WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

Some Badly Behaved Closed Queueing Networks

J. Michael Harrison
Vien Nguyen

#3691-94-MSA

May 1994

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139
Some Badly Behaved Closed Queueing Networks

J. Michael Harrison
Vien Nguyen

#3691-94-MSA

May 1994
SOME BADLY BEHAVED CLOSED QUEUEING NETWORKS

J. MICHAEL HARRISON AND VIÊN NGUYEN*

Abstract. We study several variations of a multiclass closed queueing network model, all closely related to the open priority network of Lu and Kumar. For each of our examples it is shown that, as the population size $n \to \infty$, the long-run utilization rate of each server approaches a limit strictly less than one. Thus no "bottleneck station" emerges in the heavy traffic limit. This is analogous to the "bad behavior" of open network examples developed by Bramson and by Lu and Kumar, among others, in which queue sizes grow without bound even though each service station has a traffic intensity parameter less than one. Two of the three examples considered here have preemptive-resume priority service disciplines at each station, as in the Lu-Kumar open network example, and one has FIFO service at each station, as in Bramson's open network examples.

Key words. Closed queueing networks, multiclass networks, priority networks, bottlenecks

AMS(MOS) subject classifications. 90B15, 60K25, 60K20, 60K05

1. Introduction. Over the last few years queueing theorists have been shocked to discover that there exist open network models which are unstable even though each service station has more than enough capacity to handle the load imposed on it. Perhaps the simplest example is the two-station priority network pictured in Figure 1.1, which was originally studied by Lu and Kumar [3]. Here customers arrive according to a renewal process at average rate $\lambda$, and they follow a fixed deterministic route requiring four service operations. The four operations are performed at

* Graduate School of Business, Stanford University, Stanford, CA 94305; and Sloan School of Management, M.I.T., Cambridge, MA 02139

![Figure 1.1. The Lu-Kumar priority network](image-url)
stations 1, 2, 2 and 1 successively. Customers that are waiting for or undergoing their $k^{th}$ operation will be called class $k$ customers. Each class is allowed to have its own service time distribution (thus there are four mutually independent service time sequences, also independent of the arrival process), and we denote by $m_k$ the mean service time for class $k$. Finally, there is a preemptive-resume priority service discipline at each station, the priorities being as shown in Figure 1.1.

In their original paper, Lu and Kumar [3] treated a version of this model with deterministic interarrival and service times. Assuming specific numerical values for $\lambda, m_1, \ldots$ and $m_4$, they showed that queue lengths can grow without bound in the deterministic priority network even though

$$\rho_1 \equiv \lambda(m_1 + m_4) < 1 \quad \text{and} \quad \rho_2 \equiv \lambda(m_2 + m_3) < 1. \quad (1.1)$$

Independently of this work, Rybko and Stolyar [4] analyzed the first legitimately stochastic network model shown to be unstable with traffic intensity parameters less than one. In an important recent paper, Dai and Weiss [2] studied the stochastic version of the Lu-Kumar model described above, and discovered the following additional condition:

$$\lambda(m_2 + m_4) < 1. \quad (1.2)$$

Dai and Weiss [2] showed that (1.1) and (1.2) together are sufficient for stability (that is, positive Harris recurrence) of the stochastic Lu-Kumar network. They also showed that (1.2) is necessary for the network’s deterministic fluid analog to be stable, which strongly suggests that (1.2) is necessary for stability of the queueing network as well.

It may be argued that “legitimate” networks can be made unstable by applying a “bad” priority scheme, such as those used in the Lu-Kumar and Rybko-Stolyar examples. Bramson [1] recently presented the surprising result of a multiclass network that is unstable even when its intensity parameters are less than one and stations serve customers on a FIFO basis. Whitt [7] and Seidman [6] also studied unstable deterministic networks with FIFO service.

Can a closed queueing network exhibit “bad behavior” analogous to the instability described above for open networks? In this paper we answer that question in the affirmative, as follows. We consider several closely related two-station closed networks, denoting by $n$ the fixed population size, and show that for each of the examples considered, the long-run utilization rates of both servers approach limits strictly less than one as $n \to \infty$. Thus no “bottleneck station” emerges as the population grows in size, although the existence of such a bottleneck has often been implicitly assumed in theoretical analyses of closed networks [5]. Readers should note that our “badly behaved” closed network examples are true multiclass networks, with different classes served at the same station having different mean service times, whereas most previous work on both open and closed networks
has assumed a single service time distribution for all classes served at a station.

