BOTTLENECKS IN MARKOVIAN QUEUEING NETWORKS

P.K. Pollett
Department of Mathematics
The University of Queensland
Queensland 4072
AUSTRALIA
pkp@maths.uq.edu.au

Abstract

We will consider the problem of identifying regions of congestion in closed queueing networks with state-dependent service rates. A particular queue will be called a bottleneck if the number of customers in that queue grows without bound as the total number of customers in the network becomes large. We will review methods for identifying potential bottlenecks, with a view to controlling congestion. We will see that the problem of identifying bottlenecks can be reduced to one of finding them in an isolated subnetwork with suitably modified routing intensities. Several special cases will be studied, illustrating a range of behaviour. For example, it is possible for a subnetwork to be congested, yet each queue in that subnetwork is not strictly a bottleneck.

Keywords Networks; Queueing Theory and Applications; Stochastic Models

1 INTRODUCTION

We shall be concerned with systems in which there are a fixed number of nodes J and a total number of individual units N circulating between the nodes. Such a system might describe a "job shop", where manufactured items are fashioned by various machines in turn (Jackson (1963)); it might describe the provision of spare parts for a collection of machines (Taylor and Jackson (1954)); it might describe a mining operation, where coal faces are worked in turn by a number of specialized machines (Koenigsberg (1958)). In all these examples there may be bottlenecks (regions of congestion). We shall present methods for identifying potential bottlenecks, with a view to controlling congestion.

The simplest general model for such a system of nodes is the migration process of Whittle (1968), where individuals, such as birds, migrate from one colony to another. However, it will be convenient here for us to think of the collection of nodes as being a queueing network and the individuals as being customers that require service at the various nodes. We may (as in Kelly (1975)) allow customers to be labelled according to their route through the network (which may be fixed or random), or we may allow them to have a general service-time distribution (as in Kelly (1976), for example). All we shall require here is that the steady-state (joint) distribution π of the numbers of customers $n = (n_1, n_2, \ldots, n_J)$ at the various nodes has the ubiquitous product form

$$\pi(n) = B_N \prod_{j=1}^{J} \frac{\alpha_j^{n_j}}{\prod_{r=1}^{n_j} \phi_j(r)}, \qquad n \in S, \tag{1}$$

where S is the finite subset of Z_+^J with $\sum_j n_j = N$ and B_N is a normalizing constant chosen so that π sums to unity over S. Here α_j is proportional to the amount of service requirement (in items per minute) coming into queue j, that is, the mean arrival rate (in customers per minute) multiplied by the expected service time (in items per customer); this will actually be *equal to* $\alpha_j B_N / B_{N-1}$. The quantity $\phi_j(n)$ is the service effort at queue j when there are n customers present (measured in items per minute). We shall assume that $\phi_j(0) = 0$ and $\phi_j(n) > 0$ whenever $n \geq 1$. The quantities $\alpha_1, \alpha_2, \ldots, \alpha_J$ are usually determined by a set of traffic equations, which govern the way in which customers are routed through the network. An irreducibility assumption on the traffic equations will give us $\alpha_j > 0$ for all j, and we may assume, without loss of generality, that $\sum_j \alpha_j = 1$. The form of the service rates is determined by the queueing discipline. For example, when $\phi_j(n) = n$, every customer at queue j gets the *same* service effort (this arises in the *infinite-server queue* and the certain *loss systems*), while if $\phi_j(n) = \min\{n, s_j\}$ (for $n \geq 1$), then at most s_j customers receive service, each at the same rate (this covers *first-come first-served* and *processor-sharing* disciplines, as well as some *preemptive-resume* disciplines such as *last-come first-served with preemption*). In these latter cases α_j / s_j is called the *traffic intensity* at queue j. Further details, as well as a full description of the examples cited above, can be found in Kelly (1979).

2 SIMPLE BOTTLENECKS

To illustrate what we mean by a bottleneck, and the kinds of behaviour that can occur, we shall first study in detail some of the cases mentioned above. First we shall consider the case

when the service rates are linear functions; imagine a network of infinite-server queues.

Example 1 Suppose that $\phi_j(n) = n$ for each j. We see from (1) that π is the multinomial distribution

$$\pi^{(N)}(n) = \binom{N}{n_1\ n_2\ \ldots\ n_J} \alpha_1^{n_1}\alpha_2^{n_2}\ldots\alpha_J^{n_J}, \qquad n \in S.$$

(Since we shall be varying N, the total number of customers in the network, we shall make the dependence of π on N explicit in our notation.) It follows that the marginal distribution of n_j, the number of customers at queue j, is binomial: in an obvious notation,

$$\pi_j^{(N)}(n) = \binom{N}{n} \alpha_j^n (1-\alpha_j)^{N-n}, \qquad n = 0, 1, \ldots, N.$$

It is clear that for *every* j, $\Pr(n_j \geq m) \to 1$ for each $m \geq 0$ as $N \to \infty$. We deduce that every queue becomes congested as the total number of customers becomes large. This is not surprising, because no queueing occurs; customers do not hinder one another as they pass through the network.

