Large Fluctuations in a Deterministic Multiclass Network of Queues

Ward Whitt

Management Science, Vol. 39, No. 8 (Aug., 1993), 1020-1028.

Large Fluctuations in a Deterministic Multiclass Network of Queues

Ward Whitt
AT&T Bell Laboratories, Murray Hill, New Jersey 07974-0636

In this paper we investigate a relatively simple deterministic four-class two-queue multiclass open network of single-server FIFO queues with traffic intensity one at each queue. Our purpose is to better understand the effect of feedback with class-dependent service times at the queues. The example is sufficiently tractable that we are able to describe its transient behavior in great detail. The transient behavior depends strongly on the initial conditions and, for some initial conditions, the sample paths of the queue-length processes at individual stations have sudden large fluctuations (a large jump up followed immediately by a large jump down). These large fluctuations occur because batches of customers with short service times build up in the queues. Consistent with recent work by Dai and Wang (1993) on Brownian network models, these fluctuations rule out conventional heavy-traffic limit theorems. We show how to obtain proper heavy-traffic limits for this example by weakening the topology or enlarging the space of prospective limits (and changing the topology). This example also dramatically demonstrates a disadvantage of the FIFO discipline compared to other disciplines like head-of-the-line processor-sharing (HOL-PS) among the classes at each queue (under which, the large fluctuations do not occur). Finally, the critical arrival rate for stability in our example actually depends on the service discipline, being even lower if the classes with longer service times are given high priority at each queue. This phenomenon can occur in the network setting because individual queues can be empty when there is work in the network.

(*Multiclass Queueing Networks; Open Queueing Networks; Class-Dependent Service Times; Heavy-Traffic Limits; Diffusion Approximations; Priorities; Stability Criteria*)

1. Introduction

A useful model for many applications (e.g., manufacturing systems and communication networks) is the multiclass open network of queues (MONQ), in which each class has its own (possibly non-Poisson, possibly void) exogenous arrival process and its own (possibly nonexponential) service times. Transitions among classes occur according to a transient homogeneous Markov chain: Upon completing service, a class-i customer becomes a class-j customer with probability p_{ij}, and leaves the network with probability $1 - \sum_j p_{ij}$, independent of all previous events. The queue is embodied in the class, so that each class visits only one queue. For further discussion here, we otherwise make everything standard; i.e., let all queues have one server, unlimited waiting space and the first-in first-out (FIFO) service discipline. Moreover, let all service times and all exogenous interarrival times be mutually independent; let the exogenous interarrival times for each class be independent and identically distributed (i.i.d.) and let the service times of each class be i.i.d., but allow different classes to have different service-time distributions at each queue. Let all interarrival-time and service-time distributions have finite means and variances.

The MONQ model above uses the standard description of classes, as in Harrison and Nguyen (1990, 1993). A convenient (mathematically equivalent) alternative is based on a framework of deterministic routes, as in Segal and Whitt (1989). A deterministic route contains the (usual) sequence of successive queues to be visited,

0025-1909/93/3908/1020$01.25

the exogenous interarrival-time distribution (if any) and the service-time distributions at each queue on the route. (For two-moment approximation methods as in Segal and Whitt (1989), only the first two moments of each distribution need be given.) This deterministic-route framework is appealing because a route in the model typically corresponds to the natural notion of a type in the application, e.g., a product in a manufacturing line. Finally, the deterministic-route framework can include probabilistic transitions from one queue visit on one deterministic route to another queue visit on another deterministic route. This makes the deterministic-route framework fully equivalent to the standard multiclass Markovian approach, while allowing it to more directly match the application (which often has a non-Markovian character without exploiting classes). However, here we will use the conventional terminology.

In the stated generality, it is usually not possible to calculate performance measures of interest for this MONQ model. Thus, these MONQ's are not yet well enough understood. We contribute to a better understanding of MONQ's by carefully examining one example and a few variants. In particular, we are interested in the effect of class-dependent service times with feedback (customers can return to a queue where they previously received service).

Large Fluctuations and Instability

We find that feedback together with significant differences in class-dependent service times can have a dramatic impact. Even though customers may arrive in a very regular way (e.g., deterministically and evenly spaced), batches of customers with short service times can periodically build up in the queues. When these customers reach the server, there can be a sudden surge of departures. Moreover, a surge of departures from one queue can be followed quickly by a surge of arrivals to that queue. These surges can lead to dramatic fluctuations in the queue-length processes at the stations in the network.