To be more specific, we will focus on the closed version of the Lu-Kumar model pictured in Figure 1.2, with two single-server stations, four customer classes, and the deterministic cyclic routing shown. Two very simple examples are treated in the next section, each characterized by extreme distributional assumptions (some service times are small and bounded above, others are large and bounded below), which eliminate the need for sophisticated analysis. The first of our elementary examples uses the Lu-Kumar preemptive-resume priority scheme (see Figure 1.1), and our second example shows that essentially the same system behavior may be observed with FIFO scheduling, at least if one allows some service times to vanish as \( n \to \infty \).

In section 3 we study a precise closed analog of the Lu-Kumar open network, with preemptive-resume priority scheduling at each station, with all four service time distributions assumed to be exponential, and with mean service times assumed to satisfy

\[
(1.3) \quad m_2 + m_4 > \max(m_1 + m_4, m_2 + m_3).
\]

Thus our final example, unlike those treated in section 2, is not characterized by exotic distributional forms or extreme parameter values, although one realizes after the fact that the priority rule assumed is a rather bad one. The limiting long-run average throughput rate (as \( n \to \infty \)) is explicitly calculated, and this corresponds to less-than-full utilization of each server.

2. Two simple examples.

Priority service. With the routing picture in Figure 1.2, first assume that classes 4 and 2 have preemptive-resume priority at stations 1 and 2, respectively. Let \( \{v_k(i), i = 1, 2, \ldots \} \), \( k = 1, 2, 3, 4 \), be independent sequences of i.i.d non-negative random variables representing processing times of class \( k \) job. We suppose that for some \( \delta > 0 \) the following inequalities hold al-
most surely:

\[
\begin{align*}
    v_2(1) &> \delta & v_4(1) &> \delta \\
    v_1(1) &< \delta & v_3(1) &< \delta.
\end{align*}
\]

These bounds imply (recall \( m_k \equiv E[v_k(1)] \))

\[
\max(m_1, m_3) < \min(m_2, m_4).
\]

Suppose that all \( n \) customers are queued initially at station 1 as class 1 (i.e., \( Q_1(0) = n \) and \( Q_2(0) = Q_3(0) = Q_4(0) = 0 \)), so that station 1 contains only low priority work. Set

\[
\begin{align*}
    T_0 &= \inf\{t \geq 0 : Q_1(t) = n\} = 0 \quad \text{and} \\
    T_{i+1} &= \inf\{t \geq T_i + \delta : Q_1(t) = n\}.
\end{align*}
\]

It follows from the bounds on service times (2.1) that while station 1 is processing low priority work, each class 1 customer will depart station 1 and enter station 2 before the previous customer can complete service at station 2. Consequently, assuming that station 1 is not interrupted by arrival of a high priority (class 4) customer, station 2 will be continuously busy with class 2 work after the initial service of a class 1 customer. Let \( \tau \) be the sum of the first class 1 service at station 1 and the first \( n \) class 2 services at station 2. Because of the preemptive-resume priority policy at station 2, no class 4 customers are generated between \( T_0 \) and \( T_0 + \tau \) and so at time \( T_0 + \tau \) all \( n \) customers are queued at station 2 as low priority class 1 work.

The network now evolves in exactly a symmetrical manner as before. Services of high priority (class 4) customers at station 1 begin after the initial service of a class 3 customer, and station 1 continuously serves class 4 until all \( n \) customers have been served, at which time the entire population is again queued at station 1 as class 1.

It should be clear from this discussion that

\[
E[T_1] = n(m_2 + m_4) + (m_1 + m_3).
\]

Moreover, \( T_1, T_2, \ldots \) form regeneration points for the queue length process \( Q(t) = [Q_1(t), Q_2(t), Q_3(t), Q_4(t)] \). Denoting by \( \lambda \) be the long-run throughput rate of customers, it follows from the theory of regenerative processes that

\[
\lambda = \frac{E[\text{number of customers exiting station 1 between } T_0 \text{ and } T_1]}{E[T_1]}
\]

\[
(2.5) = \frac{n}{n(m_2 + m_4) + (m_1 + m_3)},
\]

and we arrive at the main result of this subsection:

\[
\lambda \to \frac{1}{m_2 + m_4} \quad \text{as } n \to \infty.
\]
Equations (2.2) and (2.6) together imply that the long-run utilization rate for station 1 is given by

$$\lambda(m_1 + m_4) \rightarrow \frac{m_1 + m_4}{m_2 + m_4} < 1 \quad \text{as} \quad n \to \infty,$$

and similarly for station 2, the long-run utilization rate is

$$\lambda(m_2 + m_3) \rightarrow \frac{m_2 + m_3}{m_2 + m_4} < 1 \quad \text{as} \quad n \to \infty.$$

