Journal of Applied Mathematics and Stochastic Analysis
7, Number 3, 1994, 437-456

SAMPLE-PATH STABILITY CONDITIONS FOR MULTISERVER INPUT-OUTPUT PROCESSES

MUHAMMAD EL-TAHA[1]
University of Southern Maine
Department of Mathematics and Statistics
96 Falmouth Street
Portland, ME 04103, USA

SHALER STIDHAM JR.
University of North Carolina
Department of Operations Research
CB 3180, Smith Building
Chapel Hill, NC 27599-3180, USA

(Received: April, 1994; revised: 1st rev. October, 1993, 2nd rev. March, 1994)

ABSTRACT

We extend our studies of sample-path stability to multiserver input-output processes with conditional output rates that may depend on the state of the system and other auxiliary processes. Our results include processes with countable as well as uncountable state spaces. We establish rate stability conditions for busy period durations as well as the input during busy periods. In addition, stability conditions for multiserver queues with possibly heterogeneous servers are given for the workload, attained service, and queue length processes. The stability conditions can be checked from parameters of primary processes, and thus can be verified *a priori*. Under the rate stability conditions, we provide stable versions of Little's formula for single server as well as multiserver queues. Our approach leads to extensions of previously known results. Since our results are valid pathwise, non-stationary as well as stationary processes are covered.

Key words: Sample-Path Analysis, Stability, Rate Stability, Input-Output Process, Multiserver Queues, Busy Periods, Little's Formula.

AMS (MOS) subject classifications: 60G17, 60K25, 60K99.

1. Introduction

In this paper, we continue our investigation of sample-path conditions for rate stability in general input-output processes. A process is said to be *rate stable* if its evolution is $o(t)$ as $t \to \infty$

[1]The research of this author was done in part when the author was visiting INRS-Telecommunications, University of Quebec, Montreal, Canada (Summer, 1992), and completed while he was visiting the Department of Operations Research, University of North Carolina at Chapel Hill (Spring, 1994).

(see El-Taha and Stidham [3, 4, 5], Stidham and El-Taha [17], Mazumdar et al. [10], Guilleminand Mazumdar [7], Altman, Foss, Riehl, and Stidham [1]). The model investigated here is an input-output process in which the output process is composed of multiple, possibly non-homogeneous, streams. We establish conditions for rate stability that can be verified from information on input (primary) processes in a deterministic framework that makes it possible to characterize the sample-path behavior of non-stationary stochastic processes.

The issue of stability for non-stationary processes, using a sample-path framework, has been recently considered by several authors. El-Taha and Stidham [3, 4, 5] provide a sample-path characterization of (rate) stability and establish connections between rate stability and other measures of interest, such as the finiteness of the limiting average number of customers in a queueing system. Mazumdar, Guillemin, Badrinath, and Kannurpatti [10], study rate stability in the context of the workload process in a $G/G/1$ queue using sample-path arguments. Stidham and El-Taha [17] consider an input-output process with a single output stream and establish rate stability conditions using only sample-path information available from primary processes. Guillemin and Mazumdar [7] provide a pathwise proof for rate stability of the workload process in a multiserver queue with $FCFS$ discipline. See also Mazumdar et al. [9], Guillemin et al. [6], and Altman, Foss, Riehl, and Stidham [1].

This paper provides a generalization of stability results given by Stidham and El-Taha [17]. In Section 2 we prove our basic stability result by establishing sufficient conditions that are easily verifiable from conditions on primary processes. Processes with non-countable as well as integer state spaces are considered and studied within one unifying framework. Multiple output processes are allowed, where individual output streams can also be heterogeneous. In particular, in the context of a multiserver queueing system, individual servers, when active, are permitted to process work at different, possibly state-dependent rates. Section 3 focuses on full-busy-period durations and establishes their rate stability under the sufficient conditions given in Section 2. Section 4 contains applications to special cases of queueing systems and other input-output processes that extend those given by [17], as well as new ones. In particular, we give a sample-path proof for rate stability of the workload, queue length, and attained service processes in multiserver queues. We provide a counterexample to confirm an assertion that the workload process exhibits a *stronger* form of rate stability that the queue length process. Finally, in Section 5 we give sample-path proofs of Little's formula ($L = \lambda W$) for rate-stable queues, under more general sample-path conditions than previously given.

2. Basic results

The setting is that of Section 5 of El-Taha and Stidham [4] and Stidham and El-Taha [17]. That is, we consider a non-negative real-valued deterministic process, $Z = \{Z(t), t \geq 0\}$ - an *input-output* process - in which $Z(t) \geq 0$ represents quantity in a system, such as number of customers or workload in a queue. We assume that $\{Z(t), t \geq 0\}$ is right continuous with left-hand limits, and that

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

Here $A(t)$ is the cumulative input to the system in $[0, t]$, and $D(t)$: $= \sum_{i=1}^{c} D_i(t)$, where $D_i(t)$ is the cumulative output in $[0, t]$ from the i^{th} output stream, $i = 1, \ldots, c$. We assume that $\{A(t), t \geq 0\}$ and $\{D_i(t), t \geq 0\}$, $i = 1, \ldots, c$, are non-decreasing, right-continuous processes. Thus $Z(t)$ has bounded variation on finite t-intervals. Note that $D(t) \leq Z(0) + A(t)$, since $Z(t) \geq 0$.

We make no stochastic assumptions. In a stochastic setting, $\{Z(t), t \geq 0\}$ is to be interpreted as a fixed sample path (realization) of the stochastic process in question. The notion of rate

stability investigated in this paper is formally given by El-Taha and Stidham [4] (see also [3], [4], [5], [17], [10], [7], [1]). We repeat this definition for convenience.

Definition. An input-output process $\{Z(t), t \geq 0\}$ is said to be *rate stable* if $t^{-1}Z(t) \to 0$, as $t \to \infty$.

An immediate consequence of (1) (see Lemma 2.1 of Stidham and El-Taha [17]) is that in a *rate-stable* input-output process, if either the long-run average input or output rate exists, then both exist and they are equal.

Associated with the i^{th} output stream is an auxiliary right-continuous, $(0,1)$-valued process, $\{B_i(t), t \geq 0\}$, $i = 1,\ldots,c$. If $B_i(t) = 1$ then we say the i^{th} output stream is *active* at time t. The precise interpretation of an active output stream will depend on the problem context. The essential assumptions, expressed below in the hypotheses to Theorem 2.2, are that the long-run average system output rate of each stream while active is well-defined, that the total of these rates is greater than the input rate, and that $Z(t) = o(t)$ along any sequence of time points $t \to \infty$ at which not all streams are active. These assumptions may be taken in effect as a minimal definition of what it means for an output stream to be active. In applications to a $G/G/1$ queue, for example, $B_i(t) = 1(0)$ if server i is busy (idle) at time t (see Section 4). But in general we do not assume that output cannot occur from stream i while $B_i(t) = 0$, only that output occurs at a given rate while $B_i(t) = 1$.

We give a preliminary result that is of independent interest.

Lemma 2.1. *Consider the input-output process* $\{Z(t), t \geq 0\}$ *defined by* (1). *Let* α *and* δ_i, $i = 1,\ldots,c$, *be non-negative constants. Suppose*

(*i*) *the input process satisfies*

$$\limsup_{t \to \infty} t^{-1}A(t) \leq \alpha, \tag{2}$$

(*ii*) *the output process satisfies*

$$\liminf_{t \to \infty} \frac{\sum_{i=1}^{c} \int_0^t B_i(s)dD_i(s)}{\sum_{i=1}^{c} \int_0^t \delta_i B_i(s)ds} \geq 1, \tag{3}$$

where $0 < \alpha < \delta := \sum_{i=1}^{c} \delta_i < \infty$.

Then the event $\{B(t) \leq c-1\}$ *occurs infinitely often as* $t \to \infty$. *That is, for every* $t_0 \geq 0$, *there exists a* $t \geq t_0$ *such that* $B(t) \leq c-1$.

