) \\
&\quad + B_{t_n,t}^A - (D(t) - D(t_n)) \\
&= W_n + \alpha V_{n+1} + B_{t_n,t}^A - \rho B_{t_n,t}^D \\
&\quad - (1 - \rho)(D(t) - D(t_n)) \\
&\leqslant W_n + \alpha V_{n+1} + B_{t_n,t}^A - \rho B_{t_n,t}^D.
\end{aligned} \tag{12}$$

Using inequality (4) for W_n then yields:

$$\begin{aligned}
Z(t) &\leqslant \rho^{n-N} W_N + \alpha\left(\sum_{i=N+1}^{n} \rho^{n-i} V_i + V_{n+1} \right) \\
&\quad + \left(\sum_{i=N+1}^{n} \rho^{n-i} B_i^A + B_{t_n,t}^A \right) \\
&\quad - \rho\left(\sum_{i=N+1}^{n} \rho^{n-i} B_i^D + B_{t_n,t}^D \right).
\end{aligned} \tag{13}$$

Bounding the terms on the right-hand side of (13) using Lemma 2 and (10) leads to the desired result. □

2. LINEAR BURSTINESS BOUNDS

For this section, in addition to the assumptions made in Sections 1 and 2, we assume that the input and output processes have bounds that are a special case of those studied by Cruz (1991, 1992). Specifically, we assume that:

$$A(t) - A(s) \leqslant \alpha(t - s) + \sigma_A, \quad 0 \leqslant s < t, \tag{14}$$

$$D(t) - D(s) \geqslant \delta\left(\int_s^t u(\tau)\, d\tau \right) - \sigma_D, \quad 0 \leqslant s < t, \tag{15}$$

where σ_A and σ_D are nonnegative constants. In other words, for both the input and output processes, the burstiness in any interval is bounded by a constant independent of the length of the interval. We also assume that the downtime during any cycle is bounded by a constant: $V_n \leqslant V$ for all $n \geqslant 1$.

Theorem 1. *Assume $\rho < 1$ and $W_0 = 0$. Then:*

$$W_n \leqslant \frac{\alpha V}{1-\rho} + \sigma_A + \rho\sigma_D, \quad n \geqslant 1,$$

$$T_n \leqslant \frac{V}{(1-\rho)^2} + \frac{\sigma_A + \sigma_D}{\delta(1-\rho)}, \quad n \geqslant 1,$$

$$Z(t) \leqslant \frac{\alpha(2-\rho)V}{1-\rho} + \sigma_A + \rho\sigma_D, \quad t \geqslant 0.$$

Proof. It follows from (2) and (14) that $B^A_{s,t} \leqslant \sigma_A$, and from (3) and (15) that $-B^D_{s,t} \leqslant \sigma_D$, for arbitrary $s < t$. Then in Lemma 3 one can replace all the max{ . . . } terms by the appropriate term, σ_A or σ_D, replace the W_N terms by 0, and remove the ρ^{n-N} terms to increase the right-hand sides. □

Remark 2. It is known that the average cycle time in certain stationary polling systems is given by $\bar{V}/(1-\rho)$, where $\bar{V}$ is the average downtime (walking time) per cycle. (See, e.g., Altman et al. 1992. A sample-path proof in a general setting is given in Theorem 4 in the next section.) The difference between this expression and the bound obtained in Theorem 1 can be large for ρ close to 1. The following questions arise: (1) Is the bound tight? (2) Can it be improved by giving more information on our system? (3) Is the condition $\rho < 1$ necessary?

It turns out that the difference between the bound given for T_n in Theorem 1 and $\bar{V}/(1-\rho)$ is due to the weak assumptions that we have made regarding the polling regime. (Recall that we require only that all the work present in the beginning of a cycle should leave during that cycle.) The following example shows that our bound is indeed tight, in the sense that any upper bound has to be at least $\bar{V}/(1-\rho)^2$.

Consider a polling system with a single queue, an arrival process with a constant rate of $\rho < 1$, a departure process with a constant rate of 1, and constant walking times equal to V. Assume that the polling uses the following gated discipline for a very long time: when the server leaves the queue a cycle begins; only customers present at the beginning of a cycle are served in the current cycle (cf. the example of a globally gated discipline in the previous section). Since $T_{n+1} = \rho T_n + V$, it follows that the average cycle time converges to $V/(1-\rho)$. Fix $\epsilon > 0$, and let m be such that $T_m > V/(1-\rho) - \epsilon$. Then, the work in the queue at the end of the mth cycle is greater than $\rho[V/(1-\rho) - \epsilon]$. Now, assume that at the $m+1$st cycle the polling discipline changes to exhaustive, i.e., the server remains at the queue until it is empty. At the beginning of the $m+1$st cycle, the amount of work in the system satisfies $Z(t_m) > \rho[V/(1-\rho) - \epsilon]$. When the server arrives at the queue, the amount of work in the queue is $Z(t_m) + \rho V$. So the time to empty the queue is:

$$\frac{Z(t_m) + \rho V}{1-\rho} > \rho\left(\frac{V}{1-\rho} - \epsilon + V\right)(1-\rho)^{-1},$$

so that

$$T_{m+1} > \rho\left(\frac{V}{1-\rho} - \epsilon + V\right)(1-\rho)^{-1} + V = \frac{V}{(1-\rho)^2} - \frac{\rho\epsilon}{1-\rho}. \tag{16}$$

Hence, the answer to question (1) is "yes"—any upper bound has to be at least $V/(1-\rho)^2$.

The fact that the bounds were shown to be tight implies that the condition $\rho < 1$ is a necessary condition for stability: we see that arbitrarily large workloads and cycle times can be obtained as $\rho \to 1$. A partial answer to question (2) is given by Altman and Kofman (1994), who consider polling systems with a finite number of queues. The bounds are improved by restricting to fixed polling disciplines (e.g., gated or exhaustive), and then exploiting specific characteristics of the polling regime. However, this requires a case-by-case analysis of different polling regimes. The methodology used by Altman and Kofman to improve the bounds involves inductive arguments to estimate different quantities at each time that the server arrives to a queue (rather than at each time that a cycle begins).

Finally, we show that the output from the system also satisfies a Cruz-type bound (of the same type as (14)), with the same average rate as the input average rate.