In our first example, we may be equally right in saying that *all* queues are bottlenecks or *none* is. However, we shall formally define a bottleneck as follows:

Definition Queue j is said to be a *bottleneck* if, for all $m \geq 0$, $\Pr(n_j \geq m) \to 1$ as $N \to \infty$.

Next we shall consider the case when the service rates are constant; imagine a network of single-server queues, each with the standard first-come first-served discipline.

Example 2 Suppose that $\phi_j(n) = 1$ $(n \geq 1)$ for each j. In this case (1) becomes

$$\pi^{(N)}(n) = B_N \alpha_1^{n_1}\alpha_2^{n_2}\ldots\alpha_J^{n_J}, \qquad n \in S,$$

where now B_N cannot be written down explicitly. Since the service effort is the *same* at each queue, α_j is indeed the traffic intensity at queue j. If we suppose that $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{J-1} < \alpha_J$, then we should surely expect queue J, the queue with the most traffic, to be the most congested. But, what really happens as the total number of customers becomes large? This question is most easily answered using generating functions. In the general context of (1) define $\Phi_1, \Phi_2, \ldots, \Phi_J$ by

$$\Phi_j(z) = 1 + \sum_{n=1}^{\infty} \frac{\alpha_j^n}{\prod_{r=1}^n \phi_j(r)} z^n. \tag{2}$$

For the special case presently under consideration, we have $\Phi_j(z) = 1/(1-\alpha_j z)$. It is easily shown that $B_N^{-1} = <\Pi_{j=1}^J \Phi_j>_N$, where $<\cdot>_n$ takes the n^{th} coefficient of a power series, and that the marginal distribution of n_j can be evaluated as

$$\pi_j^{(N)}(n) = B_N <\Phi_j>_n <\Pi_{k\neq j}\Phi_k>_{N-n}, \qquad n = 0, 1, \ldots, N. \tag{3}$$

See Exercises 2.3.6 and 2.3.7 of Kelly (1979). For the present case we have $<\Phi_j>_n = \alpha_j^n$ and so, since in particular $<\Phi_j>_{n+m} = \alpha_j^m <\Phi_j>_n$, we find, on summing (3) over n, that $Pr(n_j \geq m) = \alpha_j^m B_N / B_{N-m}$ (this is Exercise 2.3.8 of Kelly (1979)). If we can show that $B_{N-1}/B_N \to \alpha_J$ as $N \to \infty$, we will have established that $Pr(n_J \geq m) \to 1$ (queue J is a bottleneck) and $Pr(n_j \geq m) \to (\alpha_j/\alpha_J)^m < 1$ for $j < J$ (the others are not). To see that

$B_{N-1}/B_N \to \alpha_J$, consider the power series $\Theta_i = \Phi_1 \cdots \Phi_i$, where $\Phi_j(z) = 1/(1-\alpha_j z)$. Clearly Φ_i has radius of convergence $\rho_i = 1/\alpha_i$, and, in particular, Θ_1 $(= \Phi_1)$ has radius of convergence $1/\alpha_1$. We will show, using mathematical induction, that Θ_i has radius of convergence $1/\alpha_i$ *for all* i. To this end, let $k \geq 1$ and suppose that Θ_k has radius of convergence $1/\alpha_k$. Consider the expansion

$$<\Theta_{k+1}>_m = \sum_{n=0}^{m} \alpha_{k+1}^{m-n} <\Theta_k>_n = \alpha_{k+1}^m \sum_{n=0}^{m} \rho_{k+1}^n <\Theta_k>_n \tag{4}$$

and observe that $\sum_{n=0}^{\infty} \rho_{k+1}^n <\Theta_k>_n = \Theta_k(\rho_{k+1}) < \infty$, since $\rho_{k+1} < \rho_k$. We see immediately that

$$\frac{<\Theta_{k+1}>_m}{<\Theta_{k+1}>_{m+1}} \to \frac{1}{\alpha_{k+1}}, \qquad \text{as } m \to \infty,$$

and deduce that Θ_{k+1} has radius of convergence $1/\alpha_{k+1}$. This completes the induction. Finally, we have in particular that

$$\frac{B_N}{B_{N-1}} = \frac{<\Theta_J>_{N-1}}{<\Theta_J>_N} \to \frac{1}{\alpha_J}, \qquad \text{as } N \to \infty.$$

So, certainly in the case when $\phi_j(n) = 1$ $(n \geq 1)$ for each j, if there is a queue whose traffic intensity is *strictly greater* than the others, it behaves as a bottleneck, and, for each queue j in the remainder of the network, the marginal distribution of the number of customers at that queue approaches a geometric distribution with parameter α_j/α_J in the limit as $N \to \infty$. Indeed one can show, using the same techniques, that $n_1, n_2, \ldots, n_{J-1}$ are asymptotically independent.