Proof: The proof is by contradiction. It follows from (2) and (3) that, for every $\epsilon > 0$, there exists a $T < \infty$ such that

$$A(t) \leq (\alpha + \epsilon)t, \quad t \geq T, \tag{4}$$

$$(1-\epsilon)U(t) \leq \sum_{i=1}^{c} \int_0^t B_i(s)dD_i(s), \quad t \geq T, \tag{5}$$

where $U(t) := \sum_{i=1}^{c} \int_0^t \delta_i B_i(s)ds, t \geq 0$.

Now suppose $0 < \epsilon < (\delta - \alpha)/(\delta + 1)$. Suppose that the event $\{B(t) \leq c-1\}$ does not occur infinitely often as $t \to \infty$. Then there exists a t_0 such that, for all $t \geq t_0$, $B(t) = c$, so that $B_i(t) = 1$ for all i, $1 \leq i \leq c$, and $U(t) = U(t_0) + \delta(t - t_0)$. Without loss of generality take

$t_0 \geq T$. Hence, using (4) and (5), we have for all $t \geq t_0$

$$Z(t) = Z(t_0) + (A(t) - A(t_0)) - \sum_{i=1}^{c} \int_{t_0}^{t} B_i(s)dD_i(s)$$

$$= Z(t_0) - A(t_0) + \sum_{i=1}^{c} \int_{0}^{t_0} B_i(s)dD_i(s) + A(t) - \sum_{i=1}^{c} \int_{0}^{t} B_i(s)dD_i(s)$$

$$\leq Z(t_0) - A(t_0) + \sum_{i=1}^{c} \int_{0}^{t_0} B_i(s)dD_i(s) + (\alpha + \epsilon)t - (1-\epsilon)U(t)$$

$$= Z(t_0) - A(t_0) + \sum_{i=1}^{c} \int_{0}^{t_0} B_i(s)dD_i(s) - (1-\epsilon)(U(t_0) - \delta t_0)$$

$$+ (\alpha - \delta + (\delta+1)\epsilon)t.$$

Since $(\alpha - \delta + (\delta+1)\epsilon) < 0$, by choosing t sufficiently large, we can make the quantity on the right-hand side of the last equality negative, thus leading to a contradiction of $Z(t) \geq 0$. Therefore, for every $t_0 \geq T$ there exists at least one $t > t_0$ such that $B(t) \leq c-1$. That is, the event $\{B(t) \leq c-1\}$ occurs infinitely often as $t \to \infty$. □

Remark 2.1. The above result is of independent interest. It gives sufficient conditions for the existence of construction points (points that start full busy periods) for the $\{Z(t), t \geq 0\}$ process (see Baccelli and Brémaud [2], pp. 38-46). For example, let $Z(t)$ represent the queue length at time t in a $G/G/1$ queue. Take $c = 1$ and let $B(t) = 1\{Z(t) > 0\}$. Then $\{B(t) = 0\} \equiv \{B(t) \leq c-1\}$ represents the event that the server is idle. Let $\{T_n, n \geq 1\}$ be a sequence of arrival instants such that $Z(T_n) > 0$ for all $n \geq 1$. Then the sequence $\{b_k, k \geq 1\}$ of construction points is defined by $b_0 = 0$ and $b_k := min\{T_n : T_n > b_{k-1}, Z(T_n -) = 0\}$, for $k = 1, \ldots$.

The following result, an extension of Theorem 2.2 of Stidham and El-Taha [17], gives sufficient conditions for rate stability of the process $\{Z(t), t \geq 0\}$.

Theorem 2.2. *Consider the input-output process* $\{Z(t), t \geq 0\}$ *defined by* (1). *Let* α *and* δ_i, $i = 1, \ldots, c$ *be non-negative constants. Suppose*

(*i*) *the input process satisfies*

$$\lim_{t \to \infty} t^{-1} A(t) = \alpha; \tag{6}$$

(*ii*) *the output process satisfies*

$$\lim_{t \to \infty} \frac{\sum_{i=1}^{c} \int_{0}^{t} B_i(s)dD_i(s)}{\sum_{i=1}^{c} \int_{0}^{t} \delta_i B_i(s)ds} = 1; \tag{7}$$

(*iii*) *for every nonnegative real sequence* $\{t_n\}$, $t_n \to \infty$ *as* $n \to \infty$, *such that* $B(t_n) \leq c-1$ *for all* $n \geq 1$,

$$t_n^{-1} Z(t_n) \to 0 \text{ as } n \to \infty, \tag{8}$$

where $0 \leq \alpha < \delta := \sum_{i=1}^{c} \delta_i < \infty$.

Then the process $\{Z(t), t \geq 0\}$ *defined by* (1) *is rate stable.*

Remark 2.2. Roughly speaking, condition (*iii*) states that no output stream will be inactive when the process $\{Z(t), t \geq 0\}$ takes sufficiently large values. Although conditions (7) and (8) involve secondary processes, they are easily verifiable *a priori* in many cases of interest. In applications to $G/G/c$ queues condition (8) is immediate when $c = 1$, or when $\{Z(t), t \geq 0\}$ is the queue-length process representing the number of customers in the system. When $\{Z(t), t \geq 0\}$ is the workload process in a multiserver system, condition (8) can be verified provided that the long-run average work presented to the system at transition epochs is finite. (See Section 4 below on application).

Proof: The proof is by contradiction and is similar to that of Theorem 2.2 of Stidham and El-Taha [17]. Suppose that $\{Z(t), t \geq 0\}$ is not *rate stable*. Then there exists a $\gamma > 0$ and an increasing sequence of time points $\{\tau_n, n \geq 1\}$, with $\tau_n \to \infty$ as $n \to \infty$, such that $Z(\tau_n) \geq \tau_n \gamma$ for all $n \geq 1$. Note that $B(\tau_n) = c$ for all $\tau_n \geq T$, for some $T < \infty$; otherwise we contradict (8). If $\alpha = 0$, then $\{Z(t), t \geq 0\}$ is trivially rate stable. So suppose $\alpha > 0$. Without loss of generality we assume that $Z(0) = 0$ and that $\gamma < \alpha$.

Let $U(t) := \sum_{i=1}^{c} \int_0^t \delta_i B_i(s) ds$, $t \geq 0$, and observe that $U(t) \to \infty$ as $t \to \infty$. Suppose not. Then it follows from the right-continuity of $B_i(\cdot)$ that $B_i(t) = 0$, $i = 1, \ldots, c$, and hence $B(t) = 0$, for all sufficiently large t, which contradicts $B(\tau_n) = c$ for all $\tau_n \geq T$.

Now it follows from (6) and (7) that, for every $\epsilon > 0$, there exists a $T = T(\epsilon) < \infty$ such that

$$(\alpha - \epsilon)t \leq A(t) \leq (\alpha + \epsilon)t, \quad t \geq T, \tag{9}$$

$$(1 - \epsilon)U(t) \leq \sum_{i=1}^{c} \int_0^t B_i(s) dD_i(s) \leq (1 + \epsilon)U(t), \quad t \geq T. \tag{10}$$