The system evolution is also highly sensitive to the initial conditions, as with *deterministic chaos*; see Schuster (1988). Indeed, our model is similar to the model of Erramilli and Forys (1991), which they showed exhibits chaotic behavior. Related work on essentially the same model was done by Kumar and Seidman (1990) and Lu and Kumar (1991), and is reviewed in Kumar (1993); also see Rybko and Stolyar (1992). Kumar and Seidman show that the model with class-dependent priorities can be unstable (have queue lengths going to infinity) even when the traffic intensities at each queue are strictly less than one. This instability occurs because servers can be persistently idle even when there is always work in the system. (See §5.) Kumar and Seidman also show that nonacyclic flows rather than feedback is the reason for this instability. Lu and Kumar (1991) establish positive stability results for other service disciplines.

In this context, our results are interesting primarily because we consider the FIFO discipline. Our example suggests that the natural condition, requiring all traffic intensities be strictly less than one, is the proper stability condition with the FIFO discipline, but (added in proof) remarkably this is not so; see Bramson (1993) and Seidman (1993). Our example shows unstable behavior and large fluctuations at the critical point where all traffic intensities equal one.

Sharifnia (1992) has recently shown that instability can occur in a MONQ with the FIFO discipline for another reason. In particular, instability can occur because customers with server-dependent service times are assigned to the wrong servers.

Brownian Models and Heavy-Traffic Limits

A promising way to analyze MONQ's is via Brownian models, as in Harrison (1988) and Harrison and Nguyen (1990, 1993). Brownian models are especially attractive for investigating control schemes. These Brownian models can be regarded as direct approximations or the consequence of heavy-traffic limit theorems. Heavy-traffic limits for feedforward FIFO MONQ's were established by Johnson (1983) and Peterson (1991) and heavy-traffic limits for a single multiclass FIFO queue with feedback were established by Reiman (1988), but so far heavy-traffic limits for general MONQ's have yet to be established. Indeed, Dai and Wang (1991) recently showed that a candidate Brownian model for a MONQ does not exist. Our example "explains" the Dai-Wang example; i.e., our example not only shows that there need not be convergence to Brownian models for all MONQ's in heavy traffic, but there need not be convergence in the familiar function space D of right-continuous functions with left limits, endowed with one of the usual Skorohod (1956) topologies; see Billingsley (1968) and Whitt (1980). In particular, for our example, we show that such convergence does not take place. In our example, there are fluctuations of size $O(\sqrt{n})$ in time n as $n \rightarrow \infty$, instead of $o(\sqrt{n})$ as in single-class networks. Since the heavy-traffic limit theorem for MONQs would include our deterministic example as a special case, our example is a bonafide counterexample to the theorem. However, since our example is deterministic, it still remains to determine whether fluctuations of size $O(\sqrt{n})$ in time n as $n \rightarrow \infty$ can occur in genuine stochastic models. Dai and Nguyen (1994) have shown that this is possible (added in proof).

Laws (1991) has recently done a nice heavy-traffic analysis of MONQs. He does not encounter the difficulty above because he assumes that all service times at each queue are i.i.d.

The reason that the normalized queue-length processes in our MONQ example fail to converge is because they have large fluctuations (a large jump up followed immediately by a jump down), which rule out conventional heavy-traffic limits. (Of course, the large fluctuations in turn are a consequence of the model structure.) The normalized queue-length processes behave like the sequence of functions $\{x_n \colon n \geq 1\}$ in $D[0, 1]$, defined by

$$x_n(t) = 1_{[2^{-1}+n^{-1}, 2^{-1}+2n^{-1})}(t), \quad 0 \leq t \leq 1, \qquad (1.1)$$

where 1_A is the indicator function of the set A. It is well known and easy to verify that $\{x_n\}$ in (1.1) does not converge as $n \rightarrow \infty$ in $D[0, 1]$ with the standard Skorohod (1956) J_1 topology in Billingsley (1968) or in any of the other Skorohod (1956) topologies. This is easy to see, because these topologies all reduce to uniform convergence over bounded time intervals when the limit function is continuous.

There are two natural approaches to this problem. The first approach is to *ignore the jump by weakening the topology*. For example, we could obtain convergence to the zero function if we defined convergence to mean pointwise convergence or pointwise convergence at almost all t (with respect to Lebesgue measure). Alternatively, with the familiar L^p norm on $[0, 1]$, i.e.,

$$\|x\|_p = \left(\int_0^1 |x(t)|^p dt\right)^{1/p}, \qquad (1.2)$$

we have $\|x_n - x\|_p \rightarrow 0$ as $n \rightarrow \infty$ for any p, $0 < p < \infty$, where $x(t) = 0$, $0 \leq t \leq 1$. This approach can also be applied to our MONQ example.