Theorem 2. *Under the assumptions of Theorem 1, for all $0 \leqslant t < t'$:*

$$D(t') - D(t) \leqslant \alpha(t'-t) + \frac{\alpha(2-\rho)V}{1-\rho} + \sigma_A + \rho\sigma_D.$$

Proof. For $0 \leqslant t < t'$, we have

$$D(t') - D(t) \leqslant Z(t) + A(t') - A(t) = Z(t) + \alpha(t'-t) + B^A_{t,t'}. \tag{17}$$

Suppose $t_n \leqslant t < t_{n+1}$, where $n \geqslant 0$. Using (13), we obtain:

$$D(t') - D(t) \leqslant \alpha(t'-t) + \frac{\alpha V}{1-\rho} + \alpha V + \left(\sum_{i=1}^{n} \rho^{n-i} B^A_i + B^A_{t_n,t} + B^A_{t,t'}\right) - \rho\left(\sum_{i=1}^{n} \rho^{n-i} B^D_i + B^D_{t_n,t}\right).$$

The desired result then follows, again using Lemma 2. □

Remark 3. The bounds obtained in this section may be useful in the analysis and design of communications systems. The characterization of the input process in terms of the average rate and burstiness is typical of the traffic in a communications network at the output of a spacer or a leaky bucket. The bound for the workload in the system can be used to obtain upper bounds on the size of the buffers required so as to guarantee no losses. The strict bound on the cycle time is also quite desirable in communications applications, especially in order to guarantee some service quality for synchronous traffic (such as voice or video). Indeed, several token-ring protocols possess some distributed mechanism to enforce a strict upper bound on the cycle time; an example is the *FDDI* protocol. The fact that an upper bound for the cycle times is obtained for the Cruz-type arrival process suggests that previous distributed mechanisms to enforce such constraints can be replaced by shaping the input flows to a token ring using leaky buckets. Moreover, the fact that the output process satisfies a Cruz-type bound may be used in analyzing networks, where the output from one token ring can be the input to another element of the network. (Applications to token rings are discussed in more detail in Altman and Kofman and Altman et al. 1994.)

3. LIMITING AVERAGE RATES

In this section we consider systems in which the input and output processes have limiting average rates. Specifically, in the same spirit as in Stidham and El-Taha, we make the following assumptions:

$$\lim_{t\to\infty} \frac{A(t)}{t} = \alpha, \tag{18}$$

$$\lim_{t\to\infty} \frac{\int_0^t u(s)\,dD(s)}{\int_0^t u(s)\,ds} = \delta. \tag{19}$$

We interpret δ as the sample-path version of the conditional departure rate, given that the system is "up." We also assume that:

$$\lim_{n\to\infty} \frac{V_n}{t_n} = 0. \tag{20}$$

Theorem 3. *Assume* (18), (19), (20), *and* $\rho = \alpha/\delta < 1$. *Then*:

$$\lim_{n\to\infty} \frac{W_n}{t_n} = 0, \tag{21}$$

$$\lim_{n\to\infty} \frac{T_n}{t_n} = 0, \tag{22}$$

$$\lim_{t\to\infty} \frac{Z(t)}{t} = 0. \tag{23}$$

Proof. Let $\epsilon > 0$ be given, and suppose N is large enough so that for all $i \geqslant N$ and $t \geqslant t_N$:

$$(\alpha - \epsilon)t \leqslant A(t) \leqslant (\alpha + \epsilon)t,$$

$$(\delta - \epsilon) \int_0^t u(s)\,ds \leqslant \int_0^t u(s)\,dD(s) \leqslant (\delta + \epsilon) \cdot \int_0^t u(s)\,ds,$$

$$V_i \leqslant \epsilon t_i.$$

Then, for $n > N$ and $t_N \leqslant s < t \leqslant t_n$,

$$A(t) - A(s) \leqslant \alpha t + \epsilon t - \alpha s + \epsilon s \leqslant \alpha(t - s) + 2\epsilon t_n,$$

so that $B^A_{s,t} \leqslant 2\epsilon t_n$. Similarly:

$$D(t) - D(s) \geqslant \int_s^t u(x)\,dD(x) \geqslant \delta\left(\int_s^t u(x)\,dx\right) - 2\epsilon t_n,$$

so that $-B^D_{s,t} \leqslant 2\epsilon t_n$. Replacing the max$\{\dots\}$ terms in Lemma 3 with $2\epsilon t_n$ (in the case of (7) and (8)), or $2\epsilon t_{n+1}$ (in the case of (9)), and replacing V with ϵt_n or ϵt_{n+1}, and dividing by t_n or t_{n+1} (as the case may be) gives:

$$\frac{W_n}{t_n} \leqslant \frac{\rho^{n-N} W_N}{t_n} + \left[\frac{\alpha(1 - \rho^{n-N})}{1 - \rho} + 2(1 - \rho)\right]\epsilon,$$

$$\frac{T_n}{t_n} \leqslant \frac{\rho^{n-N} W_n}{t_n \alpha(1 - \rho)} + \left[\frac{(1 - \rho^{n-N})}{(1 - \rho)^2} + \frac{4}{\delta(1 - \rho)}\right]\epsilon,$$

$$\frac{Z(t)}{t_{n+1}} \leqslant \frac{\rho_{n-N} W_N}{t_{n+1}} + \left[\frac{\alpha(2 - \rho^{n-N})}{1 - \rho} + 2(1 - \rho)\right]\epsilon,$$

for $t_n \leqslant t \leqslant t_{n+1}$.

Taking limits as $n \to \infty$, (21) and (22) follow. Define $N(t) := \max\{n : t_n < t\}$. Then:

$$\frac{Z(t)}{t} = \frac{Z(t)}{t_{N(t)+1}} \cdot \frac{t_{N(t)+1}}{t} \leqslant \frac{Z(t)}{t_{N(t)+1}} \cdot \frac{t_{N(t)+1}}{t_{N(t)}}, \tag{24}$$

and thus (23) follows from (22). □

Remark 4. The condition (23) is a kind of pathwise stability condition, called *rate stability* by El-Taha and Stidham (see also Stidham and El-Taha). Given $\lim_{t\to\infty} A(t)/t = \alpha < \infty$, it is equivalent to equality of the input and output rates:

$$\lim_{t\to\infty} \frac{D(t)}{t} = \lim_{t\to\infty} \frac{A(t)}{t} = \alpha.$$

As an elementary consequence, we obtain the following result for the long-run fraction of time that the system is down.