Let $a_n := sup\{s : s < \tau_n, B(s) \leq c - 1\}$. Then it follows that $B(s) = c$ and hence $B_i(s) = 1$, $i = 1, \ldots, c$, for all $a_n < s \leq \tau_n$. Hence

$$D(\tau_n) = D(a_n) + \sum_{i=1}^{c} \int_{a_n}^{\tau_n} B_i(s) dD_i(s), \tag{11}$$

$$U(\tau_n) = U(a_n) + \delta(\tau_n - a_n). \tag{12}$$

Moreover, the above arguments and Lemma 2.1 show that $a_n \to \infty$ as $n \to \infty$. Now choose $\epsilon < min\{\gamma/(2\delta + 3), (\delta - \alpha)/(\delta + 1)\}$. For n sufficiently large, we have $a_n > T(\epsilon)$. For such n, using (9), (10), (11), and (12) it follows that

$$\begin{aligned} Z(a_n -) &= Z(\tau_n) + (A(a_n -) - A(\tau_n)) - (D(a_n -) - D(\tau_n)) \\ &= Z(\tau_n) - A(\tau_n) + A(a_n -) + \sum_{i=1}^{c} \int_0^{\tau_n} \delta_i B_i(s) dD_i(s) - \sum_{i=1}^{c} \int_0^{a_n} \delta_i B_i(s) dD_i(s) \\ &\geq \gamma \tau_n - (\alpha + \epsilon)\tau_n + (\alpha - \epsilon)a_n + (1 - \epsilon)U(\tau_n) - (1 + \epsilon)U(a_n) \end{aligned}$$

$$= \gamma\tau_n - (\alpha+\epsilon)\tau_n + (\alpha-\epsilon)a_n + (1-\epsilon)[U(a_n) + \delta(\tau_n - a_n)] - (1+\epsilon)U(a_n)$$

$$= \gamma\tau_n - 2\epsilon a_n - 2\epsilon U(a_n) + [\delta - \alpha - (\delta+1)\epsilon](\tau_n - a_n)$$

$$> \gamma\tau_n - 2\epsilon a_n - 2\epsilon U(a_n)$$

$$\geq (\gamma - (2\delta+2)\epsilon)a_n$$

$$> [(2\delta+3)\epsilon - (2\delta+2)\epsilon]a_n$$

$$= \epsilon a_n.$$

But by (8), $Z(a_n -) \leq \epsilon a_n$. Thus we have a contradiction and the proof is complete. □

Theorem 2.2 remains valid for other variants of condition (7), as in Corollary 2.4 below. We also point out that δ_i may be interpreted as the long-run average amount of work that can be processed per unit time by the i^{th} stream while active. Thus our formulation allows queueing models with heterogeneous servers.

In general we allow for the possibility that some output can occur from the i^{th} stream while it is inactive. In may applications (e.g., the $G/G/c$ queue with heterogeneous servers) this is not the case: either the i^{th} stream is active and producing output at rate δ_i or it is inactive and producing no output. In such cases, the following corollary of Theorem 2.2 provides additional results. (It actually makes a slightly weaker assumption.)

Corollary 2.3. *Suppose conditions* (6), (7) *and* (8) *of Theorem 2.2 are satisfied, and* $0 < \alpha < \delta$. *Suppose also that*

$$\lim_{t\to\infty} t^{-1} \sum_{i=1}^{c} \int_0^t (1 - B_i(s))dD_i(s) = 0. \tag{13}$$

Then

$$\lim_{t\to\infty} t^{-1} \sum_{i=1}^{c} \delta_i \int_0^t (1 - B_i(s))ds = \delta - \alpha. \tag{14}$$

Proof: In (1), divide by t and take limits as $t\to\infty$, using Theorem 2.2 to obtain

$$\lim_{t\to\infty} t^{-1}U(t) = \alpha, \tag{15}$$

from which the desired result follows by subtracting both sides from δ. □

Remark 2.4. Suppose that the conditions of Corollary 2.3 hold and that the following limit is well defined, for each $i = 1,\ldots,c$:

$$p_i(0) := \lim_{t\to\infty} t^{-1} \int_0^t (1 - B_i(s))ds.$$

Then it follows from (14) that

$$\sum_{i=1}^{c} (\delta_i/\delta)p_i(0) = 1 - \rho, \tag{16}$$

where $\rho := \alpha/\delta$. The left-hand side of this equation can be interpreted as the long-run weighted fraction of time an output stream is inactive. Thus, (16) gives an extension to multiserver systems of the well-known formula for the fraction of time the server is idle in a single-server

facility.

Remark 2.5. If the limit in (13) exists but does not equal zero, then one can obtain an extension of the above corollary. Specifically, suppose that

$$\lim_{t\to\infty} \frac{\int_0^t \sum_{i=1}^c (1 - B_i(s)dD_i(s)}{\int_0^t \sum_{i=1}^c \delta_i(1 - B_i(s))ds} = \beta. \tag{17}$$

Then $\beta < 1$ and

$$\lim_{t\to\infty} t^{-1} \sum_{i=1}^c \delta_i \int_0^t (1 - B_i(s))ds = \frac{\delta - \alpha}{1 - \beta}. \tag{18}$$

This formula allows for output to occur when output streams are idle, i.e., when $B_i(t) = 0$.

Another variant of Theorem 2.2 that is useful, particularly when considering state- and/or time-dependent service rates, is given in the following corollary.

Corollary 2.4. *Consider the multiserver input-output process* $\{Z(t), t \geq 0\}$ *as in Theorem 2.2. Suppose that conditions* (6) *and* (8) *of Theorem 2.2 are satisfied, and condition* (7) *is replaced by*

$$\lim_{t\to\infty} \frac{\int_0^t B_i(s)dD_i(s)}{\int_0^t B_i(s)ds} = \delta_i, \; i = 1,\ldots,c. \tag{19}$$

Then the process $\{Z(t), t \geq 0\}$ *defined by* (1) *is rate stable.*

3. Busy period fluctuations

In this section we show that, under the conditions for rate stability of $\{Z(t), t \geq 0\}$, the sequence of durations of full busy periods is also rate stable. A *full busy period* in a multiserver input-output process begins when all servers become active and ends at the next time point when at least one server becomes inactive.

Lemma 2.1 shows the existence of infinitely many full busy periods - more precisely, the existence of an infinite sequence $\{t_n\}$ such that $t_n \to \infty$ as $n \to \infty$ and $B(t_n) \leq c - 1$, for all $n \geq 1$. Now let $b_0 = 0$ and, for $n = 1, 2, \ldots$, define

$$b_n := inf\{t > b_{n-1} : B(t-) \leq c - 1, B(t) = c\},$$

$$e_n := inf\{t > b_n : B(t-) = c, B(t) \leq c - 1\},$$

$$B_n := e_n - b_n,$$

$$A_n := A(e_n) - A(b_n).$$

We interpret b_n and e_n, respectively, as the beginning and end of the n^{th} full busy period. We interpret B_n as the length of the n^{th} full busy period and A_n as the input during the n^{th} busy period. By Lemma 2.1, both b_n and $e_n \to \infty$ as $n \to \infty$.

Theorem 3.1. *Consider the input-output process defined by* (1). *Suppose that the conditions of Corollary 2.3 are satisfied. Then*

(*i*) $B_n/b_n \to 0, \quad n \to \infty,$
(*ii*) $A_n/b_n \to 0, \quad n \to \infty.$

Proof: (*i*) Rewrite B_n/b_n as

$$B_n/b_n = \frac{B_n \sum_{i=1}^{c} \delta_i \int_0^{b_n} (1 - B_i(s))ds}{b_n \sum_{i=1}^{c} \delta_i \int_0^{e_n} (1 - B_i(s))ds}$$

$$= \frac{B_n(\delta b_n - \sum_{i=1}^{c} \delta_i \int_0^{b_n} B_i(s)ds)}{b_n \sum_{i=1}^{c} \delta_i \int_0^{e_n} (1 - B_i(s))ds}$$

$$= \frac{B_n \delta b_n - B_n \sum_{i=1}^{c} \delta_i \int_0^{b_n} B_i(s)ds - b_n \sum_{i=1}^{c} \delta_i \int_0^{b_n} B_i(s)ds + b_n \sum_{i=1}^{c} \delta_i \int_0^{b_n} B_i(s)ds}{b_n \sum_{i=1}^{c} \delta_i \int_0^{e_n} (1 - B_i(s))ds}$$

$$= \frac{e_n}{\sum_{i=1}^{c} \delta_i \int_0^{e_n} (1 - B_i(s))ds} \left(\frac{\sum_{i=1}^{c} \delta_i \int_0^{e_n} B_i(s)ds}{e_n} - \frac{\sum_{i=1}^{c} \delta_i \int_0^{b_n} B_i(s)ds}{b_n} \right).$$

By Corollary 2.3 the first term on the right-hand side of the last equality converges to $(\delta - \alpha)^{-1}$, and the second and third terms both converge to α, as $n \to \infty$. Thus $B_n/b_n \to 0$ as $n \to \infty$.