The second approach is to *focus on the jump by enlarging the space of possible limits* (and changing the topology as well). This can be done by identifying each function x in D with the closure of its graph, i.e.,

$$\Gamma_x = \{(t, r) : x(t) = r \quad \text{or} \quad x(t-) = r, 0 \leq t \leq 1\}, \qquad (1.3)$$

with $x(0-)$ not defined, as in Pomarede (1976). Over closed bounded time intervals, the graphs are compact subsets of R^2. We can then use the Hausdorff metric on the space of compact subsets of $[0, 1] \times R$, as is often done in the theory of random sets, i.e.,

$$m(A, B) = \max\{\sup_{x \in B} \inf_{y \in A} \|x - y\|, \sup_{y \in A} \inf_{x \in B} \|x - y\|\}; \qquad (1.4)$$

see page 15 of Matheron (1975). With the Hausdorff metric, it is easy to see that we have convergence $m(\Gamma_{x_n}, \Gamma) \rightarrow 0$ as $n \rightarrow \infty$, where

$$\Gamma = \{(t, 0) : 0 \leq t \leq 1\} \cup \{(\tfrac{1}{2}, \tfrac{1}{2})\}. \qquad (1.5)$$

Again, this same approach works for our MONQ example. We believe that with these approaches it will be possible to establish nondegenerate heavy-traffic limits for general MONQs, but the task seems challenging.

Organization of the Paper

Here is how the rest of this paper is organized. In §2 we introduce our basic deterministic four-class two-queue MONQ example. It has both queues critically stable; i.e., the traffic intensity at queue i is $\rho_i = 1$ for each i. In §3 we describe the transient behavior of this example. In §4 we consider variants of the basic model that are stable. There we study the heavy-traffic behavior of a sequence of associated stable MONQ's indexed by n, with $\rho_{ni} < 1$ for all n but $\rho_{ni} \rightarrow 1$ as $n \rightarrow \infty$. In §5 we discuss the variant of the basic model with class-dependent priorities.

These examples are admittedly very special, but minor perturbations of the models satisfy standard assumptions, so that the observed behavior can occur in conventional stochastic settings. For example, it is possible to construct related stochastic renewal arrival processes by letting each arrival not come with some small probability, where successive arrivals are treated independently. Thus, we conclude that we are describing important phenomena, which can occur in real systems.

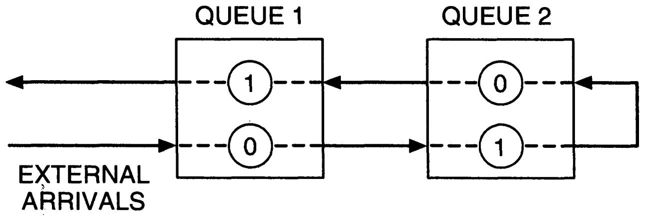
2. The Basic Deterministic Example

Our basic example consists of two queues, each with single server, unlimited waiting space and the first-in first-out (FIFO) service discipline. All customers follow the route (1, 2, 2, 1) with the class changing after each service completion, as depicted in Figure 1. Thus, there are four customer classes, two associated with queue 1 and two associated with queue 2. We shall refer to the four classes as classes 0 and 1 at queue 1, and classes 0 and 1 at queue 2. The service times are all deterministic, with the vector of successive service times being (0, 1, 0, 1); i.e., at both queues, the service times are j for class j (where j is 0 or 1).

There is an external arrival process only for class 0 at queue 1. Customers of class 0 at queue 1 arrive exogenously at every positive integer time point; i.e., the interarrival times are 1. Each customer of class 0 at queue 1 after completing service goes to queue 2 as a class 1 customer. After completing service, each class 1 customer at queue 2 becomes a class 0 customer at queue 2. After completing service, each class 0 customer at queue 2 becomes a class 1 customer at queue 1. Finally, each class 1 customer at queue 1 leaves the network after completing service.

Notice that all events take place at integer time points. Since we may have multiple transitions at each integer, we must specify the order for events scheduled for the same epoch. For this example, this scheduling can make quite a difference. We first admit the new arrival and move him to queue 2 if queue 1 is empty. Then we perform services at queue 2 and finally we perform services at queue 1. Since no customer has two consecutive 0 service times, we need go no further.

Figure 1 The Class Transition Diagram



We shall look at the system at integer time points after all transitions at that time have been made. The state of the system at any such time can be represented by an ordered pair of finite (possibly empty) sequences of positive integers. The pair

$$((n_{11}, n_{12}, \ldots, n_{1k}), (n_{21}, n_{22}, \ldots, n_{2m}))$$

denotes that queue 1 contains $n_{11} + n_{12} + \cdots + n_{1k}$ customers, with the first n_{11} being class 1, the next n_{12} being class 0, the next n_{13} being class 1 and so forth, while queue 2 contains $n_{21} + n_{22} + \cdots + n_{2m}$ customers, with the first n_{21} being class 1, the next n_{22} being class 0 and so forth. There are k such groups of consecutive 1's or 0's in queue 1, while there are m in queue 2. Let ϕ denote an empty queue.