This is the import of Whittle (1968). Whittle saw that the simple arguments presented above could be extended to cover the general case of state-dependent service rates. He proved, under certain mild assumptions, that if there is a queue whose generating function has unique minimal radius of convergence, then, not only is that queue a bottleneck, but the remainder of the network behaves as an *open* network in the limit as N gets large, with the bottleneck queue providing the exogeneous input. We shall content ourselves with the following result, which gives conditions for a particular queue to be a bottleneck:

Proposition 1 Suppose that Φ_j, defined by (2), has radius of convergence ρ_j and that $\rho_J < \rho_{J-1} \leq \rho_{J-2} \leq \cdots \leq \rho_1$. Suppose also that $<\Phi_1 \cdots \Phi_{J-1}>_{n-1}/<\Phi_1 \cdots \Phi_{J-1}>_n$ has a limit as $n \to \infty$. Then, queue J is a bottleneck.

Proof. As before, let $\Theta_i = \Phi_1 \cdots \Phi_i$ and observe that radius of convergence of Θ_i is equal to ρ_i, the smallest radius of convergence of the generating functions in the product. In particular, Θ_{J-1} has radius of convergence ρ_{J-1}. We have assumed that $<\Theta_{J-1}>_{n-1}/<\Theta_{J-1}>_n$ converges, and so it must converge to ρ_{J-1}. Hence $<\Theta_{J-1}>_{n-m}/<\Theta_{J-1}>_n$ converges to ρ_{J-1}^m. From (3) we see that the marginal distribution of n_J can be written

$$\pi_J^{(N)}(n) = B_N <\Phi_J>_n <\Theta_{J-1}>_{N-n}, \qquad n = 0, 1, \ldots, N. \tag{5}$$

Since we have the expansion

$$B_N^{-1} = <\Theta_J>_N = \sum_{m=0}^{N} <\Phi_J>_m <\Theta_{J-1}>_{N-m},$$

we may deduce that

$$(B_N<\Theta_{J-1}>_{N-n})^{-1} = \frac{<\Theta_{J-1}>_N}{<\Theta_{J-1}>_{N-n}} \sum_{m=0}^{N} <\Phi_J>_m \frac{<\Theta_{J-1}>_{N-m}}{<\Theta_{J-1}>_N}\,.$$

The term $<\Theta_{J-1}>_N/<\Theta_{J-1}>_{N-n}$ converges to ρ_{J-1}^{-n} as $N \to \infty$. But, the sum diverges since, because $\rho_J < \rho_{J-1}$, Φ_J diverges at ρ_{J-1} (by choosing N sufficiently large we can make the ratio $<\Theta_{J-1}>_{N-m}/<\Theta_{J-1}>_N$ as close as we please to ρ_{J-1}^m). It now follows that $B_N<\Theta_{J-1}>_{N-n}$, and hence $\Pr(n_J = n)$, converges to 0 for every n as $N \to \infty$. This establishes that queue J is a bottleneck.

Proposition 1 does not say anything about the limiting behaviour of the remaining queues $1, 2, \ldots, J-1$. We shall examine this question briefly by considering the marginal distribution of total number of customers in those queues. Clearly we have

$$\Pr(\textstyle\sum_{i=1}^{J-1} n_i = n) = B_N<\Theta_{J-1}>_n <\Phi_J>_{N-n}, \qquad n = 0, 1, \ldots, N, \tag{6}$$

and so it is natural to consider the expansion

$$(B_N<\Phi_J>_{N-n})^{-1} = \frac{<\Phi_J>_N}{<\Phi_J>_{N-n}} \sum_{m=0}^{N} <\Theta_{J-1}>_m \frac{<\Phi_J>_{N-m}}{<\Phi_J>_N}\,. \tag{7}$$

If we assume that $\phi_J(n)$ has a limit as $n \to \infty$ (as in Whittle (1968)), then a formal argument, based on the fact that $<\Phi_J>_{N-m}/<\Phi_J>_N \to \rho_J^m$, suggests that (7) is asymptotically $\rho_J^{-n}\Theta_{J-1}(\rho_J)$, because $\sum_{m=0}^{\infty} \rho_J^m <\Theta_{J-1}>_m = \Theta_{J-1}(\rho_J) < \infty$, and hence that

$$\Pr(\textstyle\sum_{i=1}^{J-1} n_i = n) \to \rho_J^n <\Theta_{J-1}>_n / \Theta_{J-1}(\rho_J).$$

Indeed, a similar argument, using the marginal distribution of n_j $(j < J)$, and based on an assumption that $<\Pi_{k\neq j}\Phi_k>_{n-1}/<\Pi_{k\neq j}\Phi_k>_n$ has a limit as $n \to \infty$, gives

$$\Pr(n_j = n) \to \rho_J^n <\Phi_j>_n / \Phi_j(\rho_J)\,,$$

since $\sum_{n=0}^{\infty} \rho_J^n <\Phi_j>_n = \Phi_j(\rho_J) < \infty$. If these arguments could be justified, we would conclude that, for each $j < J$, $\Pr(n_j \geq m)$ converges to a limit as $N \to \infty$, which is strictly less than 1 for every m, and, hence, that none of queues $1, 2, \ldots, J-1$ are bottlenecks. The argument *can* be justified in a variety of circumstances, but, it would appear, not in general. The problem can be reduced to one of determining the limiting behaviour of convolutions of positive sequences, as follows.