(*ii*) We have

$$\frac{A_n}{b_n} = \frac{A(e_n) \cdot e_n}{e_n \cdot b_n} - \frac{A(b_n)}{b_n}$$

$$= \frac{A(e_n)}{e_n} \frac{b_n + B_n}{b_n} - \frac{A(b_n)}{b_n}.$$

The result follows by taking limits as $n \to \infty$ and appealing to part (*i*) and (6). □

Remark 3.1. Theorem 3.1 remains valid if we replace the hypothesis that the conditions of Corollary 2.3 are satisfied by the condition that $\lim_{t \to \infty} t^{-1} \sum_{i=1}^{c} \delta_i \int_0^t (1 - B_i(s))ds$ is well defined and

$$0 < \lim_{t \to \infty} t^{-} \sum_{i=1}^{c} \delta_i \int_0^t (1 - B_i(s))ds \leq \delta. \tag{20}$$

This is important because it may be possible to verify the above condition by means other than checking the conditions of Corollary 2.3.

The single-server case

Although full busy periods in multiserver queues are rate stable under the conditions of Corollary 2.3, the same need not be true of busy cycles, that is, successive visits to empty and idle state. For example, a multiserver system can be rate stable and still have one or more servers remaining active at all time. Our next result shows that in single-server queues busy cycles are rate stable.

Consider a single-server system in which $Z(t) = 0$ implies that $B(t) = B_1(t) = 0$. Then Lemma 2.1 shows the existence of infinitely many cycles; in other words it shows that the existence of an infinite sequence $\{t_n\}$ such that $t_n \to \infty$ as $n \to \infty$ and $Z(t_n) = 0$.

Now, for $n = 1, 2, \ldots$; b_n, e_n and B_n simplify to

$$b_n = inf\{t > b_{n-1} : Z(t-) = 0, Z(t) > 0\},$$

$$e_n = inf\{t > b_n : Z(t-) > 0, Z(t) = 0\},$$

$$B_n = e_n - b_n.$$

Now, define

$$I_n := b_{n+1} - e_n,$$

$$C_n := B_n + I_n,$$

$$A_n := A(b_{n+1}) - A(b_n).$$

As before, b_n and e_n are the beginning and the end, respectively, of the n^{th} busy period. We also interpret B_n, I_n, C_n, and A_n, respectively, as the length of the n^{th} busy period, idle period, and busy cycle, and the input during the n^{th} busy cycle. Under the conditions of Lemma 2.1, $b_n \to \infty$ as $n \to \infty$.

Theorem 3.2. *Consider the input-output process* $\{Z(t), t \geq 0\}$ *defined by* (1), *with* $c = 1$. *Suppose that*

$$\lim_{t\to\infty} t^{-1} \int_0^t 1\{Z(s) = 0\} dD(s) = 0. \tag{21}$$

Suppose also that the input process satisfies

$$\lim_{t\to\infty} t^{-1} A(t) = \alpha, \tag{22}$$

and the output process satisfies

$$\lim_{t\to\infty} \frac{\int_0^t 1\{Z(s) > 0\} dD(s)}{\int_0^t 1\{Z(s) > 0\} ds} = \delta, \tag{23}$$

where $0 \leq \alpha < \delta$. *Then,*

(*i*) $C_n / b_n \to 0$, *as* $n \to \infty$,
(*ii*) $A_n / b_n \to 0$, *as* $n \to \infty$.

Proof: First note that the conditions of Corollary 2.3 are satisfied with $c = 1$ and

$B(t) = B_1(t) = 1\{Z(t) > 0\}$. To prove part (i), we first show that $I_n/e_n \to 0$ as $n \to \infty$. Now,

$$I_n/e_n = \frac{I_n \int_0^{e_n} 1\{Z(s) > 0\} ds}{\int_0^{b_{n+1}} 1\{Z(s) > 0\} ds}$$

$$= \frac{I_n (e_n - \int_0^{e_n} 1\{Z(s) = 0\} ds)}{\int_0^{b_{n+1}} 1\{Z(s) > 0\} ds}$$

$$= \frac{I_n e_n - I_n \int_0^{e_n} 1\{Z(s) = 0\} ds - e_n \int_0^{e_n} 1\{Z(s) = 0\} ds + e_n \int_0^{e_n} 1\{Z(s) = 0\} ds}{\int_0^{b_{n+1}} 1\{Z(s) > 0\} ds}$$

$$= \frac{b_{n+1}}{\int_0^{b_{n+1}} 1\{Z(s) > 0\} ds} \left(\frac{\int_0^{b_{n+1}} 1\{Z(s) = 0\} ds}{b_{n+1}} - \frac{\int_0^{e_n} 1\{Z(s) = 0\} ds}{e_n} \right).$$

By Corollary 2.3, the first term on the right-hand-side of the last equality converges to $(\rho)^{-1} = \delta/\alpha$, and the second and third terms both converge to $1 - \rho$ as $n \to \infty$. Thus $I_n/e_n \to 0$ as $n \to \infty$. Similarly one can show (as in Theorem 3.1) that

$$B_n/b_n = \frac{e_n}{\int_0^{e_n} 1\{Z(s) = 0\} ds} \left(\frac{\int_0^{e_n} 1\{Z(s) > 0\} ds}{e_n} - \frac{\int_0^{b_n} 1\{Z(s) > 0\} ds}{b_n} \right).$$

Thus $B_n/b_n \to 0$ as $n \to \infty$. Then (i) follows by noting that

$$C_n/b_n = B_n/b_n + I_n/b_n$$

$$= B_n/b_n + \frac{I_n/e_n}{1 - B_n/e_n},$$

which converges to 0 as $n \to \infty$. The proof of part (ii) is similar to that of Theorem 3.1 (ii). □

Remark 3.2. Theorem 3.2 remains valid if we replace the hypothesis that the conditions of Corollary 2.3 are satisfied by the condition that the limit

$$p(0) := \lim_{t \to \infty} t^{-1} \int_0^t 1\{Z(s) = 0\} ds$$

is well defined and $0 < p(0) \leq 1$. This is a specialization of the condition (20) given in Remark 3.1.

4. Applications to multiserver queues

In this section we give several applications to show that the sufficient conditions for stability given in Theorem 2.2 can be verified from conditions on primary quantities, that is, the arrival processes, number of servers, and server requirements in the system model. Several special cases are given, in which we establish *a priori* sufficient conditions for stability and other fundamental relationships for multiserver queues. In all applications we assume that servers are kept busy whenever possible. We also assume a work-conserving queue discipline.

The $G/G/c$ queue with heterogeneous servers is defined by the sequence $\{(A_n, S_n), n \geq 1\}$, where A_n is the time between the $n-1^{st}$ and n^{th} arrivals and S_n is the service requirement of the n^{th} arrival. (Customers need not be served in order of arrival but a server is never idle when customers are waiting.) It is also assumed that server i works at nonnegative rate (speed) δ_i, with $\delta = \sum_{i=1}^{c} \delta_i$, $0 < \delta < \infty$. Let $N(t) = max\{n: \sum_{k=1}^{n} A_k \leq t\}$ denote the number of customers that arrive in $[0, t]$.

4.1 The workload in a $G/G/c/\infty$ queue with heterogeneous servers

In this special case we have $\{Z(t), t \geq 0\} \equiv (W(t), t \geq 0\}$, and $A(t) = \sum_{k=1}^{N(t)} S_k$. Here $\{W(t), t \geq 0\}$ is the workload process. The i^{th} output process is given by $D_i(t) = \int_0^t \delta_i B_i(s) ds$, for all $t \geq 0$. Thus we have

$$W(t) = \sum_{k=1}^{N(t)} S_k - \int_0^t \sum_{i=1}^{c} \delta_i B_i(s) ds. \tag{24}$$

Note that under these conditions the limit in (7) exists and is equal to 1.