Let $X_i(n)$ be the state of queue i at time n and let $X(n) = (X_1(n), X_2(n))$ be the state of the entire system. Note that the sequence of future states $\{X(k) : k \geq n\}$ is completely determined by the present state $X(n)$ for any n and any $X(n)$. Hence, we have an appropriate state description.

We shall also be interested in the numbers of each class at each queue. Let $Q_i^j(n)$ be the number of class j customers at queue i after all the events at time n. Let $Q_i(n) = Q_i^0(n) + Q_i^1(n)$, $Q^j(n) = Q_1^j(n) + Q_2^j(n)$ and $Q(n) = Q_1(n) + Q_2(n) = Q^0(n) + Q^1(n)$.

3. The Transient Behavior of the Basic Example

In this example it is natural to define the traffic intensity ρ_i at queue i as the total expected service requirement at queue i per external arrival (since the arrival rate is 1). Since each arrival requires exactly 1 unit of service at each queue (during his incarnation as a class 1 customer at that queue), we have $\rho_1 = \rho_2 = 1$. For a single queue with a work-conserving discipline, the traffic intensity $\rho_i = 1$ is typically the critical level for stability; i.e., for $\rho_i > 1$ the number in system typically goes to infinity w.p.1 as $t \to \infty$, while for $\rho_i < 1$ the number in system typically converges in distribution to a proper limit. Thus, in this example the queues are at least nominally at the critical value for stability and so we say that the MONQ is in heavy traffic.

We first note that the number in system is nondecreasing, increasing at most by 1 each transition.

LEMMA 3.1. *For all n and all initial conditions, $Q(n) \le Q(n+1) \le Q(n) + 1$.*

PROOF. There is one class 0 arrival to queue 1 every transition. There is a departure from the network of at most one class 1 customer from queue 1. This occurs if and only if queue 1 was previously nonempty. □

We next note that it is possible for the system to reach a fixed point or a fixed cycle. For any finite string of numbers $n_1, n_2, \ldots, n_k$, let $\{n_1, n_2, \ldots, n_k\}_m$ denote the finite string obtained by concatenating m identical copies of the given string; e.g., $(\{1, 2\}_3, 1, 1)$ denotes the sequence $(1, 2, 1, 2, 1, 2, 1, 1)$.

PROPOSITION 3.1. *If $X(0) = ((m), (\{m, m\}_k, m))$ for any $k \ge 1$ and $m \ge 1$, then $X(nm) = X(0)$ for all n.*

PROOF. Easy to verify. □

Proposition 3.1 can be interpreted as demonstrating that the arrival rate 1 is the *critical arrival rate for stability* in this example. To formulate this notion precisely in general, we need to define a family of models indexed by the arrival rate. For this purpose, let a reference collection of external arrival processes $\{A_j(t) \colon t \ge 0\}$ be given, with $A_j(t)$ counting the number of class j arrivals in the interval $[0, t]$. Without loss of generality (by choosing the time units), let the total arrival rate be 1; i.e.,

$$\sum_j A_j(t)/t \to 1 \quad \text{as} \quad t \to \infty. \tag{3.1}$$

Then define the system with arrival rate λ by using arrival processes $A_{\lambda j}$ defined by simply scaling time by λ; i.e.,

$$A_{\lambda j}(t) = A_j(\lambda t), \quad t \ge 0, \tag{3.2}$$

and otherwise keeping the model fixed.

In this setting, we say that a *sufficient condition* for λ^* to be a critical arrival rate for stability is to have, for each positive number l, a number m and initial conditions such that

$$l \le Q_{\lambda^*}(t) \le l + m \quad \text{for } all\ t, \tag{3.3}$$

where $Q_\lambda(t)$ is the number in system with arrival rate λ.

It remains to show that if such a critical arrival rate exists, then it is unique and that the customary stability and instability hold for $\lambda < \lambda^*$ and $\lambda > \lambda^*$; e.g., see §1 of Heyman and Whitt (1984). Here we only establish (3.3). Note that Proposition 3.1 implies that $Q(n) = 2k + 2$ for all n when $m = 1$, so the arrival rate 1 is indeed a critical arrival rate for stability according to (3.3). We *conjecture* that $\{Q_\lambda(t) : \ge 0\}$ is tight (see Billingsley 1968) for all $\lambda < 1$ and that $Q_\lambda(t) \to \infty$ as $t \to \infty$ for all $\lambda > 1$ in this example.