Corollary 1. *Under the conditions of Theorem* 3:

$$\liminf_{t\to\infty} \frac{\int_0^t v(s)\,ds}{t} \geqslant 1 - \rho. \tag{25}$$

If in addition $\int_0^t v(s)\,dD(s) = 0$ *for all* $t \geqslant 0$, *then*:

$$\lim_{t\to\infty} \frac{\int_0^t v(s)\,ds}{t} = 1 - \rho. \tag{26}$$

Proof. We have:

$$Z(0) + A(t) = Z(t) + D(t) \geqslant Z(t) + \int_0^t u(s)\,dD(s), \tag{27}$$

so that

$$\frac{Z(0)}{t} + \frac{A(t)}{t} \geqslant \frac{Z(t)}{t} + \frac{\int_0^t u(s)\,dD(s)}{\int_0^t u(s)\,ds} \cdot \frac{\int_0^t u(s)\,ds}{t}.$$

Taking limits and using (23), one obtains:

$$\alpha \geqslant \delta\left(\limsup_{t\to\infty} \frac{\int_0^t u(s)\,ds}{t}\right),$$

from which (25) follows immediately. When $\int_0^t v(s)\, dD(s) = 0$, we have $D(t) = \int_0^t u(s)\, dD(s)$ for all $t \geq 0$; hence, (26) holds. □

In the following theorem we derive an expression for the long-run average cycle length.

Theorem 4. *Under the assumptions of Theorem* 3:

$$\liminf_{n\to\infty} \frac{t_n}{n} \leq (1-\rho)^{-1} \liminf_{n\to\infty} \frac{\sum_{i=1}^n V_i}{n}, \tag{28}$$

$$\limsup_{n\to\infty} \frac{t_n}{n} \geq (1-\rho)^{-1} \limsup_{n\to\infty} \frac{\sum_{i=1}^n V_i}{n}. \tag{29}$$

If, in addition, $\lim_{n\to\infty} n^{-1} \sum_{i=1}^n V_i = V < \infty$, *then*:

$$\lim_{n\to\infty} \frac{t_n}{n} = \frac{V}{1-\rho}. \tag{30}$$

Proof. Let n and m be positive integers, where $n > m$. Then,

$$W_m + A(t_n) - A(t_m) = W_n + D(t_n) - D(t_m),$$

or, equivalently,

$$W_m + \alpha(t_n - t_m) + B^A_{t_m,t_n}$$
$$= W_n + \delta(t_n - t_m) + B^D_{t_m,t_n} - \delta \sum_{i=m+1}^{n} V_i.$$

Rearranging terms, we have,

$$(\delta - \alpha)t_n + W_n - B^A_{t_m,t_n} + B^D_{t_m,t_n} = \delta \sum_{i=1}^{n} V_i + E_m, \tag{31}$$

where $E_m := W_m - \delta \sum_{i=1}^m V_i + (\delta - \alpha)t_m$. Now let ϵ and N be given as in Theorem 3, and suppose that $n > m \geq N$. Then, $|B^A_{t_M,t_n}| \leq 2\epsilon t_n$ and $|B^D_{t_M,t_n}| \leq 2\epsilon t_n$. Moreover, from (21) it follows that $0 \leq W_n/t_n \leq \epsilon$ for sufficiently large n. Using these bounds in (31), and dividing by n, we obtain

$$(\delta - \alpha - 5\epsilon)\frac{t_n}{n} \leq \frac{\delta \sum_{i=1}^n V_i}{n} + \frac{E_m}{n} \leq (\delta - \alpha + 5\epsilon)\frac{t_n}{n}.$$

Now fix $m \geq N$ and let $n \to \infty$. Since E_m is constant and ϵ was arbitrary, we conclude that:

$$(\delta - \alpha) \liminf_{n\to\infty} \frac{t_n}{n} \leq \delta \liminf_{n\to\infty} \frac{\sum_{i=1}^n V_i}{n}$$
$$\leq \delta \limsup_{n\to\infty} \frac{\sum_{i=1}^n V_i}{n} \leq (\delta - \alpha) \limsup_{n\to\infty} \frac{t_n}{n},$$

from which the desired result follows upon dividing by $\delta - \alpha$. □

Remark 5. Note that the proof of Theorem 4 did not use our basic assumption that $D(t_n) - D(t_{n-1}) \geq W_{n-1}$, for all $n \geq 1$, except implicitly when we invoked Theorem 3 to conclude that (21) holds, that is,

$$\lim_{n\to\infty} \frac{W_n}{t_n} = 0.$$

In fact, the proof demonstrates that Theorem 3 holds for any system in which (18), (19), and (20) hold, with $\alpha < \delta$, provided that (21) also holds. In particular, it suffices that the system be rate stable: $Z(t)/t \to 0$ as $t \to \infty$.

The fact that the average cycle length equals $V/(1 - \rho)$ for a large class of polling systems seems to be part of the folklore, but we could not find a previous pathwise proof in the literature. A proof is given in Altman et al. (1992), but the proof there uses the theory of stationary point processes. The assumptions are in some sense orthogonal to ours, in that they assume stationarity, but not ergodicity, of the processes involved.

Remark 6. The pathwise assumptions in this section and the last are usually easier to verify in polling systems if the cycles are defined in the "natural" way. That is, each cycle ends when the server completes the walking time between the last queue and first queue. Then as long as the walking time between each of the queues is bounded or has a long-run average, the total walking time in each cycle will be bounded or have a long-run average. The situation in systems where the cycles are not defined in the "natural" way becomes more complicated. For example, as we observed in Remark 1 in Section 1, we can define the cycles so that the Output Assumption is trivially satisfied, but then (typically) the bounds or growth conditions on downtimes per cycle, or the constraint that $\rho < 1$, will not be satisfied.

To illustrate, consider a system with two queues and cycles defined trivially as above. The service intervals and walking times are always one time unit. The server can serve two customers per service interval at queue 1, and one customer per service interval at queue 2. All arrivals occur during the walking time between queue 2 and queue 1, with four customers arriving at queue 1, and one customer arriving at queue 2. We define the server to be "up" when it is serving at queue 2. Thus, $\rho = 5/8 < 1$. However, it is not difficult to show that for any initial work load, the "artificial" cycles we defined to meet the output assumption will start to double in length each successive cycle after some initial number of cycles. As a result, $\{V_n\}$ is unbounded, and $V_n/t_n \to 3/8$ as $n \to \infty$. The restrictions on V_n in this and the previous sections are not met.