Theorem 4.1. *Consider the workload process* $\{W(t), t \geq 0\}$ *in the multiserver queues described by* (24). *Suppose*

(*i*) $\lim_{t \to \infty} t^{-1} N(t) = \lambda$, *where* $0 \leq \lambda < \infty$;

(*ii*) $\lim_{n \to \infty} n^{-1} \sum_{k=1}^{n} S_k = S$, *where* $0 \leq S < \infty$,

where $\alpha := \lambda S < \delta := \sum_{i=1}^{c} \delta_i$. *Then the workload process* $\{W(t), t \geq 0\}$ *is rate stable.*

Proof: Note that conditions (6) and (7) of Theorem 2.2 are satisfied. To verify condition (8), consider a sequence $\{t_n, n \geq 1\}$ such that $B(t_n) \leq c - 1$, for all $n \geq 1$. As in the proof of Theorem 2.2, we an assume $Z(0) = 0$ and $\alpha > 0$ without loss of generality. Then

$$Z(t_n)/t_n \leq (c-1) t_n^{-1} \max_{1 \leq k \leq N(t_n)} S_k. \tag{25}$$

It follows from assumption (*ii*) that $n^{-1} S_n \to 0$ and hence $n^{-1} \max_{1 \leq k \leq n} S_k \to 0$ as $n \to \infty$ (cf. Lemma 2.1 of Guillemin and Mazumdar [7], Lemma 15 of Serfozo [11]). Therefore, taking limits in (25) as $n \to \infty$ and appealing to (*i*) establishes the condition (8). Theorem 2.2 then implies that $\{W(t), t \geq 0\}$ is rate stable. □

Remark 4.1. Note that batch arrivals and/or bulk service are allowed, provided the batch

and/or bulk size is bounded. Theorem 4.1 extends results in [17], [6], [7], and [10]. Guillemin and Mazumdar [7] provide an alternate pathwise proof of Theorem 4.1 in the special case when the servers are homogeneous and the discipline is $FCFS$.

Corollary 4.2. *Suppose that the hypotheses of Theorem 4.1 hold. Then,*

$$\lim_{t\to\infty} t^{-1}\sum_{i=1}^{c}\delta_i\int_0^t (1-B_i(s))ds = \delta-\lambda. \tag{26}$$

Proof: The proof follows immediately from Corollary 2.3 and Theorem 4.1. □

Remark 4.2. Corollary 4.2 gives a set of conditions under which

$$0<\lim_{t\to\infty} t^{-1}\sum_{i1}^{c}\delta_i\int_0^t (1-B_i(s))ds \le \delta.$$

Consequently, it follows from Remark 3.1 following Theorem 3.1 that the busy period durations and the inputs during busy periods are rate stable for this queueing system.

The limit in (26) can be interpreted as the long-run weighted average number of idle servers. In the case where $\delta_i = 1$, $i = 1,2,\ldots,c$ (i.e., all servers work at unit rate on the average), this interpretation is more familiar. In the single-server queue, when $\lambda S < 1$, (26) reduces to $p(0) = 1-\rho$, where $\rho := \lambda S$.

It is worth noting that in relation (26), service rates can be time as well as state dependent. Only the long-run average rates are required to be constant. This is important in applications since, quite often, servers (human servers in particular) work at varying services rates.

We have provided a pure sample-path proof of (26) making assumptions only on "primary" quantities, thus generalizing the result for a $G/G/1$ queue in Stidham and El-Taha [17], Guillemin et al. [6], and Mazumdar et al. [10].

4.2 Number of customers in a $G/G/c/\infty$ queue with heterogeneous servers

In this special case $\{Z(t), t\ge 0\} \equiv \{L(t), t\ge 0\}$ is the queue-length process. That is, $L(t)$ is the number of customers in the system (including customers in service). Customers need not be served in order of arrival, but a server is never idle when customers are waiting. Let $A(t) \equiv N(t)$ denote the number of customers that arrive and $D(t) = \sum_{i=1}^{c}\int_0^t B_i(s)dD_i(s)$ the number that depart in $[0,t]$. Then $Z(t) = L(t)$ defined by (1) is the number of customers in the system at time t.

Theorem 4.3. *Consider the multiserver $G/G/c$ queue described in this subsection. Suppose*

$$\lim_{n\to\infty} n^{-1}\sum_{k=1}^{n} A_k = \lambda^{-1}, \quad 0\le\lambda<\infty,$$

$$\lim_{n\to\infty} n^{-1}\sum_{k=1}^{n} S_k = S, \quad 0\le S<\infty,$$

and that $\lambda S < \delta$. Then the queue-length process $\{L(t), t\ge 0\}$ is rate stable.

Proof: First note that by Corollary 4.2, the hypotheses of the theorem imply that

$$0<\lim_{t\to\infty} t^{-1}\sum_{i=1}^{c}(\delta_i/\delta)\int_0^t (1-B_i(s))ds \le 1.$$

(Recall that we used the analysis of the workload process to establish this.) Therefore $b_n \to \infty$ as $n \to \infty$. Moreover, by choosing $N(t) = A(t)$ in Theorem 3.1 and appealing to Remark 3.1, we obtain,

$$(N(e_n) - N(b_n))/b_n \to 0 \text{ as } n \to \infty,$$

that is, the number of arrivals in successive full busy periods is rate stable. Now let

$$b_t = \sum_{n=1}^{\infty} b_n 1\{b_n \le t < b_{n+1}\},$$

$$e_t = \sum_{n=1}^{\infty} e_n 1\{b_n \le t < b_{n+1}\},$$

so that b_t and e_t are the beginning and end, respectively, of the full busy period containing t if t falls in a full busy period. If t does not fall in a full busy period, then b_t and e_t are the beginning and end of the full busy period preceding t. Now

$$\begin{aligned} t^{-1}L(t) &\le (N(e_t) - N(b_t) + c - 1)/t \\ &\le (N(e_t) - N(b_t))/b_t + (c-1)/t, \end{aligned}$$

which $\to 0$ as $t \to \infty$. □

Batch arrivals and/or bulk service are permitted in this model by setting the appropriate A_n and S_n equal to zero. Note that the proof of Theorem 4.3 did not require the existence of the limit in (7). The existence of this limit can be shown to follow as a consequence of the theorem.

Corollary 4.4. *Under the conditions of Theorem 4.3, the limit in* (7) *exists, that is*

$$\lim_{t\to\infty} \frac{\sum_{i=1}^{c} \int_0^t \delta_i B_i(s) dD_i(s)}{\sum_{i=1}^{c} \int_0^t \delta_i B_i(s) ds} = 1.$$

Proof: Without loss of generality let $S = 1$. With $Z(t) = L(t)$ and

$$U(t) = \textstyle\sum_{i=1}^{c} \int_0^t \delta_i B_i(s) ds,$$

rewrite (1) as

$$t^{-1}L(t) = t^{-1}N(t) - t^{-1}U(t)[U(t)^{-1}D(t)]$$

then take limits as $t \to \infty$. By (15) and Lemma 4.3, $U(t)/t \to \lambda$ as $t \to \infty$. Thus $U(t)^{-1}D(t) \to 1$ as $t \to \infty$. □

4.3 Attained service process

Let $\{C(t), t \ge 0\}$ be a process such that $C(t)$ represents the cumulative service received by customers present at time t, that is

$$C(t) = \sum_{i=1}^{c} \int_0^t \delta_i B_i(s) ds - \sum_{k:\, D_k \le t} S_k. \tag{27}$$

This process is obviously rate stable for a $FCFS$ discipline. In the following corollary we show

that under the conditions of Theorem 4.3 and the condition that once a customer begins service its service cannot be interrupted, the process $\{C(t), t \geq 0\}$ is rate stable (i.e., $C(t) = o(t)$ as $t \to \infty$). We also show that the long-run average service time of departures coincides with that of arrivals under the conditions of Theorem 4.3.

Corollary 4.5. *Under the conditions of Theorem 4.3, assuming that each server serves one customer at a time with no preemption,*

(*i*) *the process* $\{C(t), t \geq 0\}$ *is rate stable; and*
(*ii*) $D(t)^{-1} \sum_{k: D_k \leq t} S_k \to S$ *as* $t \to \infty$.