We now show that the number in system can grow without bound with the given unit arrival rate. From Lemma 3.1 and its proof, we know that this will occur if and only if queue 1 is empty infinitely often.

PROPOSITION 3.2. *If $X(0) = ((n), (n+1))$ for any $n \ge 1$, then $X(k)$ evolves as described in Table 1, with $X(5n+6) = ((n+1), (n+2))$.*

PROOF. It is easy to verify the successive steps in Table 1. □

REMARK 3.1. Proposition 3.2 has an extension to the model in which batches of size m arrive every m units of time. If $X(0) = ((nm), (nm + m))$, then

$$X(5nm + 6m) = ((nm + m), (nm + 2m)). \quad \square$$

We now consider what happens when we start empty. Let $T(((n-1), (n)))$ be the first time that the system state reaches $((n-1), (n))$ and let $T(n)$ be the first time that the total number in the system reaches n.

PROPOSITION 3.3. *If $X(0) = (\varnothing, \varnothing)$, then $X(7) = ((1), (2))$, after which the evolution is described by Table 1. As a consequence,*

Table 1 The Evolution from State $((n), (n+1))$ to State $((n+1), (n+2))$

Time	State	Time	State
0	$((n), (n+1))$	$3n+3$	$((n+2), (n+1))$
$n-1$	$((1, n-1), (2, n-1))$	$4n+3$	$((2, n), (1, n))$
n	$(\varnothing, (1, n, n))$	$4n+4$	$((1, n+1, n+1), \varnothing)$
$n+1$	$((n), (n+1, 1))$	$4n+5$	$((n+1, 1), (n+1))$
$2n$	$((1, n-1), (2, n))$	$5n+5$	$((1, n+1), (1, n))$
$2n+1$	$(\varnothing, (1, n+1, n))$	$5n+6$	$((n+1), (n+2))$
$2n+2$	$((n+1), (n+1, 1))$		
$3n+2$	$((1, n), (1, n+1))$		

Table 2 The First Eleven States Starting out Empty

Time	State	Time	State
0	$(\varnothing, \varnothing)$	6	$((1, 1), (1))$
1	$(\varnothing, (1))$	7	$((1), (2))$
2	$(\varnothing, (1, 1))$	8	$(\varnothing, (1, 1, 1))$
3	$((1), (1, 1))$	9	$((1), (2, 1))$
4	$((2), (1))$	10	$(\varnothing, (1, 2, 1))$
5	$((1, 1, 1), \varnothing)$	11	$((2), (2, 1))$

(a) $Q(n) \to \infty$ *as* $n \to \infty$;
(b) $T(((n-1), (n))) = (5n^2 - 3n)/2$ *for all* $n \geq 2$;
(c) $T(2n) = (5n^2 - n)/2$ *for all* $n \geq 1$;
(d) $|Q(n) - 2\sqrt{2n/5} - 1| \leq 4$ *for all* n;
(e) $Q(n)/\sqrt{n} \to 2\sqrt{2/5}$ *as* $n \to \infty$,
(f) $Q(nt)/\sqrt{n} \to 2\sqrt{2t/5}$ *as* $n \to \infty$ *uniformly for* t *in compact intervals*.

PROOF. It is easy to check that the first eleven states are as in Table 2. Then apply Proposition 3.2 starting with $X(7) = ((1), (2))$. Hence,

$$T(((n-1), (n))) = 7 + \sum_{k=1}^{n-2} (5k + 6) = \frac{5n^2 - 3n}{2}.$$

From Table 1, note that $Q(k) = 2n$ for the first time exactly n transitions after reaching the state $((n-1), (n))$, and thus (c) is true. By (c), $|Q(k) - 2n - 1| \leq 1$ for

$$\frac{5n^2 - n}{2} \leq k \leq \frac{5(n+1)^2 - (n+1)}{2}.$$

Since $(5n^2 - n)/2 \geq 5(n-1)^2/2$ and $[5(n+1)^2 - (n+1)]/2 \leq 5(n+1)^2/2$,

$$\frac{5(n-1)^2}{2} \leq k \leq \frac{5(n+1)^2}{2}$$

and

$$2\sqrt{\frac{2k}{5}} - 1 \leq 2n + 1 \leq 2\sqrt{\frac{2k}{5}} + 3$$

so that (d) holds. It is easy to see that (e) and (f) follow from (d). □

From Proposition 3.3 and Tables 1 and 2, we see that the total number in system grows smoothly when we start out empty, but the number in each queue does not. From Table 1 we see that during the nth growth cycle (going from state $((n), (n+1))$ to state $((n+1), (n+2))$ the number in queue 1 experiences sudden jumps down and immediately back up again at times n (from time $n-1$ to time $n+1$) and $2n+1$, while queue 2 does similarly at time $4n+4$. Thus the analog of (f) for the individual queue lengths does *not* hold. We will explore these sudden large fluctuations further in the context of stable models in §4.

