

Chapter 5

Markov-Modulated Queues

There are many computer, communication and manufacturing systems which give rise to queueing models where the arrival and/or service mechanisms are influenced by some external processes. In such models, a single unbounded queue evolves in an environment which changes state from time to time. The instantaneous arrival and service rates may depend on the state of the environment and also, to a limited extent, on the number of jobs present.

The system state at time t is described by a pair of integer random variables, (I_t, J_t) , where I_t represents the state of the environment and J_t is the number of jobs present. The variable I_t takes a finite number of values, numbered $0, 1, \dots, N$; these are also called the environmental *phases*. The possible values of J_t are $0, 1, \dots$. Thus, the system is in state (i, j) when the environment is in phase i and there are j jobs waiting and/or being served.

The two-dimensional process $X = \{(I_t, J_t); t \geq 0\}$ is assumed to have the Markov property, i.e. given the current phase and number of jobs, the future behaviour of X is independent of its past history. Such a model is referred to as a *Markov-modulated queue*. The corresponding state space, $\{0, 1, \dots, N\} \times \{0, 1, \dots\}$ is known as a *lattice strip*.

A fully general Markov-modulated queue, with arbitrary state-dependent transitions, is not tractable. However, one can consider a subclass of models which are sufficiently general to be useful, and yet can be solved efficiently. Those models satisfy the following restrictions:

- (i) There is a threshold M , such that the instantaneous transition rates out of state (i, j) do not depend on j when $j \geq M$.
- (ii) the jumps of the random variable J are bounded.

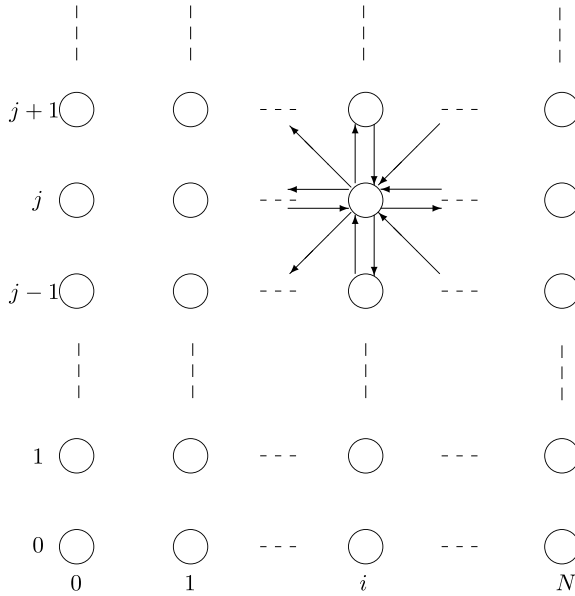


Fig. 5.1. State diagram of a QBD process.

When the jumps of the random variable J are of size 1, i.e. when jobs arrive and depart one at a time, the process is said to be of the *Quasi-Birth-and-Death* type, or QBD (the term *skip-free* is also used (Latouche *et al.*, [7]). The state diagram for this common model, showing some transitions out of state (i, j) , is illustrated in Fig. 5.1.

The requirement that all transition rates cease to depend on the size of the job queue beyond a certain threshold is not too restrictive. Note that there is no limit on the magnitude of the threshold M , although it must be pointed out that the larger M is, the greater the complexity of the solution. Similarly, although jobs may arrive and/or depart in fixed or variable (but bounded) batches, the larger the batch size, the more complex the solution.

The object of the analysis of a Markov-modulated queue is to determine the joint steady-state distribution of the environmental phase and the number of jobs in the system:

$$p_{i,j} = \lim_{t \rightarrow \infty} P(I_t = i, J_t = j); \quad i = 0, 1, \dots, N; \quad j = 0, 1, \dots \quad (5.1)$$

That distribution exists for an irreducible Markov process if, and only if, the corresponding set of balance equations has a positive solution that can be normalised.

The marginal distributions of the number of jobs in the system, and of the phase, can be obtained from the joint distribution:

$$p_{\cdot,j} = \sum_{i=0}^N p_{i,j} \quad (5.2)$$

$$p_{i,\cdot} = \sum_{j=0}^{\infty} p_{i,j}. \quad (5.3)$$

Various performance measures can then be computed in terms of these joint and marginal distributions.

The following are some examples of systems that are modelled as Markov-modulated queues.

5.1. A multiserver queue with breakdowns and repairs

A single, unbounded queue is served by N identical parallel servers (Mitrani and Avi-Itzhak, [9], Neuts and Lucantoni, [13]). Each server goes through alternating periods of being operative and inoperative, independently of the others and of the number of jobs in the system. The operative and inoperative periods are distributed exponentially with parameters ξ and η , respectively. Thus, the number of operative servers at time t , I_t , is a Markov process on the state space $\{0, 1, \dots, N\}$. This is the environment in which the queue evolves: it is in phase i when there are i operative servers.

