COMMUN. STATIST.—STOCHASTIC MODELS, 8(4), 685-717 (1992)

THE IMPACT OF A JOB BUFFER IN A TOKEN-BANK RATE-CONTROL THROTTLE

by

Arthur W. Berger

AT&T Bell Laboratories
Holmdel, NJ 07733

Ward Whitt

AT&T Bell Laboratories
Murray Hill, NJ 07974-0636

Key Words: overload control, rate control throttles, sample-path comparisons

ABSTRACT

In this paper we study a rate-control throttle with a finite-capacity token bank and a finite-capacity job buffer. The primary purpose is to gain additional insight into the impact of the job buffer. We show that the overflow processes of jobs and tokens depend on the job-buffer and token-bank capacities only through their sum, in a very strong sense. Given two throttles with arbitrary token and job arrival processes, which differ only in their initial conditions and buffer capacities, having common total capacity, there exists a random time after which the overflow processes in these two systems coincide. For given total capacity, the job buffer smooths the stream of admitted jobs, but the reduced congestion is less than might be expected. For example, the heavy-traffic limiting behavior of a downstream infinite-capacity s-server queue is unaffected by the job buffer in the throttle. We make a sample-path comparison of the throughputs at a downstream finite-capacity queue regulated by a token-bank rate-control throttle, with and without a job buffer. Given a fixed total capacity in the throttle (and thus a fixed admission rate of the throttle) and given a fixed amount of buffer space for jobs to allocate to a downstream queue and a job buffer in the throttle, the maximum throughput of jobs occurs when all the buffer capacity is allocated to the downstream queue, even though the admitted stream from the throttle is not smoothed by a job buffer. Similar results hold for systems with non-discrete flow, such as regulated Brownian motion and Markov modulated fluid models.

1. Introduction and Summary

A token-bank rate-control throttle is a rate-based input-regulation technique for congestion control. The token bank is typically a counter which increments periodically and decrements at job arrivals. Conceptually, we can think of jobs and tokens arriving in separate streams, as depicted in Figure 1, with the arrival stream of tokens typically being deterministic and evenly spaced, although we do not restrict attention to this case. An arriving job requires a token to be admitted; it is blocked and rejected (or marked, admitted and treated as a lower priority class) if there are no tokens in the token bank. Arriving tokens are put in the token bank if there is room; otherwise, they are lost. Sidi et al. [15] and Berger [1] proposed an expanded throttle in which jobs may queue in a finite buffer when the token bank is empty. Thus, with the expanded throttle there are two finite-capacity buffers, one for the tokens and one for the jobs, where at most one of the two is nonempty at any time; see §2 for a complete definition. Such modified two-buffer rate-control throttles have subsequently been considered by several authors, including Sohraby and Sidi [16], Budka and Yao [7] and Elwalid and Mitra [10].

The primary purpose of this paper is to gain additional insight into the impact of the job buffer on the performance of the throttle. For this purpose, we make several sample-path comparisons. Hence, this paper is in the same spirit as previous papers by Budka and Yao [7] and Budka [6] on sample-path comparisons for rate-control throttles, as well as Sonderman [17], [18], Whitt [21] and Cruz [8], [9] on sample-path comparisons for queueing models.

Concerning the impact of the job buffer, there are three main ideas discussed here. First, the two overflow processes essentially depend on the token-bank and job-buffer capacities only through their sum, which we call the total capacity of the throttle. This idea is a main point of Berger [1]; in §3 and §4 here we strengthen the conclusion by performing a sample-path analysis and treating general initial conditions.

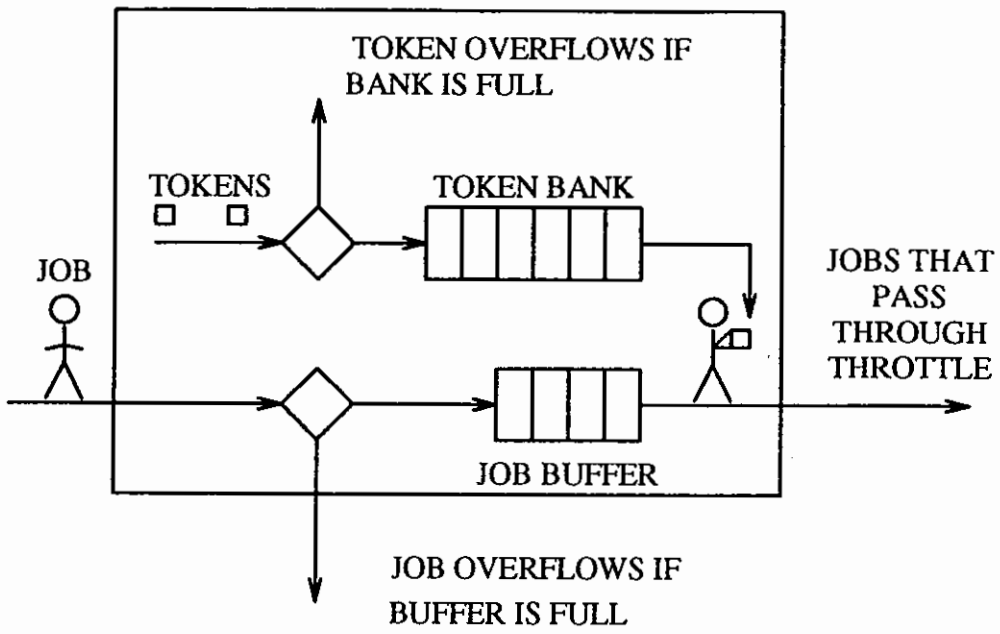

Figure 1. Diagram of rate-control throttle and downstream node.

The second idea is that (assuming that the total capacity of the throttle is fixed) the job buffer provides a benefit by smoothing the process of admitted jobs, but at the cost of introducing additional delays. This smoothing property is discussed in [15], [1], [16] and [10].

The third idea is that, while the job buffer does smooth the admitted stream to some extent, it does not reduce congestion at downstream queues as much as we might hope. This idea is advanced by Berger and Whitt [4], where it is supported by simulation. In [4] a model of LAN-to-LAN (local-area-network) traffic is regulated by a throttle with the admitted traffic being sent to a finite-capacity single-server deterministic queue. The simulation results in [4] show that the admitted stream is significantly smoothed in a short time scale but much less so in a longer time scale, as measured by the index of dispersion for intervals (IDI); see (5.6) below and [11].

We support these observations here in §5 by *proving* that the asymptotic behavior of the accepted job stream, e.g., the limiting value of the IDI, is

unchanged by the addition of the job buffer. Thus, by Theorem 1 of Iglehart and Whitt [14], the heavy-traffic limiting behavior of a downstream infinite-capacity finite-server queue is unaffected by the addition of the job buffer.

For some operating regimes of the throttle, the token arrival rate will be significantly greater than the job arrival rate and there will be very little blocking of jobs. Then the introduction of a sufficiently large job buffer in the throttle, with the total capacity of the throttle held fixed, amounts (approximately) to inserting an infinite-capacity queue before some downstream queue. Suppose that this downstream queue also has ample capacity, so that there is very little blocking; then the primary issue is delay. It is significant that a simple sample-path argument shows that the introduction of an infinite-capacity queue before another infinite-capacity queue always causes the time each job spends in the system to be greater than or equal to what it was before; see p. 358 of Suresh and Whitt [19]. (This analysis is for a single class, so that it does not touch on benefits with multiple classes.)

In §6 here we establish a related result for the rate-control throttle and a downstream finite-capacity queue. We assume that the successive service times at the downstream queue are assigned to jobs when they start service, but the job and token arrival processes can be arbitrary. We show that the blocking is always less at the downstream queue when the downstream queue has capacity $C_D + C_J$, the throttle has capacity C and there is no job buffer, than when the downstream queue has capacity C_D, the token bank has capacity $C - C_J$ and there is a job buffer with capacity C_J (assuming that $C_J \leq C$). That is, moving capacity C_J from the downstream queue to the job buffer, while holding the total capacity of the throttle fixed, decreases the throughput from the downstream queue. This means that the most efficient allocation of a fixed total buffer capacity for jobs is to assign it all to the downstream queue. (Recall that the token bank is not a buffer for jobs, and does not store a real resource; it is a counter.) In making this comparison, we include an extra arrival process to the downstream queue. This allows us to deduce a similar comparison for multiple throttles feeding a downstream queue.

To establish this comparison, we apply a result from Berger and Whitt [2], showing that more departures always come from a multi-server queue (with the first-come first-served discipline, a finite waiting room and service times assigned when service begins) when the service times are decreased and a second arrival process is added to the original arrival process. This superposition operation makes the original arrival epochs a subsequence of the new arrival epochs. Thus, the arrival processes are ordered in the subsequence stochastic ordering $\leq_2$ in Whitt [21], which is also used in Budka and Yao [7] and Budka [6] (with the appealing notation $\subseteq_{st}$).