Proof: Since $C(t) \leq c \max_{1 \leq k \leq N(t)} S_k$ and $S < \infty$, it is clear that

$$C(t)/t \leq c(N(t)/t) \max_{1 \leq k \leq N(t)} S_k/N(t).$$

Part (*i*) follows by taking limits as $t \to \infty$, and noting that $\max_{1 \leq k \leq N(t)} S_k/N(t) \to 0$ as $t \to \infty$. Suppose $S = 1$. Let $U(t) = \sum_{i=1}^{c} \int_0^t \delta_i B_i(s) ds$, and rewrite (27) as

$$D(t)^{-1} \sum_{k: D_k \leq t} S_k = D(t)^{-1} U(t) - [t^{-1} C(t)]/[t^{-1} D(t)].$$

Part (*ii*) follows by taking limits as $t \to \infty$, and appealing to part (*i*) and Corollary 4.4. If $S \neq 1$, let $\delta_i = \phi_i S$, $i = 1, \ldots, c$, and repeat the above argument. □

Part (*ii*) of Corollary 4.5 says that the symptotic mean service time of departing customers is equal to the asymptotic mean service time of arriving customers, for all stable queues with non-preemptive service discipline.

Remark 4.3. In Corollary 4.5 the assumption that each server serves one customer at a time with no preemption may be replaced by bounded service requirements. Moreover, this assumption can be removed in single-server queues because now $C(t)$ is bounded above by the duration of the busy cycle containing t.

Remark 4.4. The examples given in Stidham and El-Taha [17] for single-server queues can be easily extended to the multiserver case, using our results in this paper. Applications with output rates that are time dependent or state dependent or with rates that depends on other auxiliary processes can be established by appealing to Corollary 2.4.

4.4 Counterexample

The assertion $\lambda S < 1$ has been established as a sufficient condition for stability of the workload and queue-length processes in $G/G/1$ queues. It is worth noting that the workload process exhibits a stronger form of stability than the queue-length process in the sense that it holds under weaker conditions. We provide a counterexample to confirm this assertion. Let

$$\rho_t := t^{-1} \sum_{k=1}^{N(t)} S_k = (N(t)/t) N(t)^{-1} \sum_{k=1}^{N(t)} S_k.$$

Suppose that $\limsup_{t \to \infty} \rho_t < 1$. If either $\lim_{t \to \infty} N(t)/t$ or $\lim_{n \to \infty} n^{-1} \sum_{k=1}^{n} S_k$ does not exist, the queue-length process is not necessarily rate stable, as can be seen from the following example.

Consider a $G/G/1$-$LCFS$-PR queue, where the service requirement of the customer arriving at time t (denoted by S_t) is given by

$$S_t = \begin{cases} 1 + 1/(2^{2k-1}) & t = 2^{2k-1}, \ldots, 2^{2k} - 1; k = 1, 2, \ldots; \\ 1 & t = 2^{2k} + 2i;\ i = 1, \ldots, 2^{2k-1} - 1;\ k = 1, 2, \ldots. \end{cases}$$

Arrivals occur at those time instances where $S_t > 0$.

In this example it is clear that $S = 1$, and

$$2/3 = \liminf_{t \to \infty} \rho_t < \limsup_{t \to \infty} \rho_t = 5/6 < 1.$$

Simple manipulations show that the workload process $\{W(t), t \geq 0\}$ is *rate stable*, (i.e., $t^{-1}W(t) \to 0$ as $t \to \infty$), even though the limit of ρ_t does not exist. However, the queue-length process $\{L(t), t \geq 0\}$ is not rate stable:

$$0 = \liminf_{t \to \infty} t^{-1} L(t) < \limsup_{t \to \infty} t^{-1} L(t) = 1/2.$$

It is interesting to note that the attained-service process, $\{C(t), t \geq 0\}$, also is not rate stable:

$$0 = \liminf_{t \to \infty} t^{-1} C(t) < \limsup_{t \to \infty} t^{-1} C(t) = 1/2.$$

5. Little's Formula for Stable Queues

In this section we review and extend previous research on sample-path proofs of Little's Formula ($L = \lambda W$), focusing in particular on establishing $L = \lambda W$ in stable queueing systems with minimal *a prior* assumptions about existence and/or finiteness of the averages involved. The material in this section is based in part on Stidham [15].

In the spirit of Little [8] and Stidham [12], [13], we consider a general input-output system, fed by a discrete input process of *customers*, each of which spends a certain amount of time in the system, and then departs. We follow Whitt [18] in the problem setup and notation. The basic data are $\{(T_k, D_k), k \geq 1\}$, where $0 \leq T_k \leq T_{k+1} < \infty$, $T_k \leq D_k < \infty$, $k \geq 1$, and T_k and D_k are interpreted as the arrival time and the departure time, respectively, of customer k. We assume that $T_k \to \infty$, as $k \to \infty$, so that there are only a finite number of arrivals in any finite time interval. Let $N(t) := \#\{k: T_k \leq t\}$, $D(t) := \#\{k: D_k \leq t\}$, $t \geq 0$, so that $N(t)$ and $D(t)$ count the number of arrivals and departures, respectively, in the interval $[0, t]$. Note that, since $T_k < \infty$ for all $k \geq 1$, $N(t) \to \infty$ as $t \to \infty$. Note also that $N(t) = max\{k: T_k \leq t\}$, since $\{T_k, k \geq 1\}$ is a nondecreasing sequence. But in general we cannot write $D(t) = max\{k: D_k \leq t\}$, because $\{D_k, k \geq 1\}$ is not necessarily nondecreasing. (It is nondecreasing if the discipline is *first-in, first-out* ($FIFO$), that is, if departures occur in the same order as arrivals.) Define

$$L(t) := \#\{k: T_k \leq t < D_k\} = N(t) - D(t), \quad t \geq 0, \tag{28}$$

$$W_k := D_k - T_k, \quad k \geq 1, \tag{29}$$

so that $L(t)$ is the number of customers in the system at time t and W_k is the waiting time in the system of customer k.

If we let $I_k(t)$ denote the indicator of $T_k \leq t < D_k$, then it follows from (28) and (29) that

$$L(t) = \sum_{k=1}^{\infty} I_k(t)$$

$$W_k = \int_0^{\infty} I_k(t)dt$$

from which we obtain the basic inequality

$$\sum_{k:T_k \le t} W_k \ge \int_0^t L(s)ds \ge \sum_{k:D_k \le t} W_k, \quad t \ge 0. \tag{30}$$

Our proofs of $L = \lambda W$ use the basic inequality (30) and the following elementary lemmas (cf. Stidham [13], Stidham and El-Taha [16], El-Taha and Stidham [5]). Let $\{T_k, k \ge 1\}$ be a point process, with $0 \le T_k \le T_{k+1} < \infty$, $k \ge 1$, $T_k \to \infty$ as $k \to \infty$, and $N(t) = max\{k: T_k \le t\}$, $t \ge 0$.

Lemma 5.1. *Let $0 \le \lambda \le \infty$. Then $t^{-1}N(t) \to \lambda$ as $t \to \infty$ if and only if $k^{-1}T_k \to \lambda^{-1}$ as $k \to \infty$.*

Let $\{X_k, k \ge 1\}$ be a sequence of nonnegative numbers and define $Y(t)$: $= \sum_{k:T_k \le t} X_k$, $t \ge 0$.

Lemma 5.2. *Suppose $t^{-1}N(t) \to \lambda$ as $t \to \infty$, where $0 \le \lambda \le \infty$. Then*

(*i*) *if $n^{-1}\sum_{k=1}^{n} X_k \to X$ as $n \to \infty$, where $0 \le X \le \infty$, then $t^{-1}Y(t) \to Y = \lambda X$ as $t \to \infty$, provided that λX is well defined;*

(*ii*) *if $t^{-1}Y(t) \to Y$ as $t \to \infty$, where $0 \le Y \le \infty$, then $n^{-1}\sum_{k=1}^{n} X_k \to X = \lambda^{-1}Y$ as $n \to \infty$, provided that $\lambda^{-1}Y$ is well defined.*

We shall also need the following lemma (cf. Lemma 2.1 of Stidham [14]).