Jobs arrive according to a Poisson process, with a rate which may depend on the state of the environment, I_t . That is, when there are i operative servers, the instantaneous arrival rate is λ_i . Jobs are taken for service from the front of the queue, one at a time, by available operative servers. The required service times are distributed exponentially with parameter μ . An operative server cannot be idle if there are jobs waiting to be served. A job whose service is interrupted by a server breakdown is returned to the front of the queue. When an operative server becomes available, the service is resumed from the point of interruption, without any switching overheads. The flow of jobs is shown in Fig. 5.2.

The process $X = \{(I_t, J_t); t \geq 0\}$ is of the Quasi-Birth-and-Death type. The transitions out of state (i, j) are:

- (a) to state $(i - 1, j)$ ($i > 0$), with rate $i\xi$;
- (b) to state $(i + 1, j)$ ($i < N$), with rate $(N - i)\eta$;
- (c) to state $(i, j + 1)$ with rate λ_i ;
- (d) to state $(i, j - 1)$ with rate $\min(i, j)\mu$.

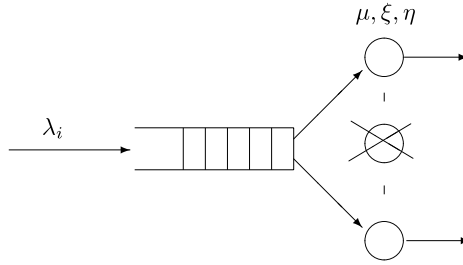


Fig. 5.2. A multiserver queue with breakdowns and repairs.

Note that only transition (d) has a rate which depends on j , and that dependency vanishes when $j \geq N$.

Remark. The breakdown and repair processes could be generalised without destroying the QBD nature of the process. For example, the servers could break down and be repaired in batches, or a server breakdown could trigger a job departure. The environmental state transitions can be arbitrary, as long as the queue changes in steps of size 1.

In this example, as in all models where the environment state transitions do not depend on the number of jobs present, the marginal distribution of the number of operative servers can be determined without finding the joint distribution first. Moreover, since the servers break down and are repaired independently of each other, that distribution is binomial:

$$p_{i,\cdot} = \binom{N}{i} \left(\frac{\eta}{\xi + \eta}\right)^i \left(\frac{\xi}{\xi + \eta}\right)^{N-i}; \quad i = 0, 1, \dots, N. \quad (5.4)$$

Hence, the steady-state average number of operative servers is equal to

$$E(X_t) = \frac{N\eta}{\xi + \eta}. \quad (5.5)$$

The overall average arrival rate is equal to

$$\lambda = \sum_{i=0}^N p_{i,\cdot} \lambda_i. \quad (5.6)$$

This gives us an explicit condition for stability. The offered load must be less than the processing capacity:

$$\frac{\lambda}{\mu} < \frac{N\eta}{\xi + \eta}. \quad (5.7)$$

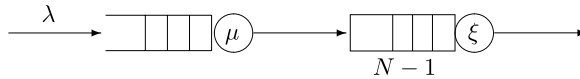


Fig. 5.3. Two nodes with a finite intermediate buffer.

5.2. Manufacturing blocking

Consider a network of two nodes in tandem, such as the one in Fig. 5.3 (Buzacott and Shanthikumar, [1], Konheim and Reiser, [6]). Jobs arrive into the first node in a Poisson stream with rate λ , and join an unbounded queue. After completing service at node 1 (exponentially distributed with parameter μ), they attempt to go to node 2, where there is a finite buffer with room for a maximum of $N-1$ jobs (including the one in service). If that transfer is impossible because the buffer is full, the job remains at node 1, preventing its server from starting a new service, until the completion of the current service at node 2 (exponentially distributed with parameter ξ). In this last case, server 1 is said to be “blocked”. Transfers from node 1 to node 2 are instantaneous.

The above type of blocking is referred to as “manufacturing blocking”. (An alternative model, which also gives rise to a Markov-modulated queue, is the “communication blocking”. There node 1 *does not start* a service if the node 2 buffer is full.)

In this system, the unbounded queue at node 1 is modulated by a finite-state environment defined by node 2. We say that the environment, I_t , is in state i if there are i jobs at node 2 and server 1 is not blocked ($i = 0, 1, \dots, N-1$). An extra state, $I_t = N$, is needed to describe the situation where there are $N-1$ jobs at node 2 and server 1 is blocked.

The above assumptions imply that the pair $X = \{(I_t, J_t); t \geq 0\}$, where J_t is the number of jobs at node 1, is a QBD process. Note that the state $(N, 0)$ does not exist: node 1 may be blocked only if there are jobs present.