As we pointed out, so far we have discussed the smoothing effect of the job buffer only for a single admitted stream coming to a downstream queue. Suppose, instead, that we have multiple streams of jobs feeding a downstream queue, each regulated by its own throttle. If the number of streams is increased, while simultaneously reducing the service time of the downstream queue, so that the server occupancy is unchanged, then the relevant time scale is reduced and, consistent with §5, the smoothing by a job buffer in the throttle can dramatically reduce the congestion in the downstream queue. The simulations in [4] show that the smoothing effect is much greater when many (e.g., 100) sources are multiplexed at the downstream queue (as typically occurs in communication network applications). However, consistent with §6 here, the simulation results in [4] also show that the total buffer capacity used in the job buffers of all the throttles is much greater than would be required at the downstream queue in order to achieve the same blocking rate.

We continue in §7 by discussing the impact of changing the total capacity of the throttle. We provide some extensions to Budka and Yao's [7] result that the throughput is an increasing concave function of the capacity. In §8 we consider the impact of changing arrival processes in a rate-control throttle with a job buffer, once again using the subsequence ordering.

Finally in §9 we point out that the results in this paper also hold for models with non-discrete flow, such as the Markov-modulated fluid model in Elwalid and

Mitra [10] and the reflected Brownian motion (RBM) in Harrison [13] and Berger and Whitt [3].

2. The General Throttle Model

Jobs and tokens arrive in separate streams. The arriving tokens are put in a *token bank* of capacity C_T, where $0 \le C_T < \infty$; if the token bank is full, then the tokens overflow and are lost. The arriving jobs are admitted immediately if there is a token in the token bank, with each admitted job taking away one token from the bank. If there is no token in the token bank, then the job is put in a *job buffer* of capacity C_J, where $0 \le C_J < \infty$. If the job buffer is full, then the job is not admitted and is said to have overflowed. (Overflowed jobs are lost, or are marked and admitted and later treated as a lower priority class. If marking is used and if job sequence is to be maintained, then the arriving job enters the full buffer while the job at the head of the buffer is marked and admitted.) Jobs in the job buffer are admitted upon subsequent token arrivals, with each job taking one token away. Hence, the admitted jobs are matched with admitted tokens.

The evolution of this throttle can be defined in terms of the two arrival processes recursively by considering successive arrival epochs (of jobs or tokens), as we show starting in (2.5) below. As discussed in §4.1 of Berger and Whitt [3], the evolution of this throttle can also be defined in terms of the two-sided regulator on pp. 21-24 of Harrison [13].

Let $A_J(t)$ and $A_T(t)$ count the number of job arrivals and token arrivals, respectively, in $(0, t]$. (We assume that $A_J(0) = A_T(0) = 0$.) It is customary to have $A_T(t)$ be deterministic, i.e., $A_T(t) = \lfloor rt \rfloor$, $t \ge 0$, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x, but we do not assume this. Let $J(t)$ and $T(t)$ represent the number of jobs in the job buffer and the number of tokens in the token bank at time t. As noted by Berger [1], since $T(t)$ and $J(t)$ cannot both be strictly positive at the same time, we can represent both processes simultaneously via $U(t)$, where

$$U(t) = T(t) - J(t) + C_J\,, \quad t \ge 0\,. \tag{2.1}$$

We obtain $T(t)$ and $J(t)$ from $U(t)$ by

$$T(t) = [U(t) - C_J]^+ , \; t \geq 0 , \tag{2.2}$$

and

$$J(t) = [C_J - U(t)]^+ , \; t \geq 0 , \tag{2.3}$$

where $[x]^+ = \max\{x, 0\}$.

Let $O_J(t)$ and $O_T(t)$ count the number of job and token overflows in $(0, t]$, respectively. Let $D(t)$ count the number of job departures (admitted jobs) in $(0, t]$, which coincides with the number of admitted tokens. We obviously have

$$\begin{aligned} D(t) &= A_J(t) - O_J(t) - J(t) + J(0) \\ &= A_T(t) - O_T(t) - T(t) + T(0) . \end{aligned} \tag{2.4}$$

The triple $(U(t), O_J(t), O_T(t))$ is the three-dimensional regulated process associated with net input process $A_T(t) - A_J(t)$ and reflecting barriers at 0 and $C \equiv C_J + C_T$, using the two-sided regulator on pp. 21-24 of Harrison [13]. To define the processes $U(t)$, $O_J(t)$ and $O_T(t)$ without directly applying the two-sided regulator, let

$$A(t) = A_T(t) + A_J(t) , \; t \geq 0 , \tag{2.5}$$

be the total arrival process, and let t_n be the n^{th} time at which $A(t)$ has a jump; i.e., let $t_0 = 0$ and let

$$t_n = \inf\{t > t_{n-1} : A(t) > A(t_{n-1})\} . \tag{2.6}$$

(We have assumed that $A_T(0) = A_J(0) = 0$. We also assume all processes are right continuous with left limits.) We must stipulate what happens with multiple arrivals at the same instant. We assume that jobs and tokens arriving simultaneously are immediately paired and admitted. Then the excess of jobs or tokens enters the system and is treated as specified above. (Other cases can be treated similarly.) Consequently, we can define the evolution of the throttle recursively by setting

$$U_n = U(t_n) \quad \text{and} \quad X_n = A_T(t_n) - A_T(t_{n-1}) - [A_J(t_n) - A_J(t_{n-1})] \ . \tag{2.7}$$

Then, for $n \geq 1$,

$$U_n = \min\{C_J + C_T \ , \max\{0, U_{n-1} + X_n\}\} \ , \tag{2.8}$$

$$O_T(t_n) = O_T(t_{n-1}) + [X_n - (U_n - U_{n-1})]^+ \ , \tag{2.9}$$

$$\text{and} \quad O_J(t_n) = O_J(t_{n-1}) + [X_n - (U_n - U_{n-1})]^- \ , \tag{2.10}$$

where $[x]^- = -\min\{x, 0\}$, $O_T(0) = O_J(0) = 0$ and $U_0 = U(0)$ is the initial condition. We assume that $0 \leq U(0) \leq C$. Then

$$U(t) = U(t_n), \ O_J(t) = O_J(t_n), \ \text{and} \ O_T(t) = O_T(t_n) \ , \ t_n \leq t < t_{n+1} \ . \tag{2.11}$$

Next $T(t)$ and $J(t)$ are obtained by combining (2.2), (2.3) and (2.11). Then $D(t)$ is obtained from (2.4). For a useful alternative defining recursion, see (8.2)–(8.4) below.

3. Insensitivity When $C_J + C_T$ is Fixed: Sample-Path Properties

In this section we begin to extend the result of Berger [1] stating that the overflow processes tend to be independent of the job-buffer capacity C_J provided that the total capacity $C_J + C_T$ remains unchanged. Toward this end, we discuss properties of the rate-control throttle that can be deduced for individual sample paths. It is significant that these results do not depend on any specific probabilistic structure. In particular, the job arrival process need not be an MAP arrival process as in [1].

First, we discuss an equivalence between systems with capacities C_J and C_T where $C = C_J + C_T$ is fixed when the initial conditions are properly related. This is a minor extension of Theorem 2 of [1]. Afterwards we will obtain results under different initial conditions.

Theorem 3.1. *The sample paths of $U(t)$, $O_J(t)$ and $O_T(t)$ remain unchanged if C_J is changed to C_J' provided that C_T is changed to $C_T' = C_T + C_J - C_J'$ and $U(0)$ is unchanged.*

Proof. The evolution of the throttle is defined by equations (2.5)–(2.11). From (2.8), we see that this evolution depends on C_J and C_T only through $C_J + C_T$ provided that $U(0)$ is unchanged. ■

Remark 3.1. Let a prime denote the new system with capacities C'_J and $C'_T = C_T + C_J - C'_J$ as in Theorem 3.1. Note that $U'(0) = U(0)$ corresponds to a *change* in the initial conditions for $T(t)$ and $J(t)$, as can be seen from (2.1). In particular,

$$U(0) = T(0) - J(0) + C_J \quad \text{and} \quad U'(0) = T'(0) - J'(0) + C'_J ,$$

so that

$$T'(0) - J'(0) = T(0) - J(0) + C_J - C'_J . \quad ■ \tag{3.1}$$

We now consider the effect of the initial condition $U(0)$, with everything else fixed. The essence of the following result is that the effect of differing initial conditions dissipates monotonically. Let the subscript n denote the value of $U(0)$, i.e., the initial conditions. (As before, we assume $A_T(0) = A_J(0) = 0$.)