Remark 1 Let (a_n) and (b_n) be sequences of strictly positive numbers and define (d_n) by $d_n = \sum_{m=0}^{n} a_m b_{n-m}/b_n$, which can be compared with the sum in (7). Our hypothesis is that the power series $A(z) = \sum a_n z^n$ and $B(z) = \sum b_n z^n$ have strictly positive radii of convergence, ρ_A and ρ_B, which satisfy $\rho = \rho_B < \rho_A$, together with the condition that $(b_{n-1}/b_n) \to \rho$. We would like to deduce that $(d_n) \to A(\rho)$ (which is finite). The case considered in the proof of Proposition 1 had $\rho_A < \rho_B = \rho$. Under this condition, we can make n sufficiently large, thus making (b_{n-m}/b_n) as close as we please to ρ^m, and, since $\rho_A < \rho$, we deduce that $(d_n) \to \infty$. To deal with the present case $\rho = \rho_B < \rho_A$, define $f_m(n)$ by

$$f_m(n) = \begin{cases} b_{n-m}/b_n & \text{if } m \leq n \\ 0 & \text{if } m > n, \end{cases}$$

so that d_n can be written as $d_n = \sum_{m=0}^{\infty} a_m f_m(n)$, and, for each m, $f_m(n) \to \rho^m$ as $n \to \infty$. By Fatou's lemma we always have that the limit infimum of (d_n) is at least $A(\rho)$. (Thus, in the context of Proposition 1, we can deduce that the limit supremum (as $N \to \infty$) of $\Pr(n_j = n)$ is *no greater* than $\rho_J^n <\Phi_j>_n / \Phi_j(\rho_J)$.) But, by using dominated convergence and monotone convergence in turn, we can delimit two criteria under which (d_n) converges to $A(\rho)$. Firstly, if there is a sequence (g_m) with $b_{n-m}/b_n \leq g_m$, for all $n \geq m$, and $\sum_{m=0}^{\infty} a_m g_m < \infty$, then the convergence of (d_n) is guaranteed. (Note that, although (b_{n-m}/b_n) converges for each m, we cannot always bound $f_m(n)$ *uniformly* in m; the simple case $b_n = (1/2)^n$ illustrates this.) On the other hand if, for all m, (b_{n-m}/b_n) is monotonic in n, then again the convergence of (d_n) is assured.

We shall illustrate Proposition 1 by considering a network of multiserver queues.

Example 3 Suppose that queue j has s_j servers, so that the traffic intensity at queue j is α_j/s_j. Since $\phi_j(n) = \min\{n, s_j\}$, we have $\phi_j(n) \to s_j$, and so $<\Phi_j>_{n-1}/<\Phi_j>_n \to s_j/\alpha_j$. Therefore ρ_j is the reciprocal of the traffic intensity at queue j. If we can verify the convergence condition of Proposition 1, it will then follow that, whenever there is a unique queue with maximal traffic intensity, it is a bottleneck. As in Proposition 1, suppose that the queues are labelled according to increasing traffic intensity. For simplicity, suppose that $\rho_J < \rho_{J-1} < \rho_{J-2} < \cdots < \rho_1$. We will prove that $<\Theta_i>_{n-1}/<\Theta_i>_n \to \rho_i$ for all i, where $\Theta_i = \Phi_1 \cdots \Phi_i$, so that in particular $<\Theta_{J-1}>_{n-1}/<\Theta_{J-1}>_n$ has a limit as $n \to \infty$, and hence that queue J, the queue with maximal traffic intensity, is a bottleneck. We already have that $<\Theta_1>_{n-1}/<\Theta_1>_n \to \rho_1$ because $\Theta_1 = \Phi_1$. For any $k \geq 1$ we have the expansion

$$<\Theta_{k+1}>_n = <\Phi_{k+1}>_n \sum_{m=0}^{n} <\Theta_k>_m \frac{<\Phi_{k+1}>_{n-m}}{<\Phi_{k+1}>_n} .$$