The transitions out of state (i, j) are:

- (a) to state $(i-1, j)$ ($0 < i < N$), with rate ξ ;
- (b) to state $(N-1, j-1)$ ($i = N, j > 0$), with rate ξ ;
- (c) to state $(i+1, j-1)$ ($0 \leq i < N-1, j > 0$), with rate μ ;
- (d) to state (N, j) ($i = N-1, j > 0$), with rate μ ;
- (e) to state $(i, j+1)$ with rate λ .

The only dependency on j comes from the fact that transitions (b), (c) and (d) are not available when $j = 0$. In this example, the j -independency threshold is $M = 1$. Note that the state $(N, 0)$ is not reachable: node 1 may be blocked only if there are jobs present.

5.3. Phase-type distributions

There is a large and useful family of distributions that can be incorporated into queueing models by means of Markovian environments (Neuts, [12]). Those distributions are “almost” general, in the sense that any distribution function either belongs to this family or can be approximated as closely as desired by functions from it.

Let I_t be a Markov process with state space $\{0, 1, \dots, N\}$ and generator matrix \tilde{A} . States $0, 1, \dots, N - 1$ are transient, while state N , reachable from any of the other states, is absorbing (the last row of \tilde{A} is 0). At time 0, the process starts in state i with probability α_i ($i = 0, 1, \dots, N - 1$; $\alpha_0 + \alpha_1 + \dots + \alpha_{N-1} = 1$). Eventually, after an interval of length T , it is absorbed in state N . The random variable T is said to have a “phase-type” (PH) distribution with parameters \tilde{A} and α_i .

The exponential distribution is obviously phase-type ($N = 1$). So is the Erlang distribution — the convolution of N exponentials. The corresponding generator matrix is

$$\tilde{A} = \begin{bmatrix} -\mu & \mu & & & \\ & -\mu & \mu & & \\ & & \ddots & \ddots & \\ & & & -\mu & \mu \\ & & & & 0 \end{bmatrix},$$

and the initial probabilities are $\alpha_0 = 1$, $\alpha_1 = \dots = \alpha_{N-1} = 0$.

Another common PH distribution is the “hyperexponential”, where $I_0 = i$ with probability α_i , and absorption occurs at the first transition. The generator matrix of the hyperexponential distribution is

$$\tilde{A} = \begin{bmatrix} -\mu_0 & & & \mu_0 \\ & -\mu_1 & & \mu_1 \\ & & \ddots & \vdots \\ & & & -\mu_{N-1} & \mu_{N-1} \\ & & & & 0 \end{bmatrix}.$$

The corresponding probability distribution function, $F(x)$, is a mixture of exponentials:

$$F(x) = 1 - \sum_{i=0}^{N-1} \alpha_i e^{-\mu_i x}.$$

The PH family is very versatile. It contains distributions with both low and high coefficients of variation. It is closed with respect to mixing and convolution: if X_1 and X_2 are two independent PH random variables with N_1 and N_2 (non-absorbing) phases respectively, and c_1 and c_2 are constants, then $c_1 X_1 + c_2 X_2$ has a PH distribution with $N_1 + N_2$ phases.

A model with a single unbounded queue, where either the interarrival intervals, or the service times, or both, have PH distributions, is easily cast in the framework of a queue in Markovian environment. Consider, for instance, the M/PH/1 queue. Its state at time t can be represented as a pair (I_t, J_t) , where J_t is the number of jobs present and I_t is the phase of the current service (if $J_t > 0$). When I_t has a transition into the absorbing state, the current service completes and (if the queue is not empty) a new service starts immediately, entering phase i with probability α_i .

The PH/PH/ n queue can also be represented as a QBD process. However, the state of the environmental variable, I_t , now has to indicate the phase of the current interarrival interval and the phases of the current services at all busy servers. If the interarrival interval has N_1 phases and the service has N_2 phases, the state space of I_t would be of size $N_1 N_2^n$.

5.4. Checkpointing and recovery in the presence of faults

The last example is not a QBD process. Consider a system where transactions, arriving according to a Poisson process with rate λ , are served in FIFO order by a single server. The service times are i.i.d. random variables distributed exponentially with parameter μ . After N consecutive transactions have been completed, the system performs a checkpoint operation whose duration is an i.i.d. random variable distributed exponentially with parameter β . Once a checkpoint is established, the N completed transactions are deemed to have departed. However, both transaction processing and checkpointing may be interrupted by the occurrence of a fault. The latter arrive according to an independent Poisson process with rate ξ . When a fault occurs, the system instantaneously rolls

back to the last established checkpoint; all transactions which arrived since that moment either remain in the queue, if they have not been processed, or return to it, in order to be processed again (it is assumed that repeated service times are resampled independently).