Theorem 3.2. *(a)* $U_{n+m}(t) = U_n(t) + m - \Delta_{m,n}(t)$, *where*

$$\Delta_{m,n}(t) = \min\{m, O_{Jn}(t) + O_{T,n+m}(t)\} \tag{3.2}$$

for all $t \geq 0$, $m \geq 0$ *and* $n \geq 0$;

(b) $[O_{Jn}(t) + O_{T,n+m}(t)] - [O_{J,n+m}(t) + O_{Tn}(t)]$ *is a nondecreasing function of* t *with*

$$0 \leq [O_{Jn}(t) + O_{T,n+m}(t)] - [O_{J,n+m}(t) + O_{Tn}(t)] \leq m$$

for all $t \geq 0$, $m \geq 0$ *and* $n \geq 0$;

(c) $U_{n+m}(t) - U_n(t)$ *is a nonincreasing function of* t *with*

$$U_n(t) \leq U_{n+m}(t) \leq U_n(t) + m$$

for all $t \geq 0$, $n \geq 0$ *and* $m \geq 0$;

(d) $O_{T,n+m}(t) - O_{Tn}(t)$ *is a nondecreasing function of* t *with*

$$O_{Tn}(t) \le O_{T,n+m}(t) \le O_{Tn}(t) + m$$

for all $t \ge 0$, $n \ge 0$ *and* $m \ge 0$;

(e) $O_{Jn}(t) - O_{J,n+m}(t)$ *is a nondecreasing function of* t *with*

$$O_{Jn}(t) - m \le O_{J,n+m}(t) \le O_{Jn}(t)$$

for all $t \ge 0$, $n \ge 0$ *and* $m \ge 0$.

Proof. Apply mathematical induction on the arrival epochs, using (2.5)–(2.11). ■

An intuitive explanation of Theorem 3.2 is as follows: At time zero, $U_{n+m}(0) = U_n(0) + m$ and the number of job and token overflows is zero. For $m > 0$ and for common arrival processes $A_J(t)$ and $A_T(t)$, $U_{n+m}(t)$ will hit the upper boundary at C before $U_n(t)$, while $U_n(t)$ will hit the lower boundary at 0 before $U_{n+m}(t)$. If at a token arrival $U_{n+m}(t)$ is at the upper boundary, while $U_n(t)$ is not, then $U_n(t)$ increments and moves closer to $U_{n+m}(t)$. Likewise, if at job arrival $U_n(t)$ is at the lower boundary, while $U_{n+m}(t)$ is not, then $U_{n+m}(t)$ decrements and moves closer to $U_n(t)$. Once the total number of job overflows for the n-system plus token overflows for the $(n + m)$-system equals m, $U_{n+m}(t)$ and $U_n(t)$ coincide.

Theorem 3.2 implies that there is a random time after which the overflow processes coincide. This time is

$$T_{n,m} = \inf\{t \ge 0 : O_{Jn}(t) + O_{T,n+m}(t) \ge m\} . \tag{3.3}$$

Corollary 1. *For all* $n \ge 0$ *and* $m \ge 0$,

(a) $U_{n+m}(t) > U_n(t)$ *for* $t < T_{n,m}$ *and* $U_{n+m}(t) = U_n(t)$ *for all* $t \ge T_{n,m}$;

(b) $O_{Jn}(t) - O_{Jn}(T_{n,m}) = O_{J,n+m}(t) - O_{J,n+m}(T_{n,m})$ *for all* $t \ge T_{n,m}$;

(c) $O_{Tn}(t) - O_{Tn}(T_{n,m}) = O_{T,n+m}(t) - O_{T,n+m}(T_{n,m})$ *for all* $t \ge T_{n,m}$.

The following result extends Theorem 1 of [1].

Corollary 2. *The quantities* $\liminf_{t\to\infty} t^{-1}O_{Tn}(t)$, $\limsup_{t\to\infty} t^{-1}O_{Tn}(t)$, $\liminf_{t\to\infty} t^{-1}O_{Jn}(t)$ *and* $\limsup_{t\to\infty} t^{-1}O_{Tn}(t)$ *are independent of* n. *Hence, if the limits exist (i.e., if* lim inf = lim sup*), then they are independent of* n.

Proof. Apply Theorem 3.2(d) and (e). ■

4. Insensitivity When $C_J + C_T$ is Fixed: Stochastic Properties

We now consider $\{A_J(t) : t \geq 0\}$ and $\{A_T(t) : t \geq 0\}$ as stochastic processes and obtain results for the stochastic processes $\{U(t) : t \geq 0\}$, $\{O_J(t) : t \geq 0\}$ and $\{O_T(t) : t \geq 0\}$, assuming that $C = C_J + C_T$ is fixed. We assume that the initial condition $U(0)$ is a random variable independent of $\{A_J(t) : t \geq 0\}$ and $\{A_T(t) : t \geq 0\}$. Let π denote the initial probability distribution, i.e.,

$$\pi(n) = P(U(0) = n), \quad 0 \leq n \leq C_J + C_T. \tag{4.1}$$

We now modify the subscript convention used in §3. Let subscripts π and C_J denote initial distribution and the capacity of the job buffer. Let the subscript n in place of π denote the special case in which $\pi(n) = 1$; e.g., $O_{J,2,C_J}(t)$ is the job overflow process starting at $U(0) = 2$ with job buffer capacity C_J.

The following result provides weak conditions for the existence of a random time after which two systems with different initial distributions and different capacities evolve identically w.p.1.

Theorem 4.1. *If there exist* π, C_J *and* C_T *for which* $P(\tau < \infty) = 1$, *where*

$$\tau = \inf\{t \geq 0 : O_{J\pi C_J}(t) + O_{T\pi C_J}(t) \geq C_J + C_T\}, \tag{4.2}$$

then for each π', C_J' *and* C_T' *with* $C_J' + C_T' = C_J + C_T$ *there is a random time* $\tilde{\tau}$ *such that*

(a) $P(\tilde{\tau} \leq \tau < \infty) = 1$,

(b) $U_{\pi' C_J'}(t) = U_{\pi C_J}(t)$ *for all* $t > \tilde{\tau}$ w.p.1,

(c) $O_{J\pi' C_J'}(t) - O_{J\pi' C_J'}(\tilde{\tau}) = O_{J\pi C_J}(t) - O_{J\pi C_J}(\tilde{\tau})$ *for all* $t > \tilde{\tau}$ w.p.1.,

(d) $O_{T\pi' C_J'}(t) - O_{T\pi' C_J'}(\tilde{\tau}) = O_{T\pi C_J}(t) - O_{T\pi C_J}(\tilde{\tau})$ *for all* $t > \tilde{\tau}$ w.p.1.

Proof. The idea of the proof is to choose $\tilde{\tau}$ so that Corollary 1 to Theorem 3.2 can be applied and so that $\tilde{\tau} \le \tau$ almost surely. Let $\tilde{\pi}$ be a probability mass function representing the joint distribution of the initial conditions in both the given unprimed system and an arbitrary primed system, i.e.,

$$\pi(j) = \sum_k \tilde{\pi}(j,k) \text{ and } \pi'(k) = \sum_j \tilde{\pi}(j,k) . \tag{4.3}$$

For any joint initial condition $(U(0) = j,\ U'(0) = k)$ $0 \le j \le C_J + C_T$ and $0 \le k \le C_J + C_T$, choose $\tilde{\tau}$ to be the random variable $T_{n,m}$ defined in (3.3) where $n = \min\{j,k\}$ and $m = \max\{j,k\} - \min\{j,k\}$. Then Corollary 1 of Theorem 3.2 applied to each possible initial condition yields parts (b)–(d). We complete the proof by showing that $T_{n,m} \le \tau$ for all m and n, so that $\tilde{\tau} \le \tau$. The idea here to that $C_T + C_J$ is an upper bound on the number of overflows (jobs or tokens) from either system (primed or unprimed) that is required before $U(t) = U'(t)$. Hence, τ in (4.2) is an upper bound on the time until $U(t) = U'(t)$ and thus is greater than or equal to $T_{n,m}$ for any initial condition (j,k). More precisely, from (3.3) and (4.2), we see that it suffices to show

$$O_{Jn}(t) + O_{T,n+m}(t) - m \ge O_{Jk}(t) + O_{Tk}(t) - C_J - C_T \text{ for all } t$$

for all n, m and k with $0 \le n \le n+m \le C_J + C_T$ and $0 \le k \le C_J + C_T$, which follows from Theorem 3.2(d) and (e). Consider separately the cases: (i) $k < n$, (ii) $n \le k \le n + m$ and (iii) $n + m < k \le C_J + C_T$. ■

We now show that if the processes $U(t)$, $O_J(t)$ and $O_T(t)$ have steady-state limits for some initial distribution π and some capacities C_J and C_T, then the steady-state limits exist and are the same for all initial distributions and all capacities C_J' and C_T' for which $C_J' + C_T' = C_J + C_T$. Let $\Rightarrow$ denote convergence in distribution; see Billingsley [5].