Lemma 5.3. *Suppose $W_n/T_n \to 0$ as $n \to \infty$. Let $0 \le L \le \infty$. Then the following are equivalent:*

$$\lim_{t\to\infty} t^{-1} \sum_{k:T_k \le t} W_k = L \tag{31}$$

$$\lim_{t\to\infty} t^{-1} \int_0^t L(s)ds = L \tag{32}$$

$$\lim_{t\to\infty} t^{-1} \sum_{k:D_k \le t} W_k = L. \tag{33}$$

Proof: Let $\epsilon > 0$ be given. Since $W_n/T_n \to 0$ as $n \to \infty$, there exists an integer N such that $k \ge N$ implies $W_k \le T_k\epsilon$. Therefore, for all $t \ge 0$,

$$\sum_{k:D_k \le t} W_k = \sum_{k:T_k + W_k \le t} W_k$$

$$\ge \sum_{k \ge N:T_k(1+\epsilon) \le t} W_k$$

$$\geq \sum_{k: T_k(1+\epsilon) \leq t} W_k - \sum_{k \leq N-1} W_k,$$

which, together with the basic inequality (30), implies

$$\sum_{k: T_k \leq t} W_k \geq \int_0^t L(s)ds \tag{34}$$

$$\geq \sum_{k: D_k \leq t} W_k \tag{35}$$

$$\geq \sum_{k: T_k(1+\epsilon) \leq t} W_k - \sum_{k \leq N-1} W_k. \tag{36}$$

First we use these inequalities to prove that (31) implies (32) and (33). Suppose (31) hold. Then

$$\lim_{t\to\infty} t^{-1} \sum_{k: T_k(1+\epsilon) \leq t} W_k = (1+\epsilon)^{-1} \lim_{t\to\infty} [t(1+\epsilon)^{-1}]^{-1} \sum_{k: T_k \leq t(1+\epsilon)^{-1}} W_k$$

$$= (1+\epsilon)^{-1} \lim_{t\to\infty} t^{-1} \sum_{k: T_k \leq t} W_k$$

$$= (1+\epsilon)^{-1} L.$$

But $\epsilon > 0$ was arbitrary. Hence the desired result follows from this equation and the inequalities (34), (35), and (36), using the fact that $\lim_{t\to\infty} t^{-1} \sum_{k \leq N-1} W_k = 0$.

Now we show that (32) implies (31). (The proof that (33) implies (31) is similar.) Suppose (32) holds. Then (34), (35), and (36) imply that

$$\liminf_{t\to\infty} t^{-1} \sum_{k: T_k \leq t} W_k \geq L$$

$$\geq \limsup_{t\to\infty} t^{-1} \sum_{k: T_k(1+\epsilon) \leq t} W_k$$

$$= (1+\epsilon)^{-1} \limsup_{t\to\infty} [t(1+\epsilon)^{-1}]^{-1} \sum_{k: T_k \leq t(1+\epsilon)^{-1}} W_k$$

$$= (1+\epsilon)^{-1} \limsup_{t\to\infty} t^{-1} \sum_{k: T_k \leq t} W_k$$

where we have used the fact the $\lim_{t\to\infty} t^{-1} \sum_{k \leq N-1} W_k = 0$ in the derivation of the second inequality. Since $\epsilon > 0$ was arbitrary, we conclude that these inequalities hold in the limit as $\epsilon \to 0$, so that (31) holds.

This completes the proof of Lemma 5.3. □

The following theorem is an immediate consequence of Lemmas 5.2 and 5.3.

Theorem 5.4. *Suppose* $t^{-1}N(t) \to \lambda$ *as* $t \to \infty$, *where* $0 \leq \lambda \leq \infty$, *and* $W_n/T_n \to 0$ *as* $n \to \infty$. *Then*

(*i*) *if* $n^{-1}\sum_{k=1}^{n} W_k \to W$ *as* $n \to \infty$, *where* $0 \leq W \leq \infty$, *then* $t^{-1}\int_0^t L(s)ds \to L$ *as* $t \to \infty$, *and* $L = \lambda W$, *provided* λW *is well defined;*

(*ii*) *if* $t^{-1}\int_0^t L(s)ds \to L$ *as* $t \to \infty$, *where* $0 \leq L \leq \infty$, *then* $n^{-1}\sum_{k=1}^{n} W_k \to W$ *as* $n \to \infty$, *and* $L = \lambda W$, *provided* $\lambda^{-1}L$ *is well defined.*

Theorem 5.4 has the following immediate corollary.

Corollary 5.5. *Suppose* $t^{-1}N(t) \to \lambda$ *as* $t \to \infty$, *where* $0 \leq \lambda < \infty$, *and* $n^{-1}\sum_{k=1}^{n} W_k \to W$, *as* $n \to \infty$, *where* $0 \leq W \leq \infty$. *Suppose also that* $n^{-1}W_n \to 0$ *as* $n \to \infty$. *Then* $t^{-1}\int_0^t L(s)ds \to L$ *as* $t \to \infty$ *and* $L = \lambda W$.

Proof: It follows from Lemma 5.1 that $n/T_n \to \lambda < \infty$, and hence $W_n/T_n \to 0$ as $n \to \infty$. The desired result then follows from Theorem 5.4. □

This corollary may be useful in cases where $W = \infty$, but $n^{-1}W_n \to 0$, as is the case, for example, in a stable $GI/GI/1$ queue in which the second moment of service time is infinite. It also leads immediately to the next corollary, which is the original sample-path version of $L = \lambda W$ contained in Stidham [12], [13].

Corollary 5.6. *Suppose* $t^{-1}N(t) \to \lambda$ *as* $t \to \infty$, *where* $0 \leq \lambda < \infty$, *and* $n^{-1}\sum_{k=1}^{n} W_k \to W$ *as* $n \to \infty$, *where* $0 \leq W < \infty$. *Then* $t^{-1}\int_0^t L(s)ds \to L$, *as* $t \to \infty$ *and* $L = \lambda W$.

Proof: Since by hypothesis $W < \infty$, it follows that $n^{-1}W_n \to 0$. The result then follows from Corollary 5.5. □

The condition that $W_n/T_n \to 0$ as $n \to \infty$ cannot be verified directly from input data such as interarrival and service times. In the following subsections, however, we show that the *rate stability* conditions established in Section 3 for single and multiserver queues are sufficient to verify that the above condition for Little's formula holds.

5.1 The single-server case

The input data for the $G/G/1$ queue consists of the sequence $\{(T_n, S_n),\ n \geq 1\}$, where T_n is the arrival instant and S_n the work requirement of customer n, $n \geq 1$. In the special case where $\{Z(t), t \geq 0\}$ is the workload process in the $G/G/1$ queue, we have $A(t) = \sum_{k=1}^{N(t)} S_k$ and $D(t) = \int_0^t \delta 1\{Z(s) > 0\}ds$, for all $t \geq 0$. Here $\{N(t), t \geq 0\}$ is the point process which counts the number of customer arrivals: $N(t) = max\{n: T_n \leq t\}$, $t \geq 0$.

Our goal is to prove that $L = \lambda W$ under minimal stability conditions on the input process, $\{A(t), t \geq 0\}$.