This system can be modelled as an unbounded queue of (uncompleted) transactions, which is modulated by an environment consisting of completed transactions and checkpoints. More precisely, the two state variables, $I(t)$ and $J(t)$, are the number of transactions that have completed service since the last checkpoint, and the number of transactions present that have not completed service (including those requiring re-processing), respectively.

The Markov-modulated queueing process $X = \{[I(t), J(t)]; t \geq 0\}$, has the following transitions out of state (i, j) :

- (a) to state $(0, j + i)$, with rate ξ ;
- (b) to state $(0, j)(i = N)$, with rate β ;
- (c) to state $(i, j + 1)$, with rate λ ;
- (d) to state $(i + 1, j - 1)(0 \leq i < N, j > 0)$, with rate μ ;

Because transitions (a), resulting from arrivals of faults, cause the queue size to jump by more than 1, this is not a QBD process.

5.5. Spectral expansion solution

Let us now turn to the problem of determining the steady-state joint distribution of the environmental phase and the number of jobs present, for a Markov-modulated queue. The solution method that we shall present is called ‘‘Spectral Expansion’’, for reasons that will become apparent.

We shall start with the most commonly encountered case, namely the QBD process, where jobs arrive and depart singly. The starting point is of course the set of balance equations which the probabilities $p_{i,j}$, defined in (5.1), must satisfy. In order to write them in general terms, the following notation for the instantaneous transition rates will be used.

- (a) Phase transitions leaving the queue unchanged: from state (i, j) to state $(k, j)(0 \leq i, k \leq N; i \neq k)$, with rate $a_j(i, k)$;
- (b) Transitions incrementing the queue: from state (i, j) to state $(k, j + 1)(0 \leq i, k \leq N)$, with rate $b_j(i, k)$;
- (c) Transitions decrementing the queue: from state (i, j) to state $(k, j - 1)(0 \leq i, k \leq N; j > 0)$, with rate $c_j(i, k)$.

It is convenient to introduce the $(N + 1) \times (N + 1)$ matrices containing the rates of type (a), (b) and (c): $A_j = [a_j(i, k)]$, $B_j = [b_j(i, k)]$ and

$C_j = [c_j(i, k)]$, respectively (the main diagonal of A_j is zero by definition; also, $C_0 = 0$ by definition). According to the assumptions of the Markov-modulated queue, there is a threshold, $M (M \geq 1)$, such that those matrices do not depend on j when $j \geq M$. In other words,

$$A_j = A; \quad B_j = B; \quad C_j = C, \quad j \geq M. \tag{5.8}$$

Note that transitions (b) may represent a job arrival coinciding with a change of phase. If arrivals are not accompanied by such changes, then the matrices B_j and B are diagonal. Similarly, a transition of type (c) may represent a job departure coinciding with a change of phase. Again, if such coincidences do not occur, then the matrices C_j and C are diagonal.

By way of illustration, here are the transition rate matrices for the model of the multiserver queue with breakdowns and repairs. In this case the phase transitions are independent of the queue size, so the matrices A_j are all equal:

$$A_j = A = \begin{bmatrix} 0 & N\eta & & & \\ \xi & 0 & (N-1)\eta & & \\ & 2\xi & 0 & \ddots & \\ & & \ddots & \ddots & \eta \\ & & & N\xi & 0 \end{bmatrix}.$$

Similarly, the matrices B_j do not depend on j :

$$B = \begin{bmatrix} \lambda_0 & & & & \\ & \lambda_1 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \lambda_N \end{bmatrix}.$$

Denoting

$$\mu_{i,j} = \min(i, j)\mu; \quad i = 0, 1, \dots, N; \quad j = 1, 2, \dots,$$

the departure rate matrices, C_j , can thus be written as

$$C_j = \begin{bmatrix} 0 & & & & \\ & \mu_{1,j} & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \mu_{N,j} \end{bmatrix}; \quad j = 1, 2, \dots,$$

These matrices cease to depend on j when $j \geq N$. Thus, the threshold M is now equal to N , and

$$C = \begin{bmatrix} 0 & & & \\ & \mu & & \\ & & \ddots & \\ & & & N\mu \end{bmatrix}.$$