Theorem 4.2. *Suppose that* $P(\tau < \infty) = 1$ *for* τ *in (4.2) and some* π, C_J *and* C_T. *Also suppose that*

$$[U_{\pi', C_J'}(s + t_1), U_{\pi', C_J'}(s + t_2), \ldots, U_{\pi', C_J'}(s + t_k)]$$

$$\Rightarrow [U^*(t_1), U^*(t_2), \ldots, U^*(t_k)], \tag{4.4}$$

$$[O_{J\pi' C_J'}(s + t_1) - O_{J\pi' C_J'}(s + t_0), \ldots, O_{J\pi' C_J'}(s + t_k) - O_{J\pi' C_J'}(s + t_{k-1})]$$

$$\Rightarrow [O_J^*(t_1) - O_J^*(t_0), \ldots, O_J^*(t_k) - O_J^*(t_{k-1})], \tag{4.5}$$

$$[O_{T\pi' C_J'}(s + t_1) - O_{T\pi' C_J'}(s + t_0), \ldots, O_{T\pi' C_J'}(s + t_k) - O_{T\pi' C_J'}(s + t_{k-1})]$$

$$\Rightarrow [O_T^*(t_1) - O_T^*(t_0), \ldots, O_T^*(t_k) - O_T^*(t_{k-1})] \tag{4.6}$$

as $s \to \infty$ *for all* $k \geq 1$ *and all* $0 \leq t_0 < t_1 < \ldots < t_k$ *for some* π' *and* C_J' *with* $C_J' + C_T' = C_J + C_T$. *Then (4.4)–(4.6) hold for all* π' *and all* C_J' *with* $C_J' + C_T' = C_J + C_T$, *with the limit being independent of the initial distribution* π' *and the capacity* C_J' *(provided that* $C_J' + C_T'$ *is fixed).*

To prove Theorem 4.2, we apply the following elementary lemma.

Lemma 4.1. *For any two random elements* X_1 *and* X_2 *and any measurable set* A,

$$|P(X_1 \in A) - P(X_2 \in A)| \leq P(X_1 \neq X_2) .$$

Proof. Write $P(X_i \in A) = P(X_i \in A, X_1 = X_2) + P(X_i \in A, X_1 \neq X_2)$. Then

$$|P(X_1 \in A) - P(X_2 \in A)| = |P(X_1 \in A, X_1 \neq X_2) - P(X_2 \in A, X_1 \neq X_2)|$$

$$\leq \max\{P(X_1 \in A, X_1 \neq X_2), P(X_2 \in A, X_1 \neq X_2)\} \leq P(X_1 \neq X_2) . \quad \blacksquare$$

Proof of Theorem 4.2. We only treat (4.4), because the other arguments are the same. Consider π'' and C_J'' with $C_J'' + C_T'' = C_J' + C_T' = C_J + C_T$. For any $x_1, \ldots, x_k$,

$$|P(U_{\pi'' C_J''}(s + t_1) \leq x_1, \ldots, U_{\pi'', C_J''}(s + t_k) \leq x_k)$$

$$- P(U^*(t_1) \leq x_1, \ldots, U^*(t_k) \leq x_k)|$$

$$\leq |P(U_{\pi'', C_J''}(s + t_1) \leq x_1, \ldots, U_{\pi'', C_J''}(s + t_k) \leq x_k)$$

$$- P(U_{\pi', C_J'}(s + t_1) \le x_1, \ldots, U_{\pi' C_J'}(s + t_k) \le x_k)|$$

$$+ |P(U_{\pi' C_J'}(s + t_1) \le x_1, \ldots, U_{\pi' C_J'}(s + t_k) \le x_k)$$

$$- P(U^*(t_1) \le x_1, \ldots, U^*(t_k) \le x_k)| \ . \tag{4.7}$$

By the definition of $\Rightarrow$, it suffices to consider only vectors $(x_1, \ldots, x_k)$ that are continuity points of the limit. For such vectors, the second term on the right in (4.7) converges to 0 as $s \to \infty$ by assumption, so that it suffices to consider the first term. By Lemma 4.1 and then Theorem 4.1(a) and (b),

$$|P(U_{\pi'' C_J''}(s + t_1) \le x_1, \ldots, U_{\pi'', C_J''}(s + t_k) \le x_k)$$

$$- P(U_{\pi' C_J'}(s + t_1) \le x_1, \ldots, U_{\pi' C_J'}(s + t_k) \le x_k)|$$

$$\le P(\tau > s) \to 0 \quad \text{as } s \to \infty \ . \quad \blacksquare$$

5. The Departure Process from the Throttle

In the previous sections we have seen that the overflow processes tend to be independent of the token-bank and job-buffer capacities C_T and C_J when $C_T + C_J$ is fixed. By Theorem 4.1, in great generality there exists a random time τ after which the overflow processes coincide in such systems. Consequently, after this random time τ, the accepted jobs coincide too. However, the timing of the job admissions is typically not the same because jobs may or may not wait in a job buffer after arrival before leaving the throttle. However, the difference is bounded, as we now show.

Let $D(t)$ be defined as in (2.4) and let D_n be the epoch of departure (admission) for the n^{th} job.

Theorem 5.1 *Consider two throttles with capacities (C_J, C_T) and (C_J', C_T') where $C_J < C_J'$ and $C_J + C_T = C_J' + C_T'$. Let the arrival processes be identical.*

(a) If both token banks start full, then tight bounds are

$$D(t) - (C_J' - C_J) \le D'(t) \le D(t) \text{ for } all\ t \ge 0 \tag{5.1}$$

and

$$D_n \le D_n' \le D_{n + C_J' - C_J} \text{ for } all\ n \ge 1\ . \tag{5.2}$$

(b) If the initial conditions are arbitrary, then tight bounds are

$$D(t) - C_T - C_J' \le D'(t) \le D(t) + C_J + C_T \text{ for } all\ t \tag{5.3}$$

and

$$D_{n - C_J - C_T} \le D_n' \le D_{n + C_T + C_J'} \text{ for } all\ n \ge C_T + C_J + 1\ . \tag{5.4}$$

Proof. From (2.4), $D(t)$ satisfies

$$D(t) = A_J(t) - O_J(t) - J(t) + J(0)$$

and similarly for $D'(t)$. In (a), since both token banks start full, $U(0) = U'(0) = C_J + C_T$, so that $U(t) = U'(t)$ for all t; see (2.1), (2.8) and Theorem 3.1. Moreover, $A_J'(t) = A_J(t)$, $O_J'(t) = O_J(t)$, $J(t) \le J'(t) \le J(t) + C_J' - C_J$ and $J'(0) = J(0) = 0$. The lower bound in (5.1) is realized if $U(t) = 0$, while the upper bound is realized if $U(t) = C_J + C_T$. The key to (5.2) is the fact that $D_n \le t$ if and only if $D(t) \ge n$ (for any sample path). To see that $D_n \le D_n'$, suppose that $D_n' = t$, which implies that $D'(t) \ge n$. By (5.1), $D(t) \ge D'(t)$. Consequently, $D(t) \ge n$, which implies that $D_n \le t = D_n'$. To see that $D_n' \le D_{n + C_J' - C_J}$, suppose that $D_{n + C_J' - C_J} = t$, which implies that $D(t) \ge n + C_J' - C_J$. By (5.1), $D'(t) \ge D(t) - (C_J' - C_J)$, so that $D'(t) \ge n$ and, thus, $D_n' \le t = D_{n + C_J' - C_J}$. As indicated by Theorem 3.2, the extreme cases in (b) are obtained by having one system start with the job buffer full and the other with the token bank full. Then one system can have up to $C_J + C_T$ more departures before $U'(t) = U(t)$. For example, to get the first inequality in (5.3), subtract $C_J + C_T$ from the first term in (5.1). ■

As an elementary consequence of Theorem 5.1, we see that the throughputs are the same. We define the *throughput* as the limit of $D(t)/t$ as $t \to \infty$.

Corollary. *Consider the two throttles in Theorem 5.1. If one of the four limits*

$$\lim_{t\to\infty} \frac{D(t)}{t}, \quad \lim_{t\to\infty} \frac{D'(t)}{t}, \quad \lim_{n\to\infty} \frac{n}{D_n} \quad \text{and} \quad \lim_{n\to\infty} \frac{n}{D'_n}$$

exists, then all four do and they are equal.

Proof. It is well known and easy to verify that $t^{-1}D(t) \to \theta$ as $t \to \infty$ if and only if $n/D_n \to \theta$ as $n \to \infty$. The rest follows from Theorem 5.1. ■

Similarly, other asymptotic properties of the departure processes are the same, in particular, the limiting values of the indices of dispersion for counts and intervals. The *index of dispersion for counts* (IDC) is the function

$$I_c(t) = \frac{\text{Var } D(t)}{ED(t)}, \ t > 0, \tag{5.5}$$

for t such that $ED(t) > 0$, where Var is the variance. The *index of dispersion for intervals* (IDI) is the function

$$I_i(n) = \frac{n \text{ Var } D_n}{(E D_n)^2}, \ n \geq 1, \tag{5.6}$$

for n such that $ED_n > 0$, where D_n is the epoch of the n^{th} departure; see [11].

The following is an easy consequence of Theorem 5.1. It closely parallels §1 of Whitt [22].