Now, since $\phi_{k+1}(n) \uparrow s_{k+1}$, we have that $<\Phi_{k+1}>_{n-1}/<\Phi_{k+1}>_n \uparrow \rho_{k+1}$. It follows that $<\Phi_{k+1}>_{n-m}/<\Phi_{k+1}>_n \uparrow \rho_{k+1}^m$. So, since $\rho_{k+1} < \rho_k$, we can use Remark 1 to deduce that $<\Theta_{k+1}>_n \sim <\Phi_{k+1}>_n \Theta_k(\rho_{k+1})$ as $n \to \infty$. It follows that $<\Theta_{k+1}>_{n-1}/<\Theta_{k+1}>_n \to \rho_{k+1}$, as required. Also, since $\phi_j(n) \uparrow s_j$, for all j, and since $\sum_{n=0}^{\infty} \rho_J^n <\Phi_j>_n = \Phi_j(\rho_J) < \infty$, we may deduce that

$$\Pr(n_j = n) \to \rho_J^n <\Phi_j>_n / \Phi_j(\rho_J) ,$$

where

$$<\Phi_j>_n = \begin{cases} \alpha_j^n/n! & \text{if } n < s_j \\ \alpha_j^n/(s_j!\, s_j^{n-s_j}) & \text{if } n \geq s_j \end{cases} = \begin{cases} \alpha_j^n/n! & \text{if } n < s_j \\ s_j^{s_j}/(s_j!\, \rho_j^n) & \text{if } n \geq s_j , \end{cases} \tag{8}$$

and hence

$$\Phi_j(z) = \sum_{n=0}^{s_j - 1} \frac{(\alpha_j z)^n}{n!} + \frac{(\alpha_j z)^s}{s!} \left(\frac{\rho_j}{\rho_j - z} \right) , \qquad |z| < \rho_j .$$

It is interesting to note that our calculations here do not depend on the particular *form* of service rates, only on their monotonicity. We have therefore proved the following result, the first part of which is a corollary of Proposition 1.

Proposition 2 Suppose that Φ_j, defined by (2), has radius of convergence ρ_j and that $\rho_J < \rho_{J-1} < \rho_{J-2} < \cdots < \rho_1$. Suppose also that $\phi_j(n)$ converges monotonically in n for all j.

Then, queue J is a bottleneck, and for each $j < J$, $\Pr(n_j = n) \to \rho_J^n <\Phi_j>_n / \Phi_j(\rho_J)$ as $N \to \infty$.

We remark that the monotonicity condition can be replaced by a boundedness condition in accordance with Remark 1, but we contend that some condition *is* needed. The condition that $\rho_1, \rho_2, \ldots, \rho_{J-1}$ are distinct can also be relaxed. If, for some $k < J-1$, $\rho_k = \rho_{k+1}$, then we require only that $\Theta_k(\rho_k) < \infty$, and this will certainly be true (by induction on k) if $\Phi_j(\rho_j) < \infty$ for all $j < J$. In Example 3 we had $\Phi_j(\rho_j) = \infty$ for all j, but the particular form (8) of the service rates allows us to draw the same conclusions; for brevity, we will not give the details here.

3 COMPOUND BOTTLENECKS

We have already seen (Example 1) that when $\phi_j(n) = n$ for all j, every queue in the network is a bottleneck. Notice that in this case each of the generating functions $\Phi_1, \Phi_2, \ldots, \Phi_J$ has infinite radius of convergence. What happens, more generally, when the generating functions corresponding to two or more queues share the same minimal radius of convergence?

Proposition 3 In the setup of Proposition 1, suppose that $\rho_L = \rho_{L+1} = \cdots = \rho_J (= \rho)$ and that $\rho < \rho_j$ for $j = 1, 2, \ldots, L-1$. Then, queues $L, L+1, \ldots, J$ behave jointly as a bottleneck in that $\Pr(\sum_{i=L}^J n_i \geq m) \to 1$ as $N \to \infty$.

Proof. Without loss of generality, suppose that $\rho_1 \geq \rho_2 \geq \ldots \rho_{L-1} \, (> \rho)$, so that in particular Θ_{L-1} has radius of convergence ρ_{L-1}. Using a simple extension of our notation, writing Θ_i^k for the product $\Phi_k \cdots \Phi_i$, we find, on summing (3) appropriately, that

$$\Pr(\textstyle\sum_{i=L}^J n_i = n) = B_N <\Theta_J^L>_n <\Theta_{L-1}^1>_{N-n}, \qquad n = 0, 1, \ldots, N. \tag{9}$$

Following a programme similar to that used in the proof of Proposition 1, we have

$$(B_N <\Theta_{L-1}^1>_{N-n})^{-1} = \frac{<\Theta_{L-1}^1>_N}{<\Theta_{L-1}^1>_{N-n}} \sum_{m=0}^N <\Theta_J^L>_m \frac{<\Theta_{L-1}^1>_{N-m}}{<\Theta_{L-1}^1>_N}.$$

The term $<\Theta_{L-1}^1>_N / <\Theta_{L-1}^1>_{N-n}$ converges to ρ_{L-1}^{-n} as $N \to \infty$, and so, as before, the sum diverges since, because $\rho < \rho_{L-1}$, Θ_J^L diverges at ρ_{L-1}. It follows directly from (9) that $\Pr(\sum_{i=L}^J n_i = n) \to 0$ for all n, and hence that $\Pr(\sum_{i=L}^J n_i \geq m) \to 1$ for all m. This completes the proof.