This example points up the delicate interplay between the Output Assumption, the requirement that $\rho < 1$, and the bounds or growth conditions on V_n. In particular, we cannot simply define cycles so that the output condition is met, and define uptime so that $\rho < 1$, if as a result the constraints on downtime per cycle end up being violated.

4. LAW OF THE ITERATED LOGARITHM

In this section we show how our pathwise analysis can be used to characterize the asymptotic behavior of W_n more precisely, in the presence of information about the rate of convergence of $A(t)/t$ to α. To keep the exposition simple, we shall confine our attention to the special case of a constant unit output rate while the system is up. That is, we assume that:

$$\frac{\int_0^t u(x)\, dD(x)}{\int_0^t u(x)\, dx} = \delta = 1, \quad t \geqslant 0. \tag{32}$$

Thus, $\rho = \alpha$. (Similar results will hold in the general case.)

Let $h(t)$ be a monotone nondecreasing function of t, such that $\lim_{t\to\infty} h(t) = \infty$. We make the following assumptions about $\{A(t),\, t \geqslant 0\}$ and $\{V_n,\, n \geqslant 1\}$:

$$\limsup_{t\to\infty} \left|\frac{A(t) - \rho t}{h(t)}\right| \leqslant 1, \tag{33}$$

$$\lim_{n\to\infty} \frac{V_n}{h(t_n)} = 0. \tag{34}$$

Theorem 5. *Assume* (32), (33), (34), *and* $\rho = \alpha < 1$. *Then*:

$$\limsup_{n\to\infty} \frac{W_n}{h(t_n)} \leqslant 2. \tag{35}$$

Proof. Let $\epsilon > 0$ be given. It follows from (33) and (34) that there exists an integer N such that, for all $t_N \leqslant s < t \leqslant t_n$,

$$A(t) - A(s) \leqslant \rho(t - s) + 2(1 + \epsilon)h(t_n),$$

and, for all $N + 1 \leqslant i \leqslant n$,

$$V_i \leqslant \epsilon h(t_n).$$

It then follows from (4) and Lemma 2 that

$$W_n \leqslant \rho^{n-N} W_N + \left(\frac{\rho\epsilon}{1 - \rho} + 2(1 + \epsilon)\right)h(t_n),$$

for all $n \geqslant N + 1$. Since ϵ was arbitrary, the desired result, (35), then follows by dividing both sides by $h(t_n)$ and letting $n \to \infty$. □

Theorem 5 applies, for example, to stochastic models in which the input process, $\{A(t),\, t \geqslant 0\}$, satisfies a Law of the Iterated Logarithm (*LIL*), in which case (33) holds with probability one with

$$h(t) = \sigma\sqrt{2t \log\log t}. \tag{36}$$

5. THE STOCHASTIC CASE

In the previous three sections we considered a single sample path for which we obtained both performance bounds and stability conditions. In this section we shall discuss stochastic models. We first note that the assumptions on the arrival and departure processes described in Section 3 (specifically, (18) and (19)) hold for any ergodic arrival and departure processes with probability 1. In that case, Theorems 1 and 4 still hold, where the equalities and inequalities should be interpreted in the almost sure sense. The assumptions in Section 4 also hold in an a.s. sense for many ergodic arrival processes, with h given in (36); hence, Theorem 5 may be interpreted in the a.s. sense for these cases as well.

Now, we consider stochastic arrival processes of the type introduced by Yaron and Sidi (1994) and Chang (1994) characterized by bounds on the tail distribution. We shall consider a general bound and then specialize to both exponential and polynomial bounds. More precisely, we assume that there exist some constant $\rho > 0$, and a nonincreasing function $G: R \to R_+$ such that for all $\sigma > 0$:

$$P\{A(t) - A(s) - \rho(t - s) > \sigma\} \leqslant G(\sigma), \tag{37}$$
$$0 \leqslant s < t < \infty.$$

We shall further restrict the arrival process and the polling discipline to ensure that (37) holds when replacing t and s by some random times. Specifically, we shall assume that it holds also for t and s chosen as the beginning and end of a cycle. That is, for all $\sigma > 0$,

$$\begin{aligned} P(B_n^A > \sigma) &= P\{A(t_{n+1}) - A(t_n) - \rho(t_{n+1} - t_n) > \sigma\} \\ &\leqslant G(\sigma), \quad n \geqslant 0. \end{aligned} \tag{38}$$

Condition (38) is natural for a quite general class of arrival processes. At the end of this section we show that this includes i.i.d. arrivals, each of which brings an amount of work that satisfies some tail condition.

Throughout this section we shall make the following simplifying assumptions: (i) departures occur at a constant unit rate, and (ii) the downtimes are uniformly bounded, i.e., there exists some constant V such that $V_n \leqslant V$ a.s.

We show below that if $\rho < 1$, then these assumptions imply that the workload at the beginning of cycles and at arbitrary times, as well as the cycle durations, have similar types of bounds on their distributions. For the case of exponential bound, this will imply that all moments of the workload in the systems are uniformly bounded in time.

Moreover, we show that the departure process has a characterization of the same type as the input process, with the same ρ. This again is important when considering a network that consists of a number of elements, each of which maps input processes of the type (37) into departure processes of the same type (with possibly different constants). Denote

$$\bar{W} := \frac{\rho V}{1 - \rho}, \quad \bar{T} := \frac{V}{(1 - \rho)^2}, \quad \bar{Z} := \frac{\rho(2 - \rho)V}{1 - \rho}.$$

Theorem 6. *Assume that* (38) *holds. Assume* $\rho < 1$ *and* $W_0 = 0$. *Then the workload at the end of the nth cycle and the duration of the nth cycle satisfy*:

$$P\{W_n - \bar{W} > \sigma\} \leqslant \sum_{i=0}^{n-1} G(a_i \rho^{-i}), \tag{39}$$

$$P\{T_n - \bar{T} > \sigma\} \leqslant \sum_{i=0}^{n-1} G(\rho(1 - \rho)a_i \rho^{-i}), \tag{40}$$

for any nonnegative constants $a_0, a_1, \ldots, a_{n-1}$ *such that* $a_0 + a_1 + \cdots + a_{n-1} \leqslant \sigma$.