Theorem 5.2. *Consider the two throttles in Theorem 5.1.*

(a) The limits

$$\lim_{t\to\infty} \frac{\text{Var } D(t)}{t} \quad \text{and} \quad \lim_{t\to\infty} \frac{ED(t)}{t}$$

exist if and only if the limits

$$\lim_{t\to\infty} \frac{\text{Var } D'(t)}{t} \quad \text{and} \quad \lim_{t\to\infty} \frac{ED'(t)}{t}$$

exist. Moreover, if they exist, then the variance limits are equal and the mean limits are equal. If the limits above all exist with $\lim_{t\to\infty} t^{-1}ED(t) > 0$, *then*

$$\lim_{t\to\infty} I_c(t) = \lim_{t\to\infty} I'_c(t) . \tag{5.7}$$

(b) The limits

$$\lim_{n\to\infty} \frac{\operatorname{Var} D_n}{n} \quad \text{and} \quad \lim_{n\to\infty} \frac{ED_n}{n}$$

exist if and only if the limits

$$\lim_{n\to\infty} \frac{\operatorname{Var} D'_n}{n} \quad \text{and} \quad \lim_{n\to\infty} \frac{ED'_n}{n}$$

exist. Moreover, if they exist, then the variance limits are equal and the mean limits are equal. If the limits all exist with $\lim_{n\to\infty} n^{-1}ED_n > 0$, *then*

$$\lim_{n\to\infty} I_i(n) = \lim_{n\to\infty} I'_i(n) . \tag{5.8}$$

Typically the limits in parts (a) and (b) of Theorem 5.2 will be related, so that the limiting values of the IDI and IDC coincide. To obtain an explicit condition, we exploit central limit theorems and uniform integrability; see p. 32 of Billingsley [5]. The random variable L below typically has a normal distribution.

Theorem 5.3. *Suppose that one of*

$$n^{-1/2}(D_n - \theta^{-1}n) \Rightarrow L \text{ as } n \to \infty \tag{5.9}$$

or

$$t^{-1/2}(D(t) - \theta t) \Rightarrow -\theta^{3/2}L \text{ as } t \to \infty \tag{5.10}$$

holds with $0 < \theta < \infty$, *where* L *is an arbitrary proper random variable. Moreover, suppose that both* $n^{-1}(D_n - \theta^{-1}n)^2$ *and* $t^{-1}(D(t) - \theta t)^2$ *are uniformly integrable. Then the limits in (5.7) and (5.8) hold and are equal.*

Proof. By Theorem 6 of Glynn and Whitt [12], the limits in (5.9) and (5.10) are equivalent. Hence, if one holds, both do. The uniform integrability implies that

the first two moments in (5.9) and (5.10) also converge. As a consequence, we obtain $n^{-1}ED_n \to \theta^{-1}$ and $n^{-1}\operatorname{Var} D_n \to \operatorname{Var} L$ as $n \to \infty$ from (5.9) and $t^{-1}ED(t) \to \theta$ and $t^{-1}\operatorname{Var} D(t) \to \theta^3 \operatorname{Var} L$ as $t \to \infty$ from (5.10). Then, by (5.5) and (5.6),

$$\lim_{n\to\infty} I_i(n) = \theta^2 \operatorname{Var} L = \lim_{t\to\infty} I_c(t) . \quad \blacksquare$$

Unfortunately, it is not possible to relate the convergence of $n^{-1}ED_n$ to the convergence of $t^{-1}ED(t)$ in general (unlike without expectations, which we used in the proof of the Corollary to Theorem 5.1).

The next result provides weak conditions for heavy-traffic limiting behavior at a downstream queue to be the same. By *heavy-traffic limiting behavior* we mean the limiting behavior of an appropriate normalization of the queue-length process as the traffic intensity ρ approaches the critical value for stability; see Iglehart and Whitt [14]. As in [14] and [5], here we use *functional central limit theorems* (FCLTs). We say that $D(t)$ satisfies a FCLT if the normalized process $n^{-\alpha}[D(nt) - \theta nt]$ for $\alpha > 0$ converges in distribution as $n \to \infty$ (in the function space $D[0, \infty)$) to a proper limit (which is typically Brownian motion; also typically $\alpha = 1/2$). We assume that the departure processes from the throttles are fixed, so that the sequence of queueing systems in the heavy-traffic limit theorem with the associated sequence of traffic intensities converging to 1 is obtained by modifying the service times at the downstream queue.

Theorem 5.4. *Consider the two throttles in Theorem 5.1. The process $D'(t)$ satisfies a FCLT if and only if $D(t)$ does, in which case the limits are the same and both processes produce the same heavy-traffic limiting behavior at a downstream queue (with infinite-capacity, finitely many servers and the first-come first-served discipline) as specified by Theorem 1 of Iglehart and Whitt [14].*

Proof. The equivalence of the FCLTs is an easy consequence of Theorem 5.1 here and Theorem 4.1 of [5]. Then apply Theorem 1 of Iglehart and Whitt [14], which states that the heavy-traffic behavior depends on the arrival process only through its FCLT. ■

6. Bounds on The Impact of a Job Buffer on a Downstream Queue

We now apply a general comparison result in Berger and Whitt [2] to bound the impact of a job buffer on a downstream queue. For the reference system, to be referred to as system 2, we assume that the downstream queue has s servers, $C_D + C_J$ extra waiting spaces ($C_D \geq 0$, $C_J \geq 1$), and the first-come first-served queue discipline. We assume that service times are assigned when service begins. The arrival process to this queue is a superposition of two component arrival processes. One component arrival process A_e is exogeneous and the other is the departure process from a rate-control throttle based on a token bank of capacity C and no job buffer, with arrival process A_J. The exogeneous arrival process allows us to draw conclusions about multiple throttles feeding the downstream queue. In particular, we can make the comparison by considering one throttle at a time.

We show that the throughput from this downstream queue decreases if we simultaneously remove buffer space C_J from the downstream queue, add a job buffer of capacity C_J to the throttle, and reduce the token-bank capacity to $C - C_J$, assuming that $C_J \leq C$. In the modified system, referred to as system 1, the throttle has a job buffer of capacity C_J but the same total throttle capacity (token bank capacity plus job buffer capacity) C, and hence the same admission rate from the throttle as in system 2, by the Corollary to Theorem 5.1. The comparison implies that the benefit of a job buffer of capacity C_J in a throttle of fixed total capacity C is less than adding capacity C_J to the downstream queue. For m throttles without job buffers feeding a downstream queue, higher throughput is attained with capacity $C_D + C_{J_1} + \ldots + C_{J_m}$ at a downstream queue than with capacity C_D at the downstream queue and a job buffer of capacity C_{J_i} in throttle i, with the total capacity of throttle i unchanged, for each i.

This result is consistent with intuition, expressing the well known advantage of statistical multiplexing. However, the generality of our result expressed via a sample-path comparison does not seem obvious or easy to prove.

Let C_k^i be the epoch that the k^{th} job to start service (in the downstream queue) is admitted to system i; let Y_k^i be the epoch that the k^{th} job to start service starts

service in system i; let Z_k^i be the epoch that the k^{th} job to start service departs from the downstream queue in system i; and let D_k^i be the epoch of the k^{th} departure from system i. Let S_k be the k^{th} service time in both systems, which we assume is assigned to the k^{th} job to start service in each system.

Theorem 6.1. *Consider the two systems defined above, with common service times, common arbitrary job arrival processes A_e and A_J, and a common token arrival process A_T, in which the service times are assigned at the downstream queue in order of service initiation. Suppose that the systems start with the downstream queue empty and their token banks full. Then*

$$Y_k^1 \geq Y_k^2 \text{ , } Z_k^1 \geq Z_k^2 \text{ and } D_k^1 \geq D_k^2 \text{ for } \textit{all } k \text{ .}$$

Proof. We apply Corollary 3 in [2]; see [2] for background. In particular, let A/A/s/c denote an s-server queue with total capacity c, i.e., with a waiting room of size $c - s$, $1 \leq s \leq c \leq \infty$, in which jobs are served in order of their arrival by the first available server without defections after entering the system. If there is a finite waiting room and the system is full when a job arrives, then the job leaves without receiving service or affecting future arrivals. The first A means that the arrival process is arbitrary, not necessarily renewal and not necessarily stationary. The second A means that the service times are also arbitrary. We consider system 2 as an A/A/s/c queue by letting the arrival process be the superposition of the two arrival processes, A_e and $A_J(t) - O_J(t)$; i.e., we consider the arrival process to the throttle after the job overflows have been deleted. For system 2, this is indeed the arrival process to the downstream queue. We consider system 1 with this same arrival process. By Theorem 3.1, the overflow processes from the throttles in systems 1 and 2 are identical.

To make the comparison, we create a new system, called system 3, that we can conveniently compare to systems 1 and 2. System 3 is a modification of system 1 such that $Y_k^1 \geq Y_k^3$ and such that Corollary 3 in [2] can be applied to systems 3 and 2 to obtain, in particular, $Y_k^3 \geq Y_k^2$ and hence conclude that $Y_k^1 \geq Y_k^2$. The remaining orderings between systems 1 and 2 follow easily from additional observations.