Since ρ_j $(j < L)$ is strictly bigger than ρ, we might expect queues $1, 2, \ldots, L-1$ not to be bottlenecks. Consider any one of these: queue j. A formal argument, based on (3), shows that $\Pr(n_j = n) \to \rho^n <\Phi_j>_n / \Phi_j(\rho)$, because $\sum_{n=0}^\infty \rho^n <\Phi_j>_n = \Phi_j(\rho) < \infty$. If this were to be justified, we would conclude that $\Pr(n_j \geq m)$ converges to a limit as $N \to \infty$, which is strictly less than 1 for every m, thus confirming that queues $1, 2, \ldots, L-1$ are not bottlenecks. In view of Remark 1, we would need $<\Pi_{k \neq j} \Phi_k>_{n-m} / <\Pi_{k \neq j} \Phi_k>_n$ to be monotonic in n, or bounded by g_m satisfying $\sum_{m=0}^\infty g_m <\Phi_j>_m < \infty$. In some cases we can use a direct argument. To illustrate this, we shall return to Example 2.

Example 2a Consider again the case when $\phi_j(n) = 1$ $(n \geq 1)$ for each j, but now suppose that $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{J-2} < \alpha_{J-1} = \alpha_J$, so that queues $J-1$ and J share a unique

maximal traffic intensity. By Proposition 3, these queues behave jointly as a bottleneck. But, is each queue *individually* a bottleneck? The same arguments can be used as before. Since, for each j, $Pr(n_j \geq m) = \alpha_j^m B_N / B_{N-m}$, and $B_{N-1}/B_N \to \alpha_J$ as $N \to \infty$, we deduce that $Pr(n_k \geq m) \to 1$, for $k = J-1$ or J, and $Pr(n_j \geq m) \to (\alpha_j/\alpha_J)^m < 1$, for $j < J-1$. So, queues $J-1$ and J are bottlenecks and the others are not.

It might be conjectured that when the generating functions corresponding to two queues share the same minimal radius of convergence, they are always bottlenecks individually. However, this is not the case. Brown and Pollett (1982) demonstrated that both of the other possibilities can occur: one or neither can be bottlenecks. The following examples were examined briefly by Brown and Pollett (1982). We shall provide further details and draw some additional conclusions.

Example 4 Consider a network with $J = 2$ queues, suppose that $\alpha_1 = \alpha_2 = 1/2$, and, for $n \geq 1$, $\phi_1(n) = (n+1)^2/n^2$ and $\phi_2(n) = 1$. Clearly Φ_1 and Φ_2 have the same radius of convergence $\rho = 2$. But, for all $n \geq 0$,

$$\Pr(n_1 = n) = \frac{B_N}{(n+1)^2}\left(\frac{1}{2}\right)^N = \frac{1}{(n+1)^2}\left\{\sum_{m=0}^{N}\frac{1}{(m+1)^2}\right\}^{-1}$$

and

$$\Pr(n_2 = n) = \frac{B_N}{(N-n+1)^2}\left(\frac{1}{2}\right)^N = \frac{1}{(N-n+1)^2}\left\{\sum_{m=0}^{N}\frac{1}{(m+1)^2}\right\}^{-1},$$

and so $\Pr(n_1 = n) \to 6/(\pi^2(n+1)^2)$ and $\Pr(n_2 = n) \to 0$. It follows that queue 2 is a bottleneck, but queue 1 is not. It is interesting to note that $\mathrm{E}\,(n_1)$ diverges as $N \to \infty$, because

$$\mathrm{E}\,(n_1) > \frac{6}{\pi^2}\sum_{n=1}^{N}\frac{n}{(n+1)^2} = \frac{6}{\pi^2}\left\{\sum_{n=1}^{N}\frac{1}{(n+1)} - \sum_{n=1}^{N}\frac{1}{(n+1)^2}\right\} \sim \frac{6}{\pi^2}\log(N)\,,$$

so that queue 1 can be interpreted as a bottleneck in a weaker sense. However, if we had chosen $\phi_1(n) = (n+1)^3/n^3$ $(n \geq 1)$, we would have $\mathrm{E}\,(n_1)$ convergent and $\mathrm{E}\,(n_1^2)$ divergent, and so forth.

