BOTTLENECKS IN MARKOVIAN QUEUEING NETWORKS

by

Phil Pollett

The University of Queensland

OUR SETTING

A closed network of queues:

- Fixed number of nodes (queues) J
- N customers circulating
- Usual Markovian assumptions in force

Examples:

- A job shop, where manufactured items are fashioned by various machines in turn.
- Provision of spare parts for a collection of machines.
- A mining operation, where coal faces are worked in turn by a number of specialized machines.

Can we identify regions of congestion (bottlenecks) from the parameters of the model?

BOTTLENECKS

Common sense:

The nodes with the smallest service effort will be the most congested.

A formal definition:

If n_j is the number of customers at node j, then this node is a *bottleneck* if, for all $m \geq 0$, $\Pr(n_j \geq m) \to 1$ as $N \to \infty$.

SIMPLE EXAMPLES

All nodes have infinitely many servers:

$\Pr(n_j = n) = \binom{N}{n}\alpha_j^n(1-\alpha_j)^{N-n}$, $n = 0, \ldots, N$, where α_j (< 1) is proportional to the arrival rate at node j divided by service rate. Clearly $\Pr(n_j = n) \to 0$ for each n as $N \to \infty$, and so *all nodes are bottlenecks*.

All nodes have a single server:

The distribution of n_j cannot be written down explicitly, but we can show that if there is a node j whose traffic intensity is *strictly greater* than the others, it is the unique bottleneck.

Moreover, for each node k in the remainder of the network, the distribution of n_k approaches a geometric distribution with parameter α_k/α_i in the limit as $N \to \infty$, and n_k, for $k \neq j$, are asymptotically *independent*.

MARKOVIAN NETWORKS

Our only assumption:

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 product form

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

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 node j (this will actually be *equal to* $\alpha_j B_N / B_{N-1}$). Suppose (wlog) that $\sum_j \alpha_j = 1$.

$\phi_j(n)$ is the service effort at node j (in items per minute) when there are n customers present. We shall assume that $\phi_j(0) = 0$ and $\phi_j(n) > 0$ whenever $n \geq 1$.

GENERATING FUNCTIONS

Our primary tool:

Define generating functions $\Phi_1, \Phi_2, \ldots, \Phi_J$ by

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

It is easily shown that $B_N^{-1} = <\prod_{j=1}^{J} \Phi_j>_N$, where $<\cdot>_n$ takes the n^{th} coefficient of a power series. The marginal distribution of n_j can be evaluated as

$$\pi_j^{(N)}(n) = B_N <\Phi_j>_n <\prod_{k \neq j} \Phi_k>_{N-n},$$

for $n = 0, 1, \ldots, N$.

SINGLE-SERVER NODES

Suppose that each node j has a single server ($\phi_j(n) = 1$ for $n \geq 1$). Then, $<\!\Phi_j\!>_n = \alpha_j^n$ and so $<\!\Phi_j\!>_{n+m} = \alpha_j^m <\!\Phi_j\!>_n$. Summing

$$\pi_j^{(N)}(n) = B_N <\!\Phi_j\!>_n <\!\textstyle\prod_{k \neq j} \Phi_k\!>_{N-n}$$

over n, and recalling that $B_N^{-1} = <\!\prod_{j=1}^J \Phi_j\!>_N$, gives $Pr(n_j \geq m) = \alpha_j^m B_N / B_{N-m}$.

Suppose that $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{J-1} < \alpha_J$, so that node J has maximal traffic intensity.

If we can prove that $B_{N-1}/B_N \to \alpha_J$ as $N \to \infty$, then $Pr(n_J \geq m) \to 1$ (node J is a bottleneck) and $Pr(n_j \geq m) \to (\alpha_j/\alpha_J)^m < 1$ for $j < J$ (the others are not).

WHY DOES $B_{N-1}/B_N \to \alpha_J$?

Define $\Theta_i = \Phi_1 \cdots \Phi_i$, where now $\Phi_j(z) = 1/(1-\alpha_j z)$. Clearly Φ_j has radius of convergence (RC) $\rho_j = 1/\alpha_j$; in particular, Θ_1 $(= \Phi_1)$ has RC $1/\alpha_1$.