We now indicate how to construct system 3 from system 1. The main idea is to regard the job buffer as part of the downstream queue in system 3. Hence, the downstream queues in both systems 2 and 3 have s servers and $C_D + C_J$ extra waiting spaces. To apply Corollary 3 in [2] to compare systems 2 and 3, system 3 must behave like an A/A/s/c queue with the first-come first-served discipline. To obtain system 3 from system 1, we first increase the service times of some jobs at the downstream queue. In particular, we increase the service times whenever the job buffer is not empty and the downstream queue is empty. Assigning longer service times allows us to regard the server at the downstream queue as never being idle when a job is in the job buffer of the throttle (which is necessary for system 3 to be regarded as an A/A/s/c model). There are two cases. If a job completing service from the downstream queue in system 1 makes the downstream queue empty with jobs in the job buffer, then the service time of this completing job is increased in system 3 by the time until the next arrival at the downstream queue; i.e., in system 3 we regard this completing job as still being in service until this condition terminates, i.e., until the next job arrives to the downstream queue to request service. The second case involves a new job entering and queueing in a previously empty job buffer when the downstream queue is empty in System 1. (For this to occur, the token bank must be empty also.) Then in system 3 we act as if this new job arrival is already in service at the downstream queue by increasing its service time by the duration of this condition. With this modification, we have $Y_k^1 \geq Y_k^3$. However, we must also consider the possibility of an arrival from the exogeneous arrival process A_e while this condition is in progress. Upon such an arrival, we switch the identities of the jobs in system 3, i.e., we put the exogeneous arrival in the job buffer and have the job in the job buffer go to the downstream queue where it continues to receive service. Thus the job from the job buffer has the longer service time. Of course, it is possible that the exogeneous job may later have a longer service time too. Note this switching of job identities also preserves the first-come first-served discipline; i.e., system 3 is constructed to have the first-come first-served discipline whereas system 1 does not.

Second, the identity switch above is actually required under more general circumstances in order to ensure that system 3 has the first-come first-served discipline. In particular, whenever a job arrives in the exogeneous stream A_e and finds space in the downstream queue while there are jobs in the job buffer, we put the first job that is in the job buffer queue into the downstream queue and we put this exogeneous job at the end of the job buffer queue. Note that all identity switching occurs only because of the exogeneous arrivals.

Third, we must account for jobs leaving the job buffer in system 1 and being lost because the original buffer of capacity C_D is full. In system 3, we represent these losses as balking from the queue of capacity $C_D + C_J$ (which is also covered by Corollary 3 to Theorem 1 in [2]). Finally, in system 3 we reject some extra jobs from the exogeneous stream A_e, because these jobs are really not allowed to go into the job buffer. In particular, we reject jobs from A_e in system 3 whenever upon arrival the job finds the downstream queue full. At this downstream queue, we thus have $S_k^3 \geq S_k^2$ for all k and $A^3 \subseteq A^2$ so that we can apply Corollary 3 to Theorem 1 in [2] to deduce that $C_k^3 \geq C_k^2$, $Y_k^3 \geq Y_k^2$, $Z_k^3 \geq Z_k^2$ and $D_k^3 \geq D_k^2$ for all k. Thus, in particular, $Y_k^3 \geq Y_k^2$ for all k and the modifications of system 1 to create system 3 yielded the ordering $Y_k^1 \geq Y_k^3$ for all k. Thus, $Y_k^1 \geq Y_k^2$ for all k. Moreover, since $Z_k^i = Y_k^i + S_k$, $Z_k^1 \geq Z_k^2$ for all k. Finally, note that D_k^i is determined from $\{Z_1^i, \ldots, Z_{k+s-1}^i\}$ by

$$D_k^i = \min_k \{Z_1^i, \ldots, Z_{k+s-1}^i\} ,$$

where $\min_k$ denotes the k^{th} smallest number (see [2]). Hence

$$D_k^1 = \min_k \{Z_1^1, \ldots, Z_{k+s-1}^1\} \geq \min_k \{Z_1^2, \ldots, Z_{k+s-1}^2\} = D_k^2 .$$

Remark 6.1 From the proof of Theorem 6.1, note that when there are no exogeneous arrivals, there are no identity switches, so that then the admission epochs to the full system of the jobs that are served coincide in systems 1 and 3, i.e., $C_k^1 = C_k^3$ for all k. Hence, then $C_k^1 \geq C_k^2$ for all k in addition to the stated conclusions of Theorem 6.1.

Remark 6.2. Theorem 6.1 has rather special initial conditions, but other initial

conditions can be introduced by having extra job arrivals in A_e and/or A_J at time 0.

Remark 6.3. If service times are not assigned at service initiation times, but the service times are i.i.d. and independent of the arrival process, then we can apply Theorem 6.1 to obtain stochastic comparisons using the stochastic subsequence ordering; see Corollary 4 of [2].

We now state the advertised consequences for a downstream queue fed by several separate rate control throttles.

Corollary. *Consider a downstream s-server queue with waiting room* $C_D + C_{J_1} + \ldots + C_{J_m}$ *fed by the superposition of departure processes from m token-bank rate-control throttles without job buffers. Let the service times at the downstream queue be assigned in order of service initiation and let the system start with the downstream queue empty and all the token banks full. Then the number of departures in* $(0,t]$ *for any* t *and thus the throughput (the limiting departure rate) from the downstream queue are greater than in a corresponding system with capacities modified as follows: There is a waiting room of size* C_D *at the downstream queue and a job buffer of capacity* C_{J_i} *in throttle i, with the total capacity of throttle i unchanged, for each i. (The remaining features of the two systems are the same.)*

7. Changing Total Capacity

In this section we consider the effect of changing the total throttle capacity $C = C_J + C_T$. Here we provide some extensions to the monotonocity results of Budka and Yao [7]. Let a subscript C indicate the total capacity. Budka and Yao [7] show in their Theorem 1 that the number of accepted jobs $D(t)$ is increasing and concave in the capacity C. From (2.8) and (2.11), we immediately obtain the following related result.

Theorem 7.1. *For all* t, $U_C(t)$ *is an nondecreasing function of C, assuming* $U_C(0)$ *remains unchanged.*

Proof. Apply mathematical induction on the arrival epochs once again with (2.8) providing the key structural property. ∎

Theorem 7.1 has a rather restrictive assumption on the initial conditions, but it disappears in the limit.

Corollary. *Assume that (4.2) holds and* $U_C(t) \Rightarrow U_C(\infty)$ *as* $t \to \infty$ *for some initial condition for all C. Then* $Ef(U_C(\infty))$ *is nondecreasing in C for every nondecreasing real-valued f.*

Proof. Assuming special initial conditions, the conclusion follows from Theorem 7.1, since $Ef(U_C(t)) \le Ef(U_{C+1}(t))$ for all t when f and $U_C(t)$ are both nondecreasing. Moreover, $Ef(U_C(t)) \to Ef(U_C(\infty))$ when $t \to \infty$ and f is a bounded continuous nondecreasing real-valued function if $U_C(t) \Rightarrow U_C(\infty)$, and stochastic order is determined by the expectations of such functions; e.g., see Theorem 2.6 of Whitt [20]. Finally, by Theorem 4.2, $U_C(t) \Rightarrow U_C(\infty)$ as $t \to \infty$ for all initial conditions if it holds for one initial condition. ■

We now proceed to obtain some more detailed information.

Theorem 7.2. *Consider two systems with total capacity C and* $C + 1$. *Let* $U_C(0) = U_{C+1}(0)$. *Let the arrival processes* $A_T(t)$ *and* $A_J(t)$ *be the same. Then*

$$(a) \quad U_{C+1}(t) = U_C(t) + [O_{T,C}(t) - O_{T,C+1}(t)] - [O_{J,C}(t) - O_{J,C+1}(t)] \text{ for } all\ t,$$

$$(b) \quad U_C(t) \le U_{C+1}(t) \le U_C(t) + 1 \text{ for } all\ t,$$

$$(c) \quad 0 \le [O_{T,C}(t) - O_{T,C+1}(t)] - [O_{J,C}(t) - O_{J,C+1}(t)] \le 1$$

for all t,

$$(d) \quad O_{T,C}(t_2) - O_{T,C}(t_1) \ge O_{T,C+1}(t_2) - O_{T,C+1}(t_1)$$

for all $t_1 < t_2$;

$$(e) \quad O_{J,C}(t_2) - O_{J,C}(t_1) \ge O_{J,C+1}(t_2) - O_{J,C+1}(t_1)$$

for all $t_1 < t_2$.