Proof. It follows from Lemma 1 that for all $n = 0, 1, \ldots,$ $W_n \leqslant \bar{W} + \sum_{i=1}^{n} \rho^{n-i} B_i^A$. Hence

$$
\begin{aligned}
P\{W_n > \bar{W} + \sigma\} &\leq P\left\{\sum_{i=1}^{n} \rho^{n-i} B_i^A > \sigma\right\} \\
&\leq \sum_{i=1}^{n} P\{\rho^{n-i} B_i^A > a_{n-i}\} \\
&= \sum_{i=1}^{n} P\{B_i^A > a_{n-i} \rho^{-n+i}\} \\
&\leq \sum_{i=1}^{n} G(a_{n-i} \rho^{-n+i}).
\end{aligned}
$$

Similarly, it follows from (11) that:

$$
\begin{aligned}
(1-\rho) T_n &\leq \frac{V}{1-\rho} + \sum_{i=1}^{n-1} \rho^{n-1-i} B_i^A + B_n^A \\
&\leq \frac{V}{1-\rho} + \sum_{i=1}^{n} \rho^{n-1-i} B_i^A.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
P\{T_n - \bar{T} > \sigma\} &\leq P\left\{\sum_{i=1}^{n} \rho^{n-1-i} B_i^A > \sigma(1-\rho)\right\} \\
&\leq \sum_{i=1}^{n} P\{\rho^{n-1-i} B_i^A > a_{n-i}(1-\rho)\} \\
&\leq \sum_{i=1}^{n} G(\rho(1-\rho) a_{n-i} \rho^{-n+i}),
\end{aligned}
$$

from which the theorem follows. □

Corollary 2. *Assume that* (38) *holds. Assume* $\rho < 1$ *and* $W_0 = 0$. *Choose some arbitrary* $\epsilon > 0$. *Consider an exponential bound on the tail distribution of the arrivals* (38):

$$G(\sigma) = K_0 \exp(-k_0 \sigma), \quad \sigma \geq 0, \tag{41}$$

where k_0, K_0 *are some positive constants. Then, the workload at the end of the nth cycle satisfies*

$$
\begin{aligned}
&P\{W_n - \bar{W} > \sigma + \epsilon\} \\
&\quad \leq K_1 \exp\{-(\sigma + \epsilon) k_1\}, \quad n \geq 1,
\end{aligned} \tag{42}
$$

for all $\sigma \geq 0$, *where* k_1 *and* K_1 *are constants given by*

$$k_1 := (1-\rho)^2 k_0, \quad K_1 := \frac{K_0}{1 - \exp(-k_1 \epsilon)}.$$

The duration of the nth cycle satisfies

$$
\begin{aligned}
&P\{T_n - \bar{T} > \sigma + \epsilon\} \\
&\quad \leq K_1' \exp\{-(\sigma + \epsilon) k_1'\}, \quad n \geq 1,
\end{aligned} \tag{43}
$$

for all $\sigma \geq 0$, *where* k_1' *and* K_1' *are constants given by*

$$k_1' := \rho(1-\rho)^3 k_0, \quad K_1' := \frac{K_0}{1 - \exp(-k_1' \epsilon)}.$$

Proof. The proof is obtained by applying Theorem 6. Set

$a_i := (i+1)\rho^i d, \quad$ where $d := (1-\rho)^2 \sigma'$.

Note that $\sum_{i=0}^{\infty} a_i = \sigma'$. We obtain in particular,

$$
\begin{aligned}
P(W_n - \bar{W} > \sigma') &\leq \sum_{i=0}^{\infty} G(a_i \rho^{-i}) \\
&= K_0 \sum_{i=0}^{\infty} \exp\{-k_0 (i+1) d\} \\
&= \frac{K_0}{1 - \exp\{-d k_0\}} \exp\{-d k_0\},
\end{aligned}
$$

for all n. For $\sigma' := \sigma + \epsilon$, we have

$$\exp\{-d k_0\} = \exp\{-(\sigma + \epsilon) k_1\} \leq \exp\{-\epsilon k_1\}, \tag{44}$$

from which the bound for W_n follows. The proof of (43) is similar. □

Corollary 3. *Assume that* (38) *holds. Assume* $\rho < 1$ *and* $W_0 = 0$. *Consider a polynomial bound on the tail distribution of the arrivals* (38):

$$G(\sigma) = K_0 \sigma^{-m}, \tag{45}$$

where K_0, m *are positive constants, with* $m > 1$. *Then, the workload at the end of the nth cycle satisfies*

$$P\{W_n - \bar{W} > \sigma\} \leq K_1 \sigma^{-m}, \quad n \geq 1, \tag{46}$$

for all $\sigma \geq 0$, *where* K_1 *is given by*

$$K_1 := K_0 (1-\rho)^{-2m} \sum_{l=1}^{\infty} l^{-m}.$$

The nth cycle duration satisfies

$$P\{T_n - \bar{T} > \sigma\} \leq K_1' \sigma^{-m}, \quad n \geq 1, \tag{47}$$

for all $\sigma \geq 0$, *where* K_1' *is given by*

$$K_1' := K_0 \rho^{-m} (1-\rho)^{-3m} \sum_{l=1}^{\infty} l^{-m}.$$

Proof. As in Corollary 2, set:

$a_i := (i+1)\rho^i d, \quad$ where $d := \sigma(1-\rho)^2$.

The proof for W_n is again obtained by applying Theorem 6, and

$$G(a_i \rho^{-i}) = G((i+1)d) = K_0 d^{-m} (i+1)^{-m}.$$

The proof for T_n is similar. □

Next, we obtain bounds on the amount of work at an arbitrary moment and on the departure process. Let $n(t)$ denote the (random) number of cycles that started prior to time t. We shall need the following assumption:

$$
\begin{aligned}
P(B_{t_{n(t)},t}^A > \sigma) &= P\{A(t) - A(t_{n(t)}) - \rho(t - t_{n(t)}) > \sigma\} \\
&\leq G(\sigma), \quad t \geq 0.
\end{aligned} \tag{48}
$$

Theorem 7. *Assume that* (38) *and* (48) *hold. Assume* $\rho < 1$ *and* $W_0 = 0$. *Then, the workload at an arbitrary time point t and the departures during an arbitrary interval* $(t, t']$ *satisfy*

$$P\{Z(t) - \bar{Z} > \sigma\} \leq \sum_{l=0}^{\infty} G(a_l \rho^{-1}) + G(a'),$$

$$P\{D(t') - D(t) - \rho(t' - t) > \sigma\}$$
$$\leq \sum_{l=0}^{\infty} G(a_l \rho^{-l}) + G(a') + G(a''),$$

for any nonnegative constants a', a'', a_0, a_1, . . . *such that* $a' + a'' + \sum_{i=0}^{\infty} a_i \leq \sigma$.