5.6. Balance equations

Using the instantaneous transition rates introduced above, the balance equations of a general QBD process can be written as

$$\begin{aligned} p_{i,j} \sum_{k=0}^N [a_j(i, k) + b_j(i, k) + c_j(i, k)] \\ = \sum_{k=0}^N [p_{k,j} a_j(k, i) + p_{k,j-1} b_{j-1}(k, i) + p_{k,j+1} c_{j+1}(k, i)], \end{aligned} \quad (5.9)$$

where $p_{i,-1} = b_{-1}(k, i) = c_0(i, k) = 0$ by definition. The left-hand side of (5.9) gives the total average number of transitions out of state (i, j) per unit time (due to changes of phase, arrivals and departures), while the right-hand side expresses the total average number of transitions into state (i, j) (again due to changes of phase, arrivals and departures). These balance equations can be written more compactly by using vectors and matrices. Define the row vectors of probabilities corresponding to states with j jobs in the system:

$$\mathbf{v}_j = (p_{0,j}, p_{1,j}, \dots, p_{N,j}); \quad j = 0, 1, \dots \quad (5.10)$$

Also, let D_j^A , D_j^B and D_j^C be the diagonal matrices whose i th diagonal element is equal to the i th row sum of A_j , B_j and C_j , respectively. Then equations (5.9), for $j = 0, 1, \dots$, can be written as:

$$\mathbf{v}_j [D_j^A + D_j^B + D_j^C] = \mathbf{v}_{j-1} B_{j-1} + \mathbf{v}_j A_j + \mathbf{v}_{j+1} C_{j+1}, \quad (5.11)$$

where $\mathbf{v}_{-1} = \mathbf{0}$ and $D_0^C = B_{-1} = 0$ by definition.

When j is greater than the threshold M , the coefficients in (5.11) cease to depend on j :

$$\mathbf{v}_j [D^A + D^B + D^C] = \mathbf{v}_{j-1} B + \mathbf{v}_j A + \mathbf{v}_{j+1} C, \quad (5.12)$$

for $j = M + 1, M + 2, \dots$

In addition, all probabilities must sum up to 1:

$$\sum_{j=0}^{\infty} \mathbf{v}_j \mathbf{e} = 1, \quad (5.13)$$

where \mathbf{e} is a column vector with $N + 1$ elements, all of which are equal to 1.

The first step is to find the general solution of the infinite set of balance equations with constant coefficients, (5.12). The latter are normally written in the form of a homogeneous vector difference equation of order 2:

$$\mathbf{v}_j Q_0 + \mathbf{v}_{j+1} Q_1 + \mathbf{v}_{j+2} Q_2 = \mathbf{0}; \quad j = M, M + 1, \dots, \quad (5.14)$$

where $Q_0 = B$, $Q_1 = A - D^A - D^B - D^C$ and $Q_2 = C$.

Associated with equation (5.14) is the so-called “characteristic matrix polynomial”, $Q(x)$, defined as

$$Q(x) = Q_0 + Q_1 x + Q_2 x^2. \quad (5.15)$$

Denote by x_k and \mathbf{u}_k the “generalised eigenvalues”, and corresponding “generalised left eigenvectors”, of $Q(x)$. In other words, these are quantities which satisfy

$$\begin{aligned} \det[Q(x_k)] &= 0, \\ \mathbf{u}_k Q(x_k) &= \mathbf{0}; \quad k = 1, 2, \dots, d, \end{aligned} \quad (5.16)$$

where $\det[Q(x)]$ is the determinant of $Q(x)$ and d is its degree. In what follows, the qualification *generalised* will be omitted.

The above eigenvalues do not have to be simple, but it is assumed that if one of them has multiplicity m , then it also has m linearly independent left eigenvectors. This tends to be the case in practice. So, the numbering in (5.16) is such that each eigenvalue is counted according to its multiplicity.

It is readily seen that if x_k and \mathbf{u}_k are any eigenvalue and corresponding left eigenvector, then the sequence

$$\mathbf{v}_{k,j} = \mathbf{u}_k x_k^j; \quad j = M, M + 1, \dots, \quad (5.17)$$

is a solution of equation (5.14). Indeed, substituting (5.17) into (5.14) we get

$$\mathbf{v}_{k,j} Q_0 + \mathbf{v}_{k,j+1} Q_1 + \mathbf{v}_{k,j+2} Q_2 = x_k^j \mathbf{u}_k [Q_0 + Q_1 x_k + Q_2 x_k^2] = \mathbf{0}.$$

By combining any multiple eigenvalues with each of their independent eigenvectors, we thus obtain d linearly independent solutions of (5.14).

On the other hand, it is known that there cannot be more than d linearly independent solutions (Gohberg *et al.*, [4]). Therefore, any solution of (5.14) can be expressed as a linear combination of the d solutions (5.17):

$$\mathbf{v}_j = \sum_{k=1}^d \alpha_k \mathbf{u}_k x_k^j; \quad j = M, M+1, \dots, \quad (5.18)$$

where α_k ($k = 1, 2, \dots, d$), are arbitrary (complex) constants.