**FIFO service.** We now consider a FIFO version of the network discussed in the previous subsection. The routing of customers remains exactly the same as shown in Figure 1.2 and customers are served on a FIFO basis at each station. Using the same notation as before, let us now impose the following bounds on service times:

\[(2.7)\]

$$v_2(1) > \delta \quad v_4(1) > \delta \quad v_1(1) < \delta/n \quad v_3(1) < \delta/n,$$

which of course implies $$\max(m_1, m_3) < \min(m_2, m_4)$$. We will find it convenient to refer to classes 1 and 3 as customers requiring “short” services and classes 2 and 4 as those requiring “long” services. Let us again assume initial conditions $$Q_1(0) = n$$ and $$Q_2(0) = Q_3(0) = Q_4(0) = 0$$; that is, all customers are initially queued at station 1 as class 1 awaiting a short service. From (2.7), one can verify that station 1 will finish serving all n class 1 customers (that is, all n customers will arrive at station 2) before station 2 completes its first long service. Thus all n class 2 services must be completed at station 2 before the first class 3 service can begin. Defining $$\tau$$ exactly as above, we find that at time $$\tau$$ all n customers are queued at station 2 as class 3. One may then argue exactly as above to conclude that (2.5) remains valid for a single FIFO system satisfying (2.7). In letting $$n \to \infty$$, we may keep the service time distributions for classes 2 and 4 (the long service operations) fixed, but (2.7) demands that service times for classes 1 and 3 vanish, implying that $$m_1, m_3 \to 0$$. Thus (2.6) remains valid in the current setting, and the long-run utilization rates at stations 1 and 2, respectively, converge as $$n \to \infty$$ to

$$\frac{m_4}{m_2 + m_4} < 1 \quad \text{and} \quad \frac{m_2}{m_2 + m_4} < 1.$$

3. A closed version of the Lu-Kumar priority network. With the routing portrayed in Figure 1.2, let us return to the assumption of preemptive-resume priority scheduling, with classes 4 and 2 having priority at stations 1 and 2, respectively, as in the Lu-Kumar open network (see Figure 1.1). Also, let all four service time distributions be exponential. With mean service times satisfying (1.3) we can (and do) choose units so that

\[(3.1)\]

$$m_2 + m_4 = 1 \quad \text{but} \quad m_1 + m_4 < 1 \quad \text{and} \quad m_2 + m_3 < 1.$$
With exponentially distributed service times and preemptive resume priority scheduling the queue length process $Q(t) = [Q_1(t), Q_2(t), Q_3(t), Q_4(t)]$ is a Markov process with finite state space, but some of its states are transient, as explained below. Suppose that both $Q_2(0) > 0$ and $Q_4(0) > 0$. Then both servers will work on priority customers at their respective stations over an initial interval $[0, \tau]$, at the end of which either $Q_2(\tau) = 0$ or $Q_4(\tau) = 0$. Moreover, no new priority customers are created during this initial period, so it is immediate that $E(\tau) < \infty$.

Suppose for the sake of concreteness that $Q_2(\tau) > 0$ and $Q_4(\tau) = 0$. Then server 2 will work on (priority) customers of class 2 over an interval $(\tau, \tau']$, during which period no new customers of class 4 can be created. Thus server 1 is either working on class 1 or else idle throughout the period, and at the end we have $Q_2(\tau') = Q_4(\tau') = 0$. The evolution of the system after time $\tau'$ can be divided into alternating “idle periods” and “busy periods” of length $I_1, B_1, I_2, B_2, \ldots$. Each idle period begins with $Q_2 = Q_4 = 0$, contains a single service completion, and ends with either $Q_2 = 1$ or $Q_4 = 1$ but not both. Using the memoryless property of the exponential distribution we have that

$$(3.2)\quad a_j \equiv E(I_j) \leq \alpha \equiv \max(m_1, m_3) \quad \text{for all } j = 1, 2, \ldots$$

Of course, the expectation in (3.2) depends on the particular initial condition $Q(0)$ assumed, but we suppress that dependence to simplify notation, and the inequality in (3.2) is understood to hold regardless of initial state.