Example 5 Consider a network with $J = 2$ queues, suppose that $\alpha_1 = \alpha_2 = 1/2$ and, for $n \geq 1$, $\phi_1(n) = \phi_2(n) = (n+1)^2/n^2$. Clearly $\Phi_1 = \Phi_2$, with radius of convergence $\rho = 2$. But, for all $n \geq 0$,

$$\Pr(n_1 = n) = \frac{1}{(n+1)^2(N-n+1)^2}\left\{\sum_{m=0}^{N}\frac{1}{(m+1)^2(N-m+1)^2}\right\}^{-1}.$$

Upon using partial fractions, we find that the term in braces reduces to

$$\frac{4}{(N+2)^3}\sum_{m=0}^{N}\frac{1}{m+1} + \frac{2}{(N+2)^2}\sum_{m=0}^{N}\frac{1}{(m+1)^2} \sim \frac{4\log(N)}{N^3} + \frac{2}{N^2}\cdot\frac{\pi^2}{6}\,,$$

and so $\Pr(n_1 = n) \to 3/(\pi^2(n+1)^2)$ as $N \to \infty$. By symmetry, *neither* queue is a bottleneck. Note also that neither n_1 nor n_2 has a proper limiting distribution as N gets large. This

is to be expected, for the system behaves as follows. When n_1 is small, n_2 is large, and vice versa, yet when this happens the rates at which customers are served at queues 1 and 2 are large and small, respectively. The system fluctuates between modes in which one or other of the queues is congested, spending a long period in each mode. The stability of each mode increases as N gets large; indeed, for a cyclic 2-node migration process with the given parameters, one can show that the equilibrium expected time to reach state $(N, 0)$ starting in state $(0, N)$ is of order N^2.

In both of the two previous examples, there were only two queues in the network. If we had appended several other queues whose generating functions had radius of convergence strictly larger than 2, our conclusions would have been no different. Indeed, given a subnetwork consisting of queues whose generating functions have the same strictly minimal radius of convergence, we can identify which of them (if any) is a bottleneck, by conditioning on the event that all N customers are located in that subnetwork. Using an elaboration of Whittle's argument (Whittle (1968)), Brown and Pollett (1982) proved the following result:

Proposition 4 Suppose that Φ_j, defined by (2), has radius of convergence ρ_j. Suppose that $\Phi_1, \Phi_2, \ldots, \Phi_K$ have the same strictly minimal radius of convergence ρ, and that $\phi_j(n)$ converges monotonically for some $j \in \{2, \ldots, K\}$. Then, queue 1 is a bottleneck if and only if

$$\Pr(n_1 \geq m \,|\, \textstyle\sum_{i=1}^K n_i = N) \to 1\,, \qquad \text{as } N \to \infty. \tag{10}$$

A sufficient condition for queue 1 to be a bottleneck is that Φ_1 diverges at its radius of convergence and that $<\Phi_2 \cdots \Phi_K>_{n-1}/<\Phi_2 \cdots \Phi_K>_n$ converges as $n \to \infty$.

Proof. To prove that (10) is necessary and sufficient for queue 1 to be a bottleneck, we will show that, as $N \to \infty$, $\Pr(n_1 = n | \sum_{i=1}^K n_i = N) \to 0$ if and only if $\Pr(n_1 = n) \to 0$. Since

$$\Pr(n_1 = n) = B_N <\Phi_1>_n <\Theta_J^2>_{N-n} = <\Phi_1>_n <\Theta_J^2>_{N-n}/<\Theta_J^1>_N\,,$$

and

$$\Pr(n_1 = n \,|\, \textstyle\sum_{i=1}^K n_i = N) = <\Phi_1>_n <\Theta_K^2>_{N-n}/<\Theta_K^1>_N\,,$$

this amounts to verifying that $<\Theta_J^2>_{N-n}/<\Theta_J^1>_N = O(<\Theta_K^2>_{N-n}/<\Theta_K^1>_N)$ as $N \to \infty$. Without loss of generality, suppose it is $\phi_K(n)$ which converges monotonically. We will then have that $<\Phi_K>_{n-l}/<\Phi_K>_n$ converges monotonically in n (to ρ^l) for each l. It follows, by monotone convergence, that d_n, given by,

$$d_n = \sum_{l=0}^{n} <\Theta_J^{K+1}>_l \frac{<\Phi_K>_{n-l}}{<\Phi_K>_n},$$

converges to $\Theta_J^{K+1}(\rho) < \infty$, and, in particular, (d_n) is bounded, and bounded away from 0. But, $<\Theta_J^K>_n = <\Phi_K>_n d_n$, and so $<\Theta_J^K>_n/<\Phi_K>_n$ is bounded similarly. We also have the two expansions

$$<\Theta_J^2>_{N-n} = \sum_{m=0}^{N-n} <\Theta_{K-1}^2>_m <\Theta_J^K>_{N-n-m}$$

and

$$<\Theta_K^2>_{N-n} = \sum_{m=0}^{N-n} <\Theta_{K-1}^2>_m <\Phi_K>_{N-n-m}\,,$$

from which it follows that $<\Theta_J^2>_{N-n} = O(<\Theta_K^2>_{N-n})$ as $N \to \infty$. A similar argument shows that $<\Theta_J^1>_N = O(<\Theta_K^1>_N)$, and it follows immediately that $<\Theta_J^2>_{N-n}/<\Theta_J^1>_N = O(<\Theta_K^2>_{N-n}/<\Theta_K^1>_N)$.