However, the only solutions that are of interest in the present context are those which can be normalised to become probability distributions. Hence, it is necessary to select from the set (5.18), those sequences for which the series $\sum \mathbf{v}_j \mathbf{e}$ converges. This requirement implies that if $|x_k| \geq 1$ for some k , then the corresponding coefficient α_k must be 0.

So, suppose that c of the eigenvalues of $Q(x)$ are strictly inside the unit disk (each counted according to its multiplicity), while the others are on the circumference or outside. Order them so that $|x_k| < 1$ for $k = 1, 2, \dots, c$. The corresponding independent eigenvectors are $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_c$. Then any normalisable solution of equation (5.14) can be expressed as

$$\mathbf{v}_j = \sum_{k=1}^c \alpha_k \mathbf{u}_k x_k^j; \quad j = M, M+1, \dots, \quad (5.19)$$

where α_k ($k = 1, 2, \dots, c$), are some constants.

The set of eigenvalues of the matrix polynomial $Q(x)$ is called its “spectrum”. Hence, expression (5.19) is referred to as the “spectral expansion” of the vectors \mathbf{v}_j . The coefficients of that expansion, α_k , are yet to be determined.

Note that if there are non-real eigenvalues in the unit disk, then they appear in complex-conjugate pairs. The corresponding eigenvectors are also complex-conjugate. The same must be true for the appropriate pairs of constants α_k , in order that the right-hand side of (5.19) be real. To ensure that it is also positive, the real parts of x_k , \mathbf{u}_k and α_k should be positive.

So far, expressions have been obtained for the vectors $\mathbf{v}_M, \mathbf{v}_{M+1}, \dots$; these contain c unknown constants. Now it is time to consider the balance equations (5.11), for $j = 0, 1, \dots, M$. This is a set of $(M+1)(N+1)$ linear equations with $M(N+1)$ unknown probabilities (the vectors \mathbf{v}_j for $j = 0, 1, \dots, M-1$), plus the c constants α_k . However, only $(M+1)(N+1) - 1$ of these equations are linearly independent, since the generator matrix of the Markov process is singular. On the other hand, an additional independent equation is provided by (5.13).

In order that this set of linearly independent equations has a unique solution, the number of unknowns must be equal to the number of equations, i.e. $(M + 1)(N + 1) = M(N + 1) + c$, or $c = N + 1$. This observation implies the following rather general result.

Proposition 5.1. *The QBD process has a steady-state distribution if, and only if, the number of eigenvalues of $Q(x)$ strictly inside the unit disk, each counted according to its multiplicity, is equal to the number of states of the Markovian environment, $N + 1$. Then, assuming that the eigenvectors of multiple eigenvalues are linearly independent, the spectral expansion solution of (5.12) has the form*

$$\mathbf{v}_j = \sum_{k=1}^{N+1} \alpha_k \mathbf{u}_k x_k^j; \quad j = M, M + 1, \dots \quad (5.20)$$

In summary, the spectral expansion solution procedure consists of the following steps:

1. Compute the eigenvalues of $Q(x)$, x_k , inside the unit disk, and the corresponding left eigenvectors \mathbf{u}_k . If their number is other than $N + 1$, stop; a steady-state distribution does not exist.
2. Solve the finite set of linear equations (5.11), for $j = 0, 1, \dots, M$, and (5.13), with \mathbf{v}_M and \mathbf{v}_{M+1} given by (5.20), to determine the constants α_k and the vectors \mathbf{v}_j for $j < M$.
3. Use the obtained solution in order to determine various moments, marginal probabilities, percentiles and other system performance measures that may be of interest.

Careful attention should be paid to step 1. The “brute force” approach which relies on first evaluating the scalar polynomial $\det[Q(x)]$, then finding its roots, may be very inefficient for large N . An alternative which is preferable in most cases is to reduce the quadratic eigenvalue-eigenvector problem

$$\mathbf{u}[Q_0 + Q_1x + Q_2x^2] = \mathbf{0}, \quad (5.21)$$

to a linear one of the form $\mathbf{u}Q = x\mathbf{u}$, where Q is a matrix whose dimensions are twice as large as those of Q_0 , Q_1 and Q_2 . The latter problem is normally solved by applying various transformation techniques. Efficient routines for that purpose are available in most numerical packages.

This linearisation can be achieved quite easily if the matrix $C = Q_2$ is non-singular (Jennings, [5]). Indeed, after multiplying (5.21) on the right

by Q_2^{-1} , it becomes

$$\mathbf{u}[H_0 + H_1x + Ix^2] = \mathbf{0}, \quad (5.22)$$

where $H_0 = Q_0C^{-1}$, $H_1 = Q_1C^{-1}$, and I is the identity matrix. By introducing the vector $\mathbf{y} = x\mathbf{u}$, equation (5.22) can be rewritten in the equivalent linear form

$$[\mathbf{u}, \mathbf{y}] \begin{bmatrix} 0 & -H_0 \\ I & -H_1 \end{bmatrix} = x[\mathbf{u}, \mathbf{y}]. \quad (5.23)$$

If C is singular but B is not, a similar linearisation is achieved by multiplying (5.21) on the right by B^{-1} and making a change of variable $x \rightarrow 1/x$. Then the relevant eigenvalues are those outside the unit disk.