Proof. It follows from (13) that:

$$Z(t) \leq \bar{Z} + \sum_{i=1}^{n} \rho^{n-i} B_i^A + B^A_{t_{n(t)},t}.$$

The bound for $Z(t)$ is now obtained as in the proof of Theorem 6. The bound for $D(t') - D(t)$ is obtained similarly by noting that:

$$D(t') - D(t) - \rho(t' - t) \leq Z(t) + B^A_{t,t'}.$$

□

As was done in Corollaries 2 and 3 for W_n and T_n, one may also obtain explicit exponential or polynomial bounds on the tail probabilities of $Z(t)$ and on the departure process by using Theorem 7 when such bounds hold for the arrival process.

5.1. Sufficient Conditions for the Burstiness Constraint (38)

Since (38) is slightly different from the standard burstiness constraints (37) on the arrivals (Chang 1994, Yaron and Sidi 1994), we present sufficient conditions for it to hold.

We first show that for both exponential-type and polynomial-type bounds, (38) is satisfied under the following condition. Assume that there is a constant $\rho > 0$, and a nonincreasing function $\bar{G}$ such that, for all $\sigma > 0$,

$$P\{A(t_n + t) - A(t_n) \geq \rho \cdot t + \sigma\} \leq \bar{G}(\sigma), \qquad (49)$$
$$n \geq 0,\ t \geq 0.$$

Indeed, if (49) holds, then for all $\hat{\rho} > \rho$, $\Delta > 0$, $n \geq 0$, $\sigma \geq 0$:

$$P\{A(t_{n+1}) - A(t_n) \geq \hat{\rho}(t_{n+1} - t_n) + \sigma\}$$
$$= \sum_{l=0}^{\infty} P\{A(t_{n+1}) - A(t_n) \geq \hat{\rho}(t_{n+1} - t_n) + \sigma;\ l\Delta \leq t_{n+1} - t_n \leq (l + 1)\Delta\}$$
$$\leq \sum_{l=0}^{\infty} P\{A(t_n + (l + 1)\Delta) - A(t_n) \geq \rho\Delta(l + 1) + \sigma + \Delta(\hat{\rho} l - \rho(l + 1))\}$$
$$\leq \sum_{l=0}^{\infty} \bar{G}(\sigma + \Delta(\hat{\rho} l - \rho(l + 1))).$$

Consider first the exponential bound, i.e., assume that there exist $\bar{K}_0$, $\bar{k}_0$ such that

$$\bar{G}(\sigma) = \bar{K}_0 \cdot \exp\{-\bar{k}_0 \cdot \sigma\}.$$

Then we obtain

$$P\{A(t_{n+1}) - A(t_n) \geq \hat{\rho}(t_{n+1} - t_n) + \sigma\}$$
$$\leq \sum_{l=0}^{\infty} \bar{K}_0 \exp\{-\bar{k}_0(\sigma + \Delta(\hat{\rho} l - \rho(l + 1)))\}$$
$$= \bar{C} \cdot \bar{K}_0 \cdot \exp\{-\bar{k}_0 \sigma\},$$

where

$$\bar{C} := \exp\{\Delta \rho \bar{k}_0\}/[1 - \exp\{-\Delta \bar{k}_0(\hat{\rho} - \rho)\}] < \infty.$$

For the case of polynomial bounds, i.e.,

$$\bar{G}(\sigma) = C/\sigma^{m+1},$$

for some C, $m > 0$, we get $G(x) = c/x^m$ (where $c > 0$ is some constant).

Next, we present sufficient conditions for (49). We shall assume that work arrives to the system with "customers." Denote by τ_i the time between the arrival of customer i and customer $i + 1$, $i \geq 1$. Denote by s_i the "work" that customer i brings (i.e., the service time required by it). Without loss of generality, assume that customer 1 arrives to an empty system at time $t = 0$.

Lemma 4. *Assume*:

(i) *the interarrival times of the customers* $\{\tau_i\}$ *form an i.i.d. sequence with finite positive mean*;

(ii.a) *the service times* $\{s_i\}$ *form an i.i.d. sequence*;

(ii.b) *there exists a constant* $\lambda > 0$ *such that* $E\ exp\{\lambda s_1\} < \infty$, *and the sequences* $\{\tau_i\}$ *and* $\{s_i\}$ *are independent*;

(iii) $\{t_n\}$ *are "independent of the future" in the following sense*: *for all* r, n, *an event* $B_{n,r} \equiv \{\sum_{j=1}^{r-1} \tau_j < t_n \leq \sum_{j=1}^{r} \tau_j\}$ *is independent of the* σ*-algebra generated by the sequences* $\{\{\tau_l\}_{l>r}; \{s_l\}_{l>r}\}$.

Then, there exists a constant $\rho > 0$ *such that, for all* $\sigma > 0$, *the bound* (49), *and hence* (38), *hold with exponential tail.*

If (ii.b) is replaced by the assumption that $Es_1^{m+2} < \infty$ *for some integer* m, *then* (49), *and hence* (38), *hold with a polynomial tail*: $\bar{G}(x) = C/x^{m+1}$ *and* $G(x) = c/x^m$, *for some constants* c *and* C.

Since the proof is technical, it will be left to the appendix.

Remark 7. Property (iii) is the only one that is related to the specific polling policy that is used. It is known to hold for a very large class of policies, including gated type and exhaustive type policies. (See Altman and Foss.)

Remark 8. We may generalize the arrival process considered in Lemma 4 to a regenerative type of arrival process and still get the same type of results. Consider some regeneration times T_1, T_2, ... Define $\tau_i = T_{i+1} - T_i$. The location of arrivals and the workloads they require in each regenerative period are given by a measure M defined on

($[0, \tau_i) \times \mathscr{E}$), where $\mathscr{E}$ is the set of possible values of the workloads. For any Borel sets $B_1 \subset \mathfrak{R}$, $B_2 \subset \mathscr{E}$, $M(B_1 \times B_2)$ is the amount of workload that arrives at time points in B_1 with values in B_2. Assume that τ_i, and the amount of work that arrives during a regenerative period both have finite first moments. Under fairly general conditions, this type of arrival process allows, in particular, for K independent i.i.d. arrival streams into K nodes. (See Foss and Rybko.)