To see that the second condition is sufficient, consider the expansion

$$\frac{<\Theta_K^1>_N}{<\Theta_K^2>_{N-n}} = \frac{<\Theta_K^2>_N}{<\Theta_K^2>_{N-n}} \sum_{m=0}^{N} <\Phi_1>_m \frac{<\Theta_K^2>_{n-m}}{<\Theta_K^2>_n}. \tag{11}$$

Since $<\Theta_K^2>_{n-1}/<\Theta_K^2>_n$ has a limit as $n \to \infty$, it must converge to ρ, the common radius of convergence of $\Phi_2 \ldots \Phi_K$. Thus $<\Theta_K^2>_{n-m}/<\Theta_K^2>_m \to \rho^m$ and, since Φ_1 diverges at ρ, the sum in (11) diverges as $N \to \infty$. It follows that $<\Theta_K^2>_{N-n}/<\Theta_K^1>_N \to 0$ and hence the necessary and sufficient condition is satisfied. This completes the proof.

In applying Proposition 4, we can imagine that the subnetwork, consisting of queues whose generating functions have the same strictly minimal radius of convergence, is an *isolated* closed network of queues. To see this, suppose that we have a closed migration process and λ_{ij} (for $i, j \in \{1, 2, \ldots, J\}$) is the probability that an individual which has just left node i is destined for node j ($\lambda_{ii} = 0$ and $\sum_{j=1}^{J} \lambda_{ij} = 1$). If the matrix $\Lambda = (\lambda_{ij})$ is irreducible, then there is a unique collection of positive numbers $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_J)$ which satisfy the (traffic) equations $\alpha\Lambda = \alpha$. For the closed migration process, with expected service rates all equal to 1, these are precisely the arrival rates $\alpha_1, \ldots, \alpha_J$ which appear in (1). For $K < J$, we can always choose routing probabilities λ'_{ij}, for $i, j \in \{1, 2, \ldots, K\}$, with $\lambda'_{ii} = 0$ and $\sum_{j=1}^{K} \lambda'_{ij} = 1$, in such a way that $\alpha' = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ satisfies $\alpha'\Lambda' = \alpha'$, thus giving the *same* arrival rates within the smaller network consisting of K nodes. We simply set

$$\lambda'_{ij} = \lambda_{ij} + \frac{\sum_{l=K+1}^{J} \lambda_{il} \sum_{k=K+1}^{J} \alpha_k \lambda_{kj}}{\sum_{k=1}^{K} \alpha_k \sum_{l=K+1}^{J} \lambda_{kl}}.$$

It is easy to check that these routing probabilities satisfy the required conditions.

Our final example shows that the sufficient condition of Proposition 4 is not necessary.

Example 6 Consider a network with $J = 2$ queues, suppose that $\alpha_1 = \alpha_2 = 1/2$, and, for $n \geq 1$, $\phi_1(n) = (n+1)^2/n^2$ and $\phi_2(n) = (n+1)^3/n^3$. Both $\phi_1(n)$ and $\phi_2(n)$ have limit 1, as $n \to \infty$. Also, Φ_1 and Φ_2 have common radius of convergence $\rho = 2$, and both converge at their radius of convergence. However, for all $n \geq 0$,

$$\Pr(n_1 = n) \leq \frac{1}{(n+1)^2(N-n+1)^3} \left\{ \sum_{m=0}^{M} \frac{1}{(m+1)^3(N-m+1)^2} \right\}^{-1},$$

for any $M < N$. But, for *fixed* M, the expression in braces is $O(N^{-2})$ as $N \to \infty$, and so $\Pr(n_1 = n)$ is bounded above by a quantity which is $O(N^{-1})$. Hence, queue 1 is a bottleneck, yet the sufficient condition of Proposition 4 is not satisfied.

ACKNOWLEDGEMENTS

The author is grateful to Mark Thompson for his careful reading of the manuscript, and Phil Diamond and Alan Jones for helpful conversations on this work. The support of the Australian Research Council (Grant No. A49702317) is gratefully acknowledged.

REFERENCES

Brown, T. and Pollett, P. (1982) Some distributional approximations in Markovian networks, *Adv. Appl. Probab.* **14**, 654–671.

Jackson, J. (1963) Jobshop-like queueing systems, *Mgmt. Sci.* **10**, 131–142.

Kelly, F. (1975) Networks of queues with customers of different types, *J. Appl. Probab.* **12**, 542–554.

Kelly, F. (1976) Networks of queues, *Adv. Appl. Probab.* **8**, 416–432.

Kelly, F. (1979) *Reversibility and Stochastic Networks*, Wiley, Chichester.

Koenigsberg, E. (1958) Cyclic queues, *Operat. Res. Quart.* **9**, 22–35.

Taylor, J. and Jackson, R. (1954) An application of the birth and death process to the provision of spare machines, *Operat. Res. Quart.* **5**, 95–108.

Whittle, P. (1968) Equilibrium distributions for an open migration process, *J. Appl. Probab.* **5**, 567–571.