If both B and C are singular, then the desired result is achieved by first making a change of variable, $x \rightarrow (\gamma + x)/(\gamma - x)$, where the value of γ is chosen so that the matrix $S = \gamma^2Q_2 + \gamma Q_1 + Q_0$ is non-singular. In other words, γ can have any value which is not an eigenvalue of $Q(x)$. Having made that change of variable, multiplying the resulting equation by S^{-1} on the right reduces it to the form (5.22).

The computational demands of step 2 may be high if the threshold M is large. However, if the matrices B_j ($j = 0, 1, \dots, M-1$) are non-singular (which is often the case in practice), then the vectors $\mathbf{v}_{M-1}, \mathbf{v}_{M-2}, \dots, \mathbf{v}_0$ can be expressed in terms of \mathbf{v}_M and \mathbf{v}_{M+1} , with the aid of equations (5.11) for $j = M, M-1, \dots, 1$. One is then left with equations (5.11) for $j = 0$, plus (5.13) (a total of $N+1$ independent linear equations), for the $N+1$ unknowns x_k .

Having determined the coefficients in the expansion (5.20) and the probabilities $p_{i,j}$ for $j < N$, it is easy to compute performance measures. The steady-state probability that the environment is in state i is given by

$$p_{i,\cdot} = \sum_{j=0}^{M-1} p_{i,j} + \sum_{k=1}^{N+1} \alpha_k u_{k,i} \frac{x_k^M}{1-x_k}, \quad (5.24)$$

where $u_{k,i}$ is the i th element of \mathbf{u}_k .

The conditional average number of jobs in the system, L_i , given that the environment is in state i , is obtained from

$$L_i = \frac{1}{p_{i,\cdot}} \left[\sum_{j=1}^{M-1} j p_{i,j} + \sum_{k=1}^{N+1} \alpha_k u_{k,i} \frac{x_k^M (M - Mx_k + x_k)}{(1-x_k)^2} \right]. \quad (5.25)$$

The overall average number of jobs in the system, L , is equal to

$$L = \sum_{i=0}^N p_i \cdot L_i. \tag{5.26}$$

5.7. Batch arrivals and/or departures

Consider now a Markov-modulated queue which is not a QBD process, i.e. one where the queue size jumps may be bigger than 1. As before, the state of the process at time t is described by the pair (I_t, J_t) , where I_t is the state of the environment (the operational mode) and J_t is the number of jobs in the system. The state space is the lattice strip $\{0, 1, \dots, N\} \times \{0, 1, \dots\}$. The variable J_t may jump by arbitrary, but bounded amounts in either direction. In other words, the allowable transitions are:

- (a) Phase transitions leaving the queue unchanged: from state (i, j) to state (k, j) ($0 \leq i, k \leq N; i \neq k$), with rate $a_j(i, k)$;
- (b) Transitions incrementing the queue by s : from state (i, j) to state $(k, j + s)$ ($0 \leq i, k \leq N; 1 \leq s \leq r_1; r_1 \geq 1$), with rate $b_{j,s}(i, k)$;
- (c) Transitions decrementing the queue by s : from state (i, j) to state $(k, j - s)$ ($0 \leq i, k \leq N; 1 \leq s \leq r_2; r_2 \geq 1$), with rate $c_{j,s}(i, k)$,

provided of course that the source and destination states are valid.

Obviously, if $r_1 = r_2 = 1$ then this is a Quasi-Birth-and-Death process.

Denote by $A_j = [a_j(i, k)]$, $B_{j,s} = [b_{j,s}(i, k)]$ and $C_{j,s} = [c_{j,s}(i, k)]$, the transition rate matrices associated with (a), (b) and (c), respectively. There is a threshold M , such that

$$A_j = A; \quad B_{j,s} = B_s; \quad C_{j,s} = C_s; \quad j \geq M. \tag{5.27}$$

Defining again the diagonal matrices D^A , D^{B_s} and D^{C_s} , whose i th diagonal element is equal to the i th row sum of A , B_s and C_s , respectively, the balance equations for $j > M + r_1$ can be written in a form analogous to (5.12):

$$\mathbf{v}_j \left[D^A + \sum_{s=1}^{r_1} D^{B_s} + \sum_{s=1}^{r_2} D^{C_s} \right] = \sum_{s=1}^{r_1} \mathbf{v}_{j-s} B_s + \mathbf{v}_j A + \sum_{s=1}^{r_2} \mathbf{v}_{j+s} C_s. \tag{5.28}$$

Similar equations, involving A_j , $B_{j,s}$ and $C_{j,s}$, together with the corresponding diagonal matrices, can be written for $j \leq M + r_1$.