REMARK 3.2. As mentioned earlier, one of the characteristic properties of deterministic chaos is that a minor perturbation of the initial condition leads to dramatically different behavior. It is significant that a discrete analog of this phenomenon is observed here. For example, by Proposition 3.1, $((1), (1, 1, 1))$ is a fixed point, but from the "nearby" states $s_1 \equiv ((1), (1, 1))$ and $s_2 \equiv (\varnothing, (1, 1, 1))$ the number in system grows without bound. Indeed, states s_1 and s_2 are easily seen to be the third and eighth states starting from $(\varnothing, \varnothing)$, i.e., totally empty; see Table 2. Moreover, these examples show a lack of monotonicity: More initial customers does not mean larger queue lengths in the future. □

REMARK 3.3. The sojourn times behave similarly to the queue lengths. It is easy to see that the sojourn time in the network for an external arrival at time n is $Q(n) + 1$, so it is relatively stable. However, the waiting time before beginning service at each queue is obviously the number of class-1 customers ahead of him in the queue upon his arrival (which may depend strongly on his position in a batch). Thus, the waiting times before beginning service at the individual queues exhibit large fluctuations too.

REMARK 3.4. It is instructive to compare the FIFO discipline with the head-of-the-line processor-sharing discipline (HOL-PS), which is known to have advantages in MONQ's; e.g., see Demers, Keshav and Shenker (1989) and Fendick and Rodrigues (1991). With the HOL-PS discipline at each queue, the service rate is shared among the first customers in each class at each queue if customers from both classes are present. For our model, this is equivalent to immediate service for class 0 customers at each queue. Thus, our model with HOL-PS is equivalent to the FIFO network in which each customer goes first to queue 2 and then to queue 1, having a service time of 1 at each queue. Thus, with

HOL-PS, if $Q_1^1(0) = k_1$ and $Q_2^1(0) = k_2$, then $Q_1(n) = k_1$ and $Q_2(n) = k_2$ for all $n \geq 1$, so that all irregular behavior is removed. □

4. Associated Stable Models

The model analyzed in §§2 and 3 is unstable. We now discuss its implications for stable models. In this section we indicate two ways to construct stable models from the unstable model.

Our first stable model is constructed by adding a third FIFO queue with service times 1 and an arbitrary arrival process with arrival rate $\lambda < 1$. Let all external arrivals come to this new queue before they go to the rest of the network, but otherwise let the rest of the network be the same. During the busy periods of the new queue, the rest of the network obviously has the transient behavior of the unstable deterministic model just described (with unspecified initial conditions).

In our second approach, we construct a sequence of stable models. Let a subscript n index the models. In the nth model, let the initial state be $((n), (n+1))$. We make adjustments so that the nth growth cycle in Table 1 ends at state $((n), (n+1))$ instead of $((n+1), (n+2))$, so that the cycle repeats. First, at time $5n+6$ we assume that there is no arrival, which is equivalent to assuming that the interarrival time after time $5n+5$ is 2 instead of 1. Next we assume that the arrival at time $3n+4$ has 0 service time as a class 1 customer at queue 1, while arrival $4n+5$ has 0 service time as a class 1 customer at queue 2. Since at time $5n+5$, the second customer in queue 1 is arrival $4n+5$, while the second customer is queue 2 is arrival $3n+4$, we transition to $((n), (n+1))$ instead of to $((n+1), (n+2))$ at time $5n+6$. We then let the cycle repeat. We can make the model stationary by letting the initial state be uniformly distributed over the $5n+6$ states associated with the cycle, but we will assume that we start in state $((n), (n+1))$ at time 0. The important point is that this model is stable; it is periodic with period $5n+6$.

Due to the adjustments, the interarrival times are not i.i.d., and the service times are not independent of the arrival process in the stationary model construction. However, the model still seems to have the same character. Due to the adjustments, in the nth model the arrival rate is $(5n+5)/(5n+6)$, while the average service requirement per external arrival at each queue is $(5n+4)/(5n+5)$. Hence, the traffic intensity at queue i in model n is

$$\rho_{ni} = (5n+4)/(5n+6), \qquad (4.1)$$

so that $1 - \rho_{ni}$ is of order n^{-1}.

One might object to the initial condition $((n), (n+1))$, but it is essentially the steady-state behavior associated with model n, so we believe it is natural.