Theorem 5.7. *Suppose* $t^{-1}N(t) \to \lambda$ *as* $t \to \infty$, *and* $n^{-1}\sum_{k=1}^{n} S_k \to S$ *as* $n \to \infty$, *where* $0 < \alpha := \lambda S < \delta$. *Then* $W_n/T_n \to 0$ *as* $n \to \infty$ *and hence* (*i*) *and* (*ii*) *of Theorem 5.4 hold.*

Proof: Let $b(n) := \sum_{k=1}^{\infty} b_k 1\{b_k \leq T_n < e_k\}$ and $B(n) := \sum_{k=1}^{\infty}(e_k - b_k)1\{b_k \leq T_n < e_k\}$. That is, $b(n)$ and $B(n)$ are the beginning and the duration, respectively, of the busy period that corresponds to the n^{th} arrival. It follows from Theorem 3.2 that

$$T_n^{-1}W_n \leq T_n^{-1}B(n) \leq b(n)^{-1}B(n) \to 0 \text{ as } n \to \infty.$$

The desired result then follows by appealing to Theorem 5.4. □

5.2 The multiserver case

In the special case where $\{Z(t), t \geq 0\}$ is the workload process in a $G/G/c$ queue, we have $A(t) = \sum_{k=1}^{N(t)} S_k$ and $D(t) = \sum_{i=1}^{c} \int_0^t \delta_i B_i(s)ds$, for all $t \geq 0$. Here again $\{N(t), t \geq 0\}$ is the point process which counts the number of customer arrivals and $\{S_k, k \geq 1\}$ is the sequence of work requirements of the arriving customers.

Unlike the single-server case, in multiserver queues we need the additional assumption that the queue discipline is work conserving and nonpreemptive. Let D'_k be the time instant the k^{th} arrival departs the queue (i.e., joins service). If we let $I_k(t)$ denote the indicator of $\{T_k \leq t < D'_k\}$, then it follows from (28) and (29) that

$$L_q(t) = \sum_{k=1}^{\infty} I_k(t)$$

$$W_q(k) = \int_0^{\infty} I_k(t)dt$$

where $L_q(t)$ is the number of customers in the queue at time t and $W_q(k)$ is the delay (time in the queue) of the k^{th} arrival.

Theorem 5.8. *Assume the queue discipline is work conserving and nonpreemptive. Suppose* $t^{-1}N(t) \to \lambda$ *as* $t \to \infty$, *and* $n^{-1}\sum_{k=1}^{n} S_k \to S$ *as* $n \to \infty$, *where* $0 < \alpha := \lambda S < \delta$. *Then*

(*i*) *if* $n^{-1}\sum_{k=1}^{n} W_q(k) \to W_q$ *as* $n \to \infty$, *where* $0 \leq W_q \leq \infty$, *then* $t^{-1}\int_0^t L_q(s)ds \to L_q$ *as* $t \to \infty$, *and* $L_q = \lambda W_q$, *provided* λW_q *is well defined;*

(*ii*) *if* $t^{-1}\int_0^t L_q(s)ds \to L_q$ *as* $t \to \infty$, *where* $0 \leq L_q \leq \infty$, *then* $n^{-1}\sum_{k=1}^{n} W_q(k) \to W_q$ *as* $n \to \infty$, *and* $L_q = \lambda W_q$, *provided* $\lambda^{-1}L_q$ *is well defined.*

Proof: Let $Z(t) = \sum_{k=1}^{N(t)} S_k - \sum_{i=1}^{c} \int_0^t \delta_i B_i(s)ds$ be the workload in the system at time t, $t \geq 0$. Let $b(n) := \sum_{k=1}^{\infty} b_k 1\{b_k \leq T_n < e_k\}$ and $B(n) := \sum_{k=1}^{\infty} (e_k - b_k)1\{b_k \leq T_n < e_k\}$. That is, $b(n)$ and $B(n)$ are the beginning and the duration, respectively, of the full busy period that corresponds to the n^{th} arrival. The conditions of Corollary 2.3 are satisfied. Therefore, by Theorem 3.1,

$$T_n^{-1} W_q(n) \leq T_n^{-1} B(n) \leq b(n)^{-1} B(n) \to 0 \text{ as } n \to \infty.$$

Thus the result follows by using Theorem 5.4. □

Corollary 5.9. *Suppose the conditions of Theorem 5.8 hold. Then*

(*i*) *if* $n^{-1}\sum_{k=1}^{n} W_k \to W$ *as* $n \to \infty$, *where* $0 \leq W \leq \infty$, *then* $t^{-1}\int_0^t L(s)ds \to L$ *as* $t \to \infty$, *and* $L = \lambda W$, *provided* λW *is well defined;*

(*ii*) *if* $t^{-1}\int_0^t L(s)ds \to L$ *as* $t \to \infty$, *where* $0 \leq L \leq \infty$, *then* $n^{-1}\sum_{k=1}^{n} W_k \to W$ *as* $n \to \infty$, *and* $L = \lambda W$, *provided* $\lambda^{-1}L$ *is well defined.*

Proof: Note that $W_k \leq W_q(k) + S_k / min_i \delta_i$, so that $W_k / T_k \to 0$ as $k \to \infty$, since $min_i \delta_i > 0$ and $S_k / k \to 0$ as $k \to \infty$. □

The above theorem is valid under rather weak conditions, for example, servers can be hetero-

geneous as well as homogeneous. Furthermore, although a customer cannot be preempted once in service, it may be switched form a fast to a slow server or vice versa.

We note also that in the stable versions fo $L = \lambda W$ we did not require the assumption that the sequence of departure times be finite.

References

[1] Altman, E., Foss, S.G., Riehl, E.R., and Stidham, S., Performance Bounds and Pathwise Stability for Generalized Vacation and Polling Systems, Technical Report **UNC/OR TR93-8**, Department of Operations Research, University of North Carolina at Chapel Hill 1993.

[2] Baccelli, F. and Brémaud, P., *Palm Probabilities and Stationary Queues*, Lecture Notes in Statistics **41**, Springer-Verlag, New York 1987.

[3] El-Taha, M. and Stidham, Jr., S., Sample-path analysis of stochastic discrete-event systems, *Proc. of the 30th IEEE CDC Meeting* 1991.

[4] El-Taha, M. and Stidham, Jr., S., Deterministic analysis of queueing systems with hetero geneous servers, *Theoretical Computer Science* **106** (1992), 243-264.

[5] El-Taha, M. and Stidham, Jr., S., Sample-path analysis of stochastic discrete-event systems, *Discrete Event Dynamic Systems: Theory and Appl.* **3** (1993), 325-346.

[6] Guillemin, F., Badrinath, V. and Mazumdar, R., Les techniques trajectorielles appliquées aux files d'attente non-stationnaires, submitted for publication 1991.

[7] Guillemin, F. and Mazumdar, R., On pathwise behavior of multiserver queues, *Queueing Systems: Theory and Appl.* (1993), 000-000.

[8] Little, J.D.C., A proof for the queueing formula $L = \lambda W$, *Operations Research* (1961), 383-387.

[9] Mazumdar, R., Badrinath, V., Guillemin, F. and Rosenberg, C., Pathwise rate conservation and queueing applications, submitted for publication 1991.

[10] Mazumdar, R., Guillemin, F., Badrinath, V. and Kannurpatti, R., On pathwise behavior of queues, *Operations Research Letters* **12** (1992), 263-270.

[11] Serfozo, R.F., More about Little laws for waiting times and utility processes, School of Industrial and Syst. Eng., Georgia Institute of Technology 1993.

[12] Stidham, Jr., S., $L = \lambda W$: A discounted analogue and a new proof, *Operations Research* **20** (1972), 708-732.

[13] Stidham, Jr., S., A last word on $L = \lambda W$, *Operations Research* **22** (1974), 417-421.

[14] Stidham, Jr., S., Sample-path analysis of queues, In: *Applied Prob. and Comp. Science: The Interface* (ed. by R. Disney and T. Ott), Birkhauser, Boston 1982, 41-70.

[15] Stidham, Jr., S., A comparison of sample-path proofs of $L = \lambda W$, preprint, INRIA, Sophia, Antipolis 1991.

[16] Stidham, Jr., S. and El-Taha, M., Sample-path analysis of processes with imbedded point processes, *Queueing Systems: Theory and Appl.* **5** (1989), 131-165.

[17] Stidham, Jr., S. and El-Taha, M., A note on sample-path stability conditions for input-output processes, *Operations Research Letters* **14** (1993), 1-7.

[18] Whitt, W., A review of $L = \lambda W$ and extensions, *Queueing Systems: Theory and Appl.* **9** (1991), 235-268.