As before, (5.28) can be rewritten as a vector difference equation, this time of order $r = r_1 + r_2$, with constant coefficients:

$$\sum_{\ell=0}^r \mathbf{v}_{j+\ell} Q_\ell = \mathbf{0}; \quad j \geq M. \quad (5.29)$$

Here, $Q_\ell = B_{r_1-\ell}$ for $\ell = 0, 1, \dots, r_1 - 1$,

$$Q_{r_1} = A - D^A - \sum_{s=1}^{r_1} D^{B_s} - \sum_{s=1}^{r_2} D^{C_s},$$

and $Q_\ell = C_{\ell-r_1}$ for $\ell = r_1 + 1, r_1 + 2, \dots, r_1 + r_2$.

The spectral expansion solution of this equation is obtained from the characteristic matrix polynomial

$$Q(x) = \sum_{\ell=0}^r Q_\ell x^\ell. \quad (5.30)$$

The solution is of the form

$$\mathbf{v}_j = \sum_{k=1}^c \alpha_k \mathbf{u}_k x_k^j; \quad j = M, M+1, \dots, \quad (5.31)$$

where x_k are the eigenvalues of $Q(x)$ in the interior of the unit disk, \mathbf{u}_k are the corresponding left eigenvectors, and α_k are constants ($k = 1, 2, \dots, c$). These constants, together with the probability vectors \mathbf{v}_j for $j < M$, are determined with the aid of the state-dependent balance equations and the normalising equation.

There are now $(M + r_1)(N + 1)$ so-far-unused balance equations (the ones where $j < M + r_1$), of which $(M + r_1)(N + 1) - 1$ are linearly independent, plus one normalising equation. The number of unknowns is $M(N + 1) + c$ (the vectors \mathbf{v}_j for $j = 0, 1, \dots, M - 1$), plus the c constants α_k . Hence, there is a unique solution when $c = r_1(N + 1)$.

Proposition 5.2. *The Markov-modulated queue has a steady-state distribution if, and only if, the number of eigenvalues of $Q(x)$ strictly inside the unit disk, each counted according to its multiplicity, is equal to the number of states of the Markovian environment, $N + 1$, multiplied by the largest arrival batch, r_1 . Then, assuming that the eigenvectors of multiple eigenvalues*

are linearly independent, the spectral expansion solution of (5.28) has the form

$$\mathbf{v}_j = \sum_{k=1}^{r_1*(N+1)} \alpha_k \mathbf{u}_k x_k^j; \quad j = M, M + 1, \dots \tag{5.32}$$

For computational purposes, the polynomial eigenvalue-eigenvector problem of degree r can be transformed into a linear one. For example, suppose that Q_r is non-singular and multiply (5.29) on the right by Q_r^{-1} . This leads to the problem

$$\mathbf{u} \left[\sum_{\ell=0}^{r-1} H_\ell x^\ell + I x^r \right] = \mathbf{0}, \tag{5.33}$$

where $H_\ell = Q_\ell Q_r^{-1}$. Introducing the vectors $\mathbf{y}_\ell = x^\ell \mathbf{u}$, $\ell = 1, 2, \dots, r - 1$, one obtains the equivalent linear form

$$[\mathbf{u}, \mathbf{y}_1, \dots, \mathbf{y}_{r-1}] \begin{bmatrix} 0 & & & -H_0 \\ I & 0 & & -H_1 \\ & \ddots & \ddots & \\ & & I & -H_{r-1} \end{bmatrix} = x[\mathbf{u}, \mathbf{y}_1, \dots, \mathbf{y}_{r-1}].$$

As in the quadratic case, if Q_r is singular then the linear form can be achieved by an appropriate change of variable.

5.8. A simple approximation

The spectral expansion solution can be computationally expensive. Its numerical complexity depends crucially on the number of environmental phases: that number determines the number of eigenvalues and eigenvectors that have to be evaluated, and influences the size of the set of simultaneous linear equations that have to be solved. Moreover, when N is large, there may be numerical problems concerned with ill-conditioned matrices. In some cases, both the complexity and the numerical stability of the solution are adversely affected when the system is heavily loaded.

For these reasons, it may be worth abandoning the exact solution, if one can develop a reasonable approximation which is simple, easy to implement, robust and computationally cheap. Such an approximation can be extracted from the spectral expansion solution