We now describe the asymptotic behavior of the sequence of models as $n \to \infty$. For this purpose, let $\lfloor x \rfloor$ be the greatest integer less than or equal to x. The following is an easy consequence of Table 2. Now we use the Hausdorff metric in (1.4).

PROPOSITION 4.1. *Consider the models above indexed by n.*

(a) $|Q_n(k) - 2n + 2| \leq 2$ *for all k, so that*

$$Q_n(\lfloor nt \rfloor)/n \to 2 \quad \text{as } n \to \infty \text{ uniformly in } t.$$

(b) $Q_{ni}^j(\lfloor nt \rfloor)/n \to X_i^j(t)$ *as $n \to \infty$ in the Skorohod* (1956) *J_1 topology, where*

$$X_1^0(t) = X_2^0(t) = 1 - X_1^1(t) = 1 - X_2^1(t) = t - \lfloor t \rfloor.$$

(c) $Q_{ni}(\lfloor nt \rfloor)/n \to 1$ *as $n \to \infty$ uniformly for all t in compact sets not containing an integer, but fails to converge in $D[0, \infty)$ with any of the Skorohod* (1956) *topologies.*

(d) *For $t \in [0, 5k]$, the graphs of $Q_{n1}(\lfloor nt \rfloor)/n$ converge in the Hausdorff metric as $n \to \infty$ to the set*

$$\{(t, 1) : 0 \leq t \leq 5k\} \cup \{(5j+1, 0), (5j+2, 0), (5j+4, 2), j = 0, 1, \ldots, k-1\}.$$

(e) *For $t \in [0, 5k]$, the graphs of $Q_{n2}(\lfloor nt \rfloor)/n$ converge in the Hausdorff metric to the set*

$$\{(t, 1) : 0 \leq t \leq 5k\} \cup \{(5j+1, 2), (5j+2, 2), (5j+4, 0), j = 0, 1, \ldots, k-1\}.$$

Part (c) shows that we fail to get a functional central limit theorem (FCLT) for this example with the usual topology. It may seem to rule out a functional law of large numbers (FLLN) too, but it does not, because the proper normalizations for the FCLT and FLLN are $Q_{ni}(n^2t)/n$ and $Q_{ni}(n^2t)/n^2$, respectively, when $1 - \rho_{ni} = O(n^{-1})$. Part (c) rules out this FCLT, but not the

FLLN. Part (c) also shows how we can obtain convergence to a limit in D (in fact continuous) by weakening the topology. Parts (d) and (e) show how we can obtain convergence to a limit that reflects the oscillations by enlarging the space of prospective limits and changing the topology.

5. Class-Dependent Priorities

It is also instructive to compare the FIFO discipline with class-dependent priorities at the queues. If class 0 has priority at each queue, then the irregularity is eliminated, as in Remark 3.4. However, if class 1 has priority at each queue, then the behavior is even worse, as in Kumar and Seidman (1990), Erramilli and Forys (1991) and Kumar (1993). Then the critical arrival rate for stability is evidently $\frac{1}{2}$ instead of 1. It is significant that the critical traffic intensity for stability depends on the queue discipline.

The following is an analog of Proposition 3.1 previously established by Kumar and Seidman. We find a fixed cycle with arrival rate $\frac{1}{2}$. This result shows that $\lambda = \frac{1}{2}$ is a critical arrival rate for stability according to (3.3) when the FIFO discipline is replaced by appropriate class-dependent priorities.

PROPOSITION 5.1. *Consider the MONQ example in §2, modified by having interarrival times of 2 instead of 1 and class-dependent priorities at the queues. Let class 1 have high priority at each queue. If* $X(0) = (\varnothing, n)$, *then* $X(2n) = ((2n), \varnothing)$ *and* $X(4n) = (\varnothing, (n))$.

On the other hand, if the arrival rate is 1, then there are no departures at all when we start out empty, as is implied by the following result.

PROPOSITION 5.2. *Consider the MONQ example in Proposition 5.1, but with arrival rate 1. If* $X(0) = (\varnothing, \varnothing)$, *then* $X(n) = (\varnothing, (1, n-1))$ *for all* n.

Proposition 5.2 shows that it is not easy to deduce the critical arrival rate for stability by observing the output rate associated with an unstable arrival rate. In Proposition 5.2, the external arrival rate is 1 but the departure rate is 0. We might think that the critical arrival rate for stability must be 0.[1]

[1] I thank Jim Dai and Mike Harrison for bringing my attention to Dai and Wang (1993), which was the primary motivation for this work; I thank Jim Dai and Larry Wein for bringing my attention to the related work by P. R. Kumar and his colleagues; and I thank Jeff Lagarias, Steve Pincus and the referees for helpful comments on the presentation.

References

Billingsley, P., *Convergence of Probability Measures*, Wiley, New York, 1968.