APPENDIX
PROOF OF LEMMA 4

For all r, n, we have

$$A(t_n + t) - A(t_n) \leqslant s_{r+1} + \sum_{i=1}^{\eta_r(t)} s_{r+i+1},$$

a.s. on the event $B_{n,r}$, where $\eta_r(t) = \max\{m \geqslant 0 : \tau_{r+1} + \cdots + \tau_{r+m} \leqslant t\}$, $r \geqslant 0$. Then,

$$\begin{aligned} P\{A(t_n + t) - A(t_n) > \rho \cdot t + \sigma\} &= \sum_{r=1}^{\infty} P(B_{n,r}) \cdot P\{A(t_n + t) - A(t_n) > \rho \cdot t + \sigma | B_{n,r}\} \\ &\leqslant \sum_{r=1}^{\infty} P(B_{n,r}) \cdot P\left\{ s_{r+1} + \sum_{i=1}^{\eta_r(t)} s_{r+i+1} \right. \\ &\qquad \left. > \rho \cdot t + \sigma \middle| B_{n,r} \right\} \\ &= \sum_{r=1}^{\infty} P(B_{n,r}) \cdot P\left\{ s_{r+1} + \sum_{i=1}^{\eta_r(t)} s_{r+i+1} > \rho \cdot t + \sigma \right\} \\ &= P\left\{ s_1 + \sum_{i=1}^{\eta_0(t)} s_{i+1} > \rho \cdot t + \sigma \right\}, \end{aligned}$$

where the next-to-last equality follows from condition (iii), and the last equality from assumptions (i) and (ii.a). We see that our problem is reduced to estimating the probability

$$P_{t,\sigma} := P\left\{ s_1 + \sum_{i=1}^{\eta(t)} s_{i+1} > \rho \cdot t + \sigma \right\}, \tag{50}$$

where $\eta(t) = \eta_0(t)$. Set $a = E\tau_1$, $b = Es_1$. Choose $\rho > b/a$ and set $c = \rho a/b > 1$. Fix some constants $\tilde{c}$ and $\tilde{\sigma}$ satisfying $1 < \tilde{c} < c$, $\tilde{\sigma} = \alpha\sigma$, where $0 < \alpha < 1$ is arbitrary. Denote $x = (\tilde{c}/a)t + \tilde{\sigma}$, and let $\lceil x \rceil$ denote the smallest integer greater than or equal to x. We shall decompose (50) into two parts, which we shall estimate separately:

$$P_{t,\sigma} \leqslant P_1 + P_2, \quad P_1 := P\{\eta(t) \geqslant \lceil x \rceil\},$$

$$P_2 := P\left\{ \sum_{i=1}^{\lceil x \rceil} s_i > \rho t + \sigma \right\}.$$

Let $T_n := \tau_1 + \cdots + \tau_n$, and let $\gamma > 0$ be arbitrary. Then,

$$\begin{aligned} P_1 &= P\{T_{\lceil x \rceil} \leqslant t\} = P\{\exp\{-\gamma T_{\lceil x \rceil}\} \geqslant \exp\{-\gamma t\}\} \\ &\leqslant [\psi(\gamma)]^x \exp(\gamma t) = [\psi(\gamma)]^{\tilde{\sigma}} [\psi(\gamma) \exp\{\gamma a/\tilde{c}\}]^{(\tilde{c}/a)t}, \end{aligned} \tag{51}$$

where $\psi(\gamma) := E \exp\{-\gamma\tau_1\}$, the Laplace-Stieltjes transform of τ_1 evaluated at γ. Now,

$$\begin{aligned} \psi(\gamma) &= 1 - \gamma a + o(\gamma) = 1 - \gamma a/\tilde{c} \\ &\quad - \gamma a(\tilde{c} - 1)/\tilde{c} + o(\gamma) \end{aligned}$$

as $\gamma \to 0$. Hence, for $\gamma \ll 1$, it follows from (51) that

$$P_1 \leqslant \exp\{\log \psi(\gamma)\tilde{\sigma}\} = \exp\{\log \psi(\gamma)\alpha\sigma\},$$

for all $t \geqslant 0$. (So far, we have not used any assumption on the tail probability of the service times.)

Next, we bound P_2. Let t_0 and ϵ be some strictly positive constants such that

$$(\tilde{c}t_0/a + 1)(b + \epsilon) = \rho t_0.$$

Consider first the case $\tilde{c}t/a + \tilde{\sigma} \geqslant \tilde{c}t_0/a$. Note that:

$$\begin{aligned} \lceil \tilde{c}t/a + \tilde{\sigma} \rceil (b + \epsilon) &\leqslant (\tilde{c}t/a + \tilde{\sigma} + 1)(b + \epsilon) \\ &\leqslant \frac{\tilde{c}t/a + \tilde{\sigma}}{\tilde{c}t_0/a} (\tilde{c}t_0/a + 1)(b + \epsilon) \\ &= \frac{\tilde{c}t/a + \tilde{\sigma}}{\tilde{c}t_0/a} \rho t_0 = \rho t + \rho a \tilde{\sigma}/\tilde{c}. \end{aligned}$$

Recall that in the definition of $\tilde{\sigma}$, α was arbitrary. We now choose a specific α so that $\rho a \alpha/\tilde{c} =: q$ is less than 1. Hence, $\rho a \tilde{\sigma}/\tilde{c} = q\sigma$. Denote $R := \sup_{n \geqslant 0} \Sigma_{i=1}^{n} (s_i - b - \epsilon)$. We have

$$\begin{aligned} P_2 &\leqslant P\left\{ \sum_{i=1}^{\lceil \tilde{c}t/a + \tilde{\sigma} \rceil} s_i > \rho t + \sigma \right\} \\ &\leqslant P\left\{ \sum_{i=1}^{\lceil \tilde{c}t/a + \tilde{\sigma} \rceil} (s_i - b - \epsilon) > (1 - q)\sigma \right\} \\ &\leqslant P\{R > (1 - q)\sigma\} =: Z. \end{aligned}$$

Now, it is well known (e.g., see Asmussen 1987, p. 184) that $Es_1^{m+2} < \infty$, $m \geqslant 0$, implies that $ER^{m+1} < \infty$, so that $Z \leqslant M\sigma^{-(m+1)}$ for some constant $M < \infty$; likewise, if $E \exp \lambda s_1 < \infty$ for some $\lambda > 0$, then $Z \leqslant \exp\{-M\sigma\}$ for some constant $M > 0$ (e.g., see Kingman 1970).