Proof. As before, apply mathematical induction on arrival epochs. Note that at each transition we either maintain or switch between $U_{C+1}(t_n) - U_C(t_n) = 0$ and $U_{C+1}(t_n) - U_C(t_n) = 1$. We transition from $U_{C+1}(t_n) - U_C(t_n) = 0$ to

$U_{C+1}(t_{n+1}) - U_C(t_{n+1}) = 1$ if we have an extra token overflow in the process with capacity C at time t_{n+1}. We transition from $U_{C+1}(t_n) - U_C(t_n) = 1$ to $U_{C+1}(t_{n+1}) - U_C(t_{n+1}) = 0$ if we have an extra job overflow in the process with capacity C at time t_{n+1}. Part (c) just combines (a) and (b). ■

Corollary 1. *(Budka and Yao [7]) Suppose that*

$$\lambda'_C = \lim_{t\to\infty} \frac{O_{JC}(t)}{t} \quad \text{and} \quad r'_C = \lim_{t\to\infty} \frac{O_{TC}(t)}{t}$$

are well defined. Then λ'_C and r'_C are nonincreasing in C.

Remark 7.1. It is significant that this corollary is *not* true for the standard A/A/1/C queue in which service times are associated with arrivals; see [23].

Let $\leq_{st}$ denote ordinary stochastic order.

Corollary 2. *(Budka and Yao [7]) If $U_C(t) \Rightarrow U_C(\infty)$ and $U_{C+1}(t) \Rightarrow U_{C+1}(\infty)$ as $t \to \infty$, then*

$$U_C(\infty) \leq_{st} U_{C+1}(\infty) .$$

Corollary 3. *(a) Suppose that C_T is increased to $C'_T = C_T + 1$, while C_J and $U(0)$ are unchanged. Then*

$$T(t) \leq T'(t) \leq T(t) + 1$$

and

$$J(t) - 1 \leq J'(t) \leq J(t) \text{ for all } t .$$

(b) Suppose that C_J is increased to $C'_J = C_J + 1$ while C_T and $U(0)$ are unchanged. Then

$$J(t) \leq J'(t) \leq J(t) + 1$$

and

$$T(t) - 1 \leq T'(t) \leq T(t) \text{ for all } t .$$

Corollary 4. *Suppose that*

$$T_C(t) \Rightarrow T_C(\infty) \quad \text{and} \quad J_C(t) \Rightarrow J_C(\infty) \quad \text{as } t \to \infty$$

for all C and all initial conditions.

(a) If $C_T' = C_T + 1$ *and* $C_J' = C_J$, *then*

$$T(\infty) \leq_{st} T'(\infty) \leq_{st} T(\infty) + 1$$

and

$$J(\infty) - 1 \leq_{st} J'(\infty) \leq_{st} J(\infty) .$$

(b) If $C_J' = C_J + 1$ *and* $C_T' = C_T$, *then*

$$J(\infty) \leq_{st} J'(\infty) \leq_{st} J(\infty) + 1$$

and

$$T(\infty) - 1 \leq_{st} T'(\infty) \leq_{st} T(\infty) .$$

8. Changing Arrival Processes

Budka and Yao [7] applied the subsequence ordering to compare two throttles (without a job buffer) with two different token arrival processes. Here is the deterministic variant of their result. Let $D^i(t)$ be the departure process in (2.4) in system i.

Theorem 8.1. *(Budka and Yao [7]) Consider two rate-control throttles without job buffers. If* $A_T^1 \subseteq A_T^2$, $C_T^1 \leq C_T^2$ *and* $T^1(0) \leq T^2(0)$, *then* $T^1(t) \leq T^2(t)$ *for all* t *and* $D^1 \subseteq D^2$.

We now obtain a related comparison result for throttles with a job buffer. With our greater generality, the conclusion is necessarily weaker. (It is easy to construct examples showing that neither $D^1 \subseteq D^2$ nor $U^1(t) \leq U^2(t)$ need hold.) Following [7], let $D(t)$ denote the departure process from the throttle in (2.4), let $H(t) \equiv D(t) + J(t)$ be the counting process recording the number of jobs that have arrived and not overflowed by time t, and let $I(t) \equiv D(t) + T(t)$ be the

counting process recording the number of tokens that have arrived and not overflowed by time t.

Theorem 8.2. *Consider two rate-control throttles with common token banks and job buffers that are initially empty. If $A_J^1 \subseteq A_J^2$ and $A_T^1 \subseteq A_T^2$, then $D^1(t) \le D^2(t)$, $H^1(t) \le H^2(t)$ and $I^1(t) \le I^2(t)$ for all t.*

Proof. Let $\{t_k : k \ge 1\}$ be the sequence of arrival epochs for jobs and tokens in system 2, where $t_k < t_{k+1}$. At each epoch t_k a batch of jobs and a batch of tokens arrives. (One of these batches may be empty.) By the assumed orderings $A_J^1 \subseteq A_J^2$ and $A_T^1 \subseteq A_T^2$, all arrivals in system 1 occur at these epochs too. Moreover, at each arrival epoch the batch sizes are always larger or the same size in system 2. Let $\{J_k^i : k \ge 1\}$ and $\{T_k^i : k \ge 1\}$ be the sequences of job and token batch sizes in system i at the epochs $\{t_k : k \ge 1\}$. (Note that, for any given k, as many as three of J_k^1, J_k^2, T_k^1 and T_k^2 may be zero.) We apply mathematical induction on the indices k to deduce that the orderings

$$\begin{aligned} D^1(t_k) &\le D^2(t_k) \\ D^1(t_k) + J^1(t_k) &\le D^2(t_k) + J^2(t_k) \\ D^1(t_k) + T^1(t_k) &\le D^2(t_k) + T^2(t_k) \end{aligned} \tag{8.1}$$

are maintained for all k, which implies the desired result. First, it is easy to see that (8.1) holds for $k = 1$. In establishing (8.1), recall that only one of $J^i(t_k)$ and $T^i(t_k)$ can be positive at any time. Our convention (see §2) has been to treat simultaneous job and token arrivals by first pairing and admitting $\min\{J_k^i, T_k^i\}$ and then sending the excess to the throttle. The evolution of the throttles can thus be defined by the following modification of the recursion in (4) and (5) in §V of [7]. If $T_{n+1}^i \ge J_{n+1}^i$, then

$$\begin{aligned} D^i(t_{n+1}) + J^i(t_{n+1}) \equiv H^i(t_{n+1}) &= H^i(t_n) + J_{n+1}^i \\ D^i(t_{n+1}) + T^i(t_{n+1}) \equiv I^i(t_{n+1}) &= \min\{I^i(t_n) + T_{n+1}^i,\ H^i(t_{n+1}) + C_T\} \end{aligned} \tag{8.2}$$

If $T_{n+1}^i < J_{n+1}^i$, then

$$D^i(t_{n+1}) + T^i(t_{n+1}) \equiv I^i(t_{n+1}) = I^i(t_n) + T^i_{n+1}$$

$$D^i(t_{n+1}) + J^i(t_{n+1}) \equiv H^i(t_{n+1}) = \min\{H^i(t_n) + J^i_{n+1},\ I^i(t_{n+1}) + C_J\} \ . \quad (8.3)$$

Finally, in either case,

$$D^i(t_{n+1}) = \min\{H^i(t_{n+1}),\ I^i(t_{n+1})\} \ . \quad (8.4)$$

For example, the first line of (8.2) holds because the J^i_{n+1} jobs are paired with tokens and all admitted. The second line in (8.2) holds, because we can add all T^i_{n+1} tokens until the tokens exceed the number of jobs admitted by the token-bank capacity C_T. From (8.2)–(8.4), it is immediate that (8.1) is maintained when $T^i_{n+1} \geq J^i_{n+1}$ for both i and when $T^i_{n+1} < J^i_{n+1}$ for both i, because $T^1_{n+1} \leq T^2_{n+1}$ and $J^1_{n+1} \leq J^2_{n+1}$. Hence, there are two remaining cases:

$$J^2_{n+1} > T^2_{n+1} \geq T^1_{n+1} \geq J^1_{n+1} \quad (8.5)$$

and

$$T^2_{n+1} \geq J^2_{n+1} \geq J^1_{n+1} > T^1_{n+1} \ . \quad (8.6)$$

We consider only (8.5), because the reasoning for (8.6) is essentially the same. We exploit the fact that

$$I^i(t_n) \geq H^i(t_n) - C_J \quad (8.7)$$

and

$$H^i(t_n) \geq I^i(t_n) - C_T \ . \quad (8.8)$$

To obtain (8.7) note that subtracting the number of departures, $D^i(t_n)$, from both sides of (8.7) yields

$$T^i(t_n) \geq J^i(t_n) - C_J \ ,$$

which is valid, because the left side is nonnegative and the right side is nonpositive. The same observation yields (8.8). Given (8.5), from (8.3) and then the induction assumption applied to the H's and I's (using first (8.5), then (8.7) and finally (8.2)), we obtain

$$H^2(t_{n+1}) = \min\{H^2(t_n) + J^2_{n+1},\, I^2(t_n) + T^2_{n+1} + C_J\}$$

$$\geq \min\{H^1(t_n) + J^1_{n+1},\, I^1(t_n) + J^1_{n+1} + C_J\}$$

$$\geq H^1(t_n) + J^1_{n+1} = H^1(t_{n+1}) \,. \tag{8.9}$$

Given (8.5), from (8.2), the induction assumption applied to the I's (using the subsequence ordering for the T's) and (8.3), we obtain

$$I^1(t_{n+1}) = \min\{I^1(t_n) + T^1_{n+1},\, H^1(t_{n+1}) + C_T\}$$

$$\leq I^1(t_n) + T^1_{n+1}$$

$$\leq I^2(t_n) + T^2_{n+1} = I^2(t_{n+1}) \,. \tag{8.10}$$

Finally, we obtain $D^1(t_{n+1}) \leq D^2(t_{n+1})$ from (8.4), (8.9) and (8.10). ■