A “busy period” begins with one of the two servers having a single priority customer to work on (there may be non-priority customers present at one or both stations), and it ends as soon as that server runs out of priority customers. Because of the preemptive-resume priorities and the routing pictured in Figure 1.2, no priority customers are created for the other server during such a busy period, so it ends with $Q_2 = Q_4 = 0$ once again. The key to our analysis is the following bound, whose proof is postponed until the end of this section.

**Lemma 3.1.** There exists a constant $c > 0$, not depending on $n$, such that

$$(3.3)\quad b_j \equiv E(B_j) \geq cn \quad \text{for all } j = 1, 2, \ldots$$

The transient states $q = (q_1, q_2, q_3, q_4)$ referred to earlier are those with $q_2 > 0$ and $q_4 > 0$. As we have seen, such states cannot be revisited after the initial interval $[0, \tau]$, but all other states of the Markov chain $Q(t)$ communicate and hence there is a unique stationary distribution. Let us denote by $\pi_2$ the stationary probability of those states $q$ with $q_2 > 0$, by $\pi_4$ the stationary probability of those states $q$ with $q_4 > 0$, and by $\pi_{13}$ the stationary probability of those $q$ with $q_1 + q_3 = n$ (that is, $q_2 = q_4 = 0$). From the path decomposition described above (an initial period of finite
expected duration, followed by alternating idle and busy periods, each with finite expected duration) it follows that

\[(3.4)\quad \pi_2 + \pi_4 + \pi_{13} = 1.\]

Moreover, with \(a_j \equiv E(I_j)\) and \(b_j \equiv E(B_j)\) as in (3.2) and (3.3), we have that

\[(3.5)\quad \pi_{13} = \lim_{n \to \infty} \left[ \frac{\sum_{j=1}^{n} a_j}{\sum_{j=1}^{n} a_j + b_j} \right].\]

Of course, two equivalent expressions for the long-run throughput rate \(\lambda\) are (here \(\mu_k \equiv 1/m_k\))

\[(3.6)\quad \lambda = \mu_2 \pi_2 \quad \text{and} \quad \lambda = \mu_4 \pi_4,\]

implying that \(\lambda(m_2 + m_4) = \pi_2 + \pi_4\). With our convention that \(m_2 + m_4 = 1\), this and (3.4) together imply

\[(3.7)\quad \lambda = (1 - \pi_{13})/(m_2 + m_4) = 1 - \pi_{13}.\]

Together, (3.2), (3.3) and (3.5) imply that \(\pi_{13} \to 0\) as \(n \to \infty\), so we have the following major conclusion.

Proposition 3.1. \(\lambda \to 1\) as \(n \to \infty\).

Corollary 3.1. The long-run utilization rate for server 1 is \(\lambda(m_1 + m_4) \to m_1 + m_4 < 1\) as \(n \to \infty\). Similarly, the long-run utilization rate for server 2 is \(\lambda(m_2 + m_3) \to m_2 + m_3 < 1\) as \(n \to \infty\).

It remains to prove Lemma 3.1. From (3.1) we have that

\[(3.8)\quad \mu_1 > \mu_2 \quad \text{and} \quad \mu_3 > \mu_4.\]

Let us suppose that the \(j^{th}\) busy period (its duration is \(B_j\)) begins with the creation of a class 2 customer, and that \(k\) customers of class 1 remain in existence after that creation (\(k \leq n - 1\)). Until the supply of class 1 customers has been exhausted, the queue length process \(Q_2(t)\) is precisely that of an \(M/M/1\) queue with input rate \(\mu_1\), service rate \(\mu_2\), and hence traffic intensity parameter \(\rho = \mu_1/\mu_2 > 1\). For such an \(M/M/1\) queue there is a probability \(p_{12} > 0\) that a busy period never ends. Using this and Wald's identity, we obtain a lower bound of \(p_{12} m_2 k\) for the conditional expectation of \(B_j\) given \(k\). Similarly, if the \(i^{th}\) busy period begins with creation of a class 4 customer this creation leaving \(n - k - 1\) class 3 customers behind, then the conditional expectation of \(B_j\) is bounded below by \(p_{34} m_4 (n - k)\), where \(p_{34} > 0\). It follows that (3.3) holds with

\[(3.9)\quad c = \min(p_{12} m_2, p_{34} m_4) > 0.\]
REFERENCES