Bramson, M., "Instability of FIFO Queueing Networks," Mathematics Department, University of Wisconsin, Madison, WI, 1993.

Dai, J. G. and V. Nguyen, "On the Convergence of Multiclass Queueing Networks in Heavy Traffic," *Ann. Appl. Probab.* 4 (1994), to appear.

—— and Y. Wang, "Nonexistence of Brownian Models of Certain Multiclass Queueing Networks," *Queueing Systems* 13 (1993), 41–46.

Demers, A., S. Keshav and S. Shenker, "Analysis and Simulation of a Fair Queueing Algorithm," *Proc. ACM SIGCOMM* (1989).

Erramilli, A. and L. J. Forys, "Oscillations and Chaos in a Flow Model of a Switching System," *IEEE J. Sel. Areas Commun.*, 9 (1991), 171–178.

Fendick, K. W. and M. A. Rodrigues, "A Heavy-Traffic Comparison of Shared and Segregated Buffer Schemes for Queues with the Head-of-Line Processor Sharing Discipline," *Queueing Systems*, 9 (1991), 163–190.

Harrison, J. M., "Brownian Models of Queueing Networks with Heterogeneous Customer Populations," *Proceedings of the IMA Workshop on Stochastic Differential Equations*, W. Fleming and P. L. Lions (Eds.), Springer, New York, 1988, pp. 147–186.

—— and V. Nguyen, "The QNET Method for Two-Moment Analysis of Open Queueing Networks," *Queueing Systems*, 6 (1990), 1–32.

—— and ——, "Brownian Models of Multiclass Queueing Networks: Current Status and Open Problems," *Queueing Systems* 13 (1993), 5–40.

Heyman, D. P. and W. Whitt, "The Asymptotic Behavior of Queues with Time-Varying Arrival Rates," *J. Appl. Probab.*, 21 (1984), 143–156.

Johnson, D. P., *Diffusion Approximations for Optimal Filtering of Jump Processes and for Queueing Networks*, Ph.D. Dissertation, University of Wisconsin, Madison, WI, 1983.

Kumar, P. R., "Re-entrant Lines," *Queueing Systems*, 13 (1993), 87–110.

—— and T. I. Seidman, "Dynamic Instabilities and Stabilization Methods in Distributed Real-Time Scheduling of Manufacturing Systems," *IEEE Trans. Aut. Control*, 35 (1990), 289–298.

Laws, C. N., "Resource Pooling in Queueing Networks with Dynamic Routing," *Adv. Appl. Probab.* 24 (1992), 699–726.

Lu, S. H. and P. R. Kumar, "Distributed Scheduling Based on Due

Dates and Buffer Priorities," *IEEE Trans. Aut. Control*, 36 (1991), 1406–1416.

Matheron, G., *Random Sets and Integral Geometry*, Wiley, New York, 1975.

Peterson, W. P., "A Heavy Traffic Limit Theorem for Networks of Queues with Multiple Customer Types," *Math. Oper. Res.*, 16 (1991), 90–118.

Pomarede, J. L., *A Unified Approach Via Graphs to Skorohod's Topologies on the Function Space D*, Ph.D. Dissertation, Department of Statistics, Yale University, 1976.

Reiman, M. I., "A Multiclass Feedback Queue in Heavy Traffic," *Adv. Appl. Probab.*, 20 (1988), 179–207.

Rybko, A. N. and A. L. Stolyar, "Ergodicity of Stochastic Processes Describing the Operation of Open Queueing Networks," *Problems of Information Transmission* 28 (1992), 199–220.

Schuster, H. G., *Deterministic Chaos*, VCH Verlagsgesellschaft, Weinheim, 1988.

Segal, M. and W. Whitt, "A Queueing Network Analyzer for Manufacturing," in *Teletraffic Science for New Cost-Effective Systems, Networks and Services, ITC*-12, M. Bonatti (Ed.), North-Holland, Amsterdam, 1989, pp. 1146–1152.

Seidman, T. I., "First Come, First Served Is Unstable!" Department of Mathematics, University of Maryland Baltimore County, 1993.

Sharifnia, A., "Instability of Some Distributed Scheduling Policies in Manufacturing Systems and Their Stabilization," *IEEE Trans. Aut. Control* (1993), to appear.

Skorohod, A. V., "Limit Theorems for Stochastic Processes," *Theor. Probab. Appl.*, 1 (1956), 261–290.

Whitt, W., "Some Useful Functions for Functional Limit Theorems," *Math. Oper. Res.*, 5 (1980), 67–85.

Accepted by Linda Green; received March 20, 1992. This paper has been with the author 1 1/2 months for 2 revisions.