Remark 8.1. As in [7], from (8.2)–(8.4), it follows that the triple $(H(t_n), I(t_n), D(t_n))$ is nondecreasing and concave in (C_J, C_T). This result does not quite follow directly from [7], because we treat simultaneous job and token arrivals differently here. Similarly, $(H(t_n), I(t_n), D(t_n))$ is nondecreasing and concave in $(J_1, \ldots, J_n, T_1, \ldots, T_n)$ from which second-order stochastic comparisons can be deduced. ■

9. Non-Discrete Flow Models

The throttle model in §2 has discrete jobs and tokens that arrive according to the integer-valued counting processes $A_J(t)$ and $A_T(t)$. However, the results in this paper extend to the case of continuous divisible quantities of "work" and "credit," as occur for example with the fluid model of Elwalid and Mitra [10]. Then $A_J(t)$ and $A_T(t)$ can have general nondecreasing sample paths. (As a regularity condition, we assume that these sample paths are right-continuous.) In this more general setting, we assume that the triple $(U(t), O_J(t), O_T(t))$ is defined by applying the two-sided regulator with reflecting barriers at 0 and $C \equiv C_J + C_T$ to the net-input process $X(t) \equiv A_T(t) - A_J(t)$, as on pp. 21-24 of Harrison [13] and §4 of Berger and Whitt [3]. Moreover, we assume that $T(t)$, $J(t)$ and $D(t)$ can be defined by (2.2)–(2.4). In §2 we noted that these definitions

could also be used for the discrete job-and-token model with integer-valued counting processes $A_J(t)$ and $A_T(t)$, so we are simply extending the current model.

The key property for obtaining comparison results in this more general setting is that the sample paths of $A_J(t)$ and $A_T(t)$ can be approximated uniformly in any finite interval by piecewise-constant discrete-valued functions. In particular, to achieve a uniform approximation to within n^{-1}, consider $\lfloor nA_J(t) \rfloor / n$ and $\lfloor nA_T(t) \rfloor / n$ for $t \geq 0$, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x. All the inductive proofs here apply directly to the associated piecewise-constant discrete-valued net-input processes approximating $X(t)$.

We then represent the general case with input pair $(A_T(t), A_J(t))$ as the limit of the sequence of approximating piecewise-constant discrete-valued input pairs $\{(A_T^n(t), A_J^n(t)) : n \geq 1\}$. We than have convergence of the other processes, i.e., $[X^n(t), U^n(t), O_T^n(t), O_J^n(t), T^n(t), J^n(t), D^n(t)] \rightarrow [X(t), U(t), O_T(t), O_J(t), T(t), J(t), D(t)]$ as $n \rightarrow \infty$, by continuity. In particular, with the supremum norm on any finite interval and the maximum norm on product spaces, the maps from $(A_J(t), A_T(t))$ to $X(t) \equiv A_T(t) - A_J(t)$, from $X(t)$ to $(U(t), O_J(t), O_T(t))$, from $U(t)$ to $[T(t), J(t)]$ and from $[A_J(t), O_J(t), J(t)]$ to $D(t)$ are all continuous. For the map from $X(t)$ to $(U(t), O_J(t), O_T(t))$, see Theorem 4.2 of [3]. Indeed, there it is shown that the map from $X(t)$ to $U(t)$ is Lipschitz with modulus 2. All the other maps are elementary. Hence, all previous theorems extend to this more general setting of non-discrete flows.

These results apply to all regulated stochastic flow systems (defined by applying the two-sided regulator) in which the net input process $X(t)$ is of bounded variation, because then we have $X(t) = A_T(t) - A_J(t)$ for nondecreasing $A_T(t)$ and $A_J(t)$, but the results also apply even more generally. Since the process $(U(t), O_J(t), O_T(t))$ depends only on $X(t)$, to treat them it suffices to work directly with $X(t)$. Then it suffices for $X(t)$ to have left and right limits at every t in order to do the piecewise-constant discrete-valued approximation; e.g., see

p. 110 of Billingsley [5]. For example, the results here apply to two-sided regulated Brownian motion (RBM) in Harrison [13] and Berger and Whitt [3].

Finally, it should be noted that some non-discrete models do not fall directly into this two-sided regulator framework, because the two-sided regulator treats the work and credit as divisible. We might instead have discrete jobs and continuous credit, where we do not admit portions of a job. Instead, we admit the entire job when a full unit of credit has accumulated. However, this particular model is already covered by our original model; we simply count the accumulated integer amounts of credit; i.e., we let $A_T(t)$ be the integer part of the credit that has arrived in the interval $(0, t]$. It appears that other modifications can be treated similarly.

REFERENCES

[1] A. W. Berger, Performance analysis of a rate-control throttle where tokens and jobs queue, *IEEE J. Select. Areas Commun.* **9** (1991) 165-170.

[2] A. W. Berger and W. Whitt, Comparisons of multi-server queues with finite waiting rooms, *Stochastic Models,* this issue.

[3] A. W. Berger and W. Whitt, The Brownian approximation for rate-control throttles and the G/G/1/C queue, *Discrete Event Dynamic Systems,* 2 (1992) to appear.

[4] A. W. Berger and W. Whitt, Traffic Shaping by a job buffer in a token-bank rate-control throttle, submitted for publication, 1992.

[5] P. Billingsley, *Convergence of Probability Measures,* New York: Wiley, 1968.

[6] K. C. Budka, Stochastic monotonicity and concavity properties of rate-based flow control mechanisms, *IEEE Trans. Aut. Control,* to appear.

[7] K. C. Budka and D. D. Yao, Monotonicity and convexity properties of rate control throttles, Department of Industrial Engineering and Operations

Research, Columbia University, 1990. Abbreviated version in *Proceedings of 29th IEEE conference on Decision and Control,* (1990) 883-884.

[8] R. L. Cruz, A calculus for network delay, part I: network elements in isolation, *IEEE Trans. Inf. Thy.* **37** (1991) 114-121.

[9] R. L. Cruz, A calculus for network delay, part II: network analysis, *IEEE Trans. Inf. Thy.* **37** (1991) 121-141.

[10] A. Elwalid and D. Mitra, Analysis and design of rate-based congestion control of high-speed networks, I: stochastic fluid models, access regulation, *Queueing Systems* **9** (1991) 29-64.

[11] K. W. Fendick and W. Whitt, Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue, *Proceedings of the IEEE* **77** (1989) 171-194.

[12] P. W. Glynn and W. Whitt, Ordinary CLT and WLLN versions of $L = \lambda W$, *Math. Opns. Res.* **13** (1988) 674-692.

[13] J. M. Harrison, *Brownian Motion and Stochastic Flow Systems,* New York: Wiley, 1985.

[14] D. L. Iglehart and W. Whitt, Multiple channel queues in heavy traffic, II: sequences, networks and batches, *Adv. Appl. Prob.* **2** (1970) 355-369.

[15] M. Sidi, Z. Liu, I. Cidon and I. Gopal, Congestion control through input rate regulation, GLOBECOM '89, Dallas, Texas, pp. 1746-1768, 1989.

[16] K. Sohraby and M. Sidi, On the performance of bursty and correlated sources subject to leaky bucket rate-based access control schemes, IEEE INFOCOM '90, Bal Harbor, Florida, 1990.

[17] D. Sonderman, Comparing multi-server queues with finite waiting rooms, I: same number of servers, *Adv. Appl. Prob.* **11** (1979) 439-447.

[18] D. Sonderman, Comparing multi-server queues with finite waiting rooms, II: different number of servers, *Adv. Appl. Prob.* **11** (1979) 448-455.

[19] S. Suresh and W. Whitt, The heavy-traffic bottleneck phenomenon in open queueing networks, *Oper. Res. Letters* **9** (1990) 355-362.

[20] W. Whitt, Uniform conditional stochastic order, *J. Appl. Prob.* **17** (1980) 112-123.

[21] W. Whitt, Comparing counting processes and queues, *Adv. Appl. Prob.* **13** (1981) 207-220.

[22] W. Whitt, Approximations for departure processes and queues in series, *Naval Res. Log. Qtrly* **31** (1984) 499-521.

Received: 6/14/1991
Revised: 2/2/1992
Accepted: 3/24/1992

Recommended by Brad Makrucki, Editor