Claim: Θ_i has RC $1/\alpha_i$ *for all* i, so that

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

Proof: Suppose Θ_k has RC $1/\alpha_k$ and consider

$$\begin{aligned} <\Theta_{k+1}>_m &= \sum_{n=0}^{m} \alpha_{k+1}^{m-n} <\Theta_k>_n \\ &= \alpha_{k+1}^m \sum_{n=0}^{m} \rho_{k+1}^n <\Theta_k>_n . \end{aligned}$$

Clearly $\sum_{n=0}^{\infty} \rho_{k+1}^n <\Theta_k>_n = \Theta_k(\rho_{k+1}) < \infty$, since $\rho_{k+1} < \rho_k$, and so

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

implying that Θ_{k+1} has RC $1/\alpha_{k+1}$.

THE GENERAL CASE

Message: Bottleneck behaviour depends on the relative sizes of the radii of convergence of the power series $\Phi_1, \Phi_2, \ldots, \Phi_J$.

Proposition 1: Suppose Φ_j 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

$$\frac{<\Phi_1 \cdots \Phi_{J-1}>_{n-1}}{<\Phi_1 \cdots \Phi_{J-1}>_n} \tag{1}$$

has a limit as $n \to \infty$. Then, node J is a bottleneck.

Example: Suppose node j has s_j servers, so that the traffic intensity at node j is proportional to 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 proportional to the reciprocal of the traffic intensity at node j. It can be shown that (1) holds.

COMPOUND BOTTLENECKS

What happens when the generating functions corresponding to two or more nodes *share* the *same* minimal RC?

Proposition 2: 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, nodes $L, L+1, \ldots, J$ behave *jointly* as a bottleneck in that Pr($\sum_{i=L}^{J} n_i \geq m) \to 1$ as $N \to \infty$.

It might be conjectured that when the generating functions corresponding to two nodes share the same minimal RC, they are always bottlenecks *individually*. However, while this is true when all nodes have a single server (because $Pr(n_j \geq m) \to (\rho/\rho_j)^m$), it is *not true* in general.

SOME EXAMPLES

Consider a network with $J = 2$ nodes and suppose that $\alpha_1 = \alpha_2 = 1/2$. In the following examples Φ_1 and Φ_2 have the same RC $\rho = 2$.

Only one node is a bottleneck: Suppose that $\phi_1(n) = (n+1)^2/n^2$ and $\phi_2(n) = 1$ for $n \geq 1$. Then, it can be shown that $\Pr(n_1 = n) \to 6/(\pi^2(n+1)^2)$ and $\Pr(n_2 = n) \to 0$ as $N \to \infty$.

Neither node is a bottleneck: Suppose that $\phi_1(n) = \phi_2(n) = (n+1)^2/n^2$ for $n \geq 1$. Then, $\Pr(n_1 = n) \to 3/(\pi^2(n+1)^2)$ as $N \to \infty$.

AND FINALLY ...

Proposition 3: Suppose that $\Phi_1, \Phi_2, \ldots, \Phi_K$ have the *same strictly minimal* RC ρ, and that $\phi_j(n)$ converges monotonically for some $j \in \{2, \ldots, K\}$. Then, node 1 is a bottleneck *if and only if*

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

A *sufficient condition* for node 1 to be a bottleneck is that Φ_1 diverges at its RC and

$$\frac{<\Phi_2 \cdots \Phi_K>_{n-1}}{<\Phi_2 \cdots \Phi_K>_n} \text{ converges as } n \to \infty.$$

This latter condition is not necessary: In the setup of the previous examples, suppose that $\phi_1(n) = (n+1)^2/n^2$ and $\phi_2(n) = (n+1)^3/n^3$ for $n \geq 1$. Then, Φ_1 and Φ_2 have common RC $\rho = 2$ and both *converge* at their RC. But, it can be shown that $\Pr(n_1 = n)$ is bounded above by a quantity which is $O(N^{-1})$ as $N \to \infty$, implying that node 1 is a bottleneck.