For $\tilde{c}t/a + \tilde{\sigma} < \tilde{c}t_0/a$, we have

$$\begin{aligned} P_2 &\leqslant P\left\{ \sum_{i=1}^{\lceil \tilde{c}t_0/a \rceil} s_i > \sigma \right\} \leqslant \sum_{i=1}^{\lceil \tilde{c}t_0/a \rceil} P\left\{ s_i > \frac{\sigma}{\lceil \tilde{c}t_0/a \rceil} \right\} \\ &= \lceil \tilde{c}t_0/a \rceil P\left\{ s_1 > \frac{\sigma}{\lceil \tilde{c}t_0/a \rceil} \right\}. \end{aligned}$$

Hence, P_2 has the same tail behavior as s_1. This concludes the proof. □

ACKNOWLEDGMENT

E. Altman and S. Foss were supported by the European Grant BRA-QMIPS of CEC DG XIII.

The research of E. Riehl and S. Stidham was supported by the National Science Foundation under Grant No. NCR-9406823. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

ALTMAN, E. AND S. G. FOSS. 1997. Polling on a Space With General Arrival and Service Time Distribution. *O. R. Letts.* **20,** 187–194.

ALTMAN, E., S. G. FOSS, E. R. RIEHL, AND S. STIDHAM. 1994. Sample Path Analysis of Token Rings. *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks*. In *Proceedings of the 14th International Teletraffic Congress*. Antibes Juan-les-Pins, France, 811–820.

ALTMAN, E. AND D. KOFMAN. 1994. Bounds for Performance Measures of Token Rings. *IEEE/ACM Transactions on Networking*, 292–299.

ALTMAN, E., P. KONSTANTOPOULOS, AND Z. LIU. 1992. Stability, Monotonicity, and Invariant Quantities in General Polling Systems. *Queueing Systems: Theory and Applications,* **11,** 35–58.

ALTMAN, E. AND H. LEVY. 1994. Queueing in Space. *Adv. in Appl. Probab.* **26,** 1095–1116.

ALTMAN, E. AND Z. LIU. 1994. Improving the Stability Characteristics of Asynchronous Traffic in FDDI Token Rings. In *High Speed Networks and Their Performance*, H. G. Perros and Y. Viniotis (eds.) 441–460. North-Holland, Amsterdam.

ALTMAN, E. AND F. SPIEKSMA. 1996. Ergodicity, Moment Stability and Central Limit Theorems of Station Times in Polling Systems. *Comm. Statist. Stochastic Models* **12,** 307–328.

ASMUSSEN, S. 1987. *Applied Probability Processes and Queues*. J. Wiley, New York.

BOROVKOV, A. A. 1984. *Asymptotic Methods in Queueing Theory*. J. Wiley, New York.

BOROVKOV, A. A. AND R. SCHASSBERGER. 1994. Ergodicity of a Polling System. *Stochastic Process. and Appl.* **50,** 253–262.

CHANG, C.-S. 1994. Stability, Queue Length, and Delay in Deterministic and Stochastic Queueing Networks. *IEEE Trans. on Automat. Control.* **39,** 913–931.

CRUZ, R. L. 1991. A Calculus for Network Delay, Part I: Network Elements in Isolation. *IEEE Trans. Inform. Theory,* **37,** 114–131.

CRUZ, R. L. 1992. A Calculus for Network Delay, Part II: Network Analysis. *IEEE Trans. Inform. Theory.* **37,** 132–141.

EL-TAHA, M. AND S. STIDHAM JR. 1993. Sample-Path Analysis of Stochastic Discrete-Event Systems. *Discrete Event Dynamic Systems: Theory and Applications,* **3,** 325–346.

FOSS, S. AND A. RYBKO. 1996. Stability of Multiclass Jackson—Type Networks. *Markov Process. Related Fields* **2,** 461–486.

FRICKER, C. AND M. R. JAIBI. 1994. Monotonicity and Stability of Periodic Polling Models. *Queueing Systems,* **15,** 211–238.

GEORGIADIS, L. AND W. SZPANKOWSKI. 1992. Stability of Token-Passing Rings. *Queueing Systems: Theory and Applications,* **11,** 7–33.

GEORGIADIS, L., W. SZPANKOWSKI, AND L. TASSIULAS. 1993. Stability of Ring Networks With Spatial Re-use. In *Proceedings of the Conference on Applied Probability in Engineering, Computer and Communication Sciences*. INRIA/ORSA/TIMS/SMAI, June 16–18, Paris.

KINGMAN, J. F. C. 1970. Inequalities in the Theory of Queues. *J. Royal Stat. Soc., Series B,* **32,** 102–110.

KROESE, D. P. AND V. SCHMIDT. 1992. A Continuous Polling System With General Service Times. *Anns. Appl. Prob.* **2,** 906–927.

KROESE, D. P. AND V. SCHMIDT. 1994. Single-server Queues With Spatially Distributed Arrivals. *QUESTA,* **17,** 317–345.

MASSOULIÉ, L. 1993. On the Construction of Stationary Point Processes in Polling Systems. In *Proceedings of the Conference on Applied Probability in Engineering, Computer and Communication Sciences.* INRIA/ORSA/TIMS/SMAI, June 16–18, Paris.

RESING, J. A. C. 1993. Polling Systems and Multitype Branching Processes. *Queueing Systems,* **13,** 409–426.

STIDHAM, S., JR. AND M. EL-TAHA. 1993. A Note on Sample-Path Stability Conditions for Input–Output Processes. *O. R. Letts.* **14,** 1–7.

YARON, O. AND M. SIDI. 1994. Performance and Stability of Communication Networks via Robust Exponential Bounds. *IEEE/ACM Trans. Networking,* **1,** 372–385.

ZHDANOV, V. S. AND E. A. SAKSONOV. 1979. Conditions of Existence of Steady-state Modes in Cyclic Queueing Systems. *Automat. Remote Control,* **40,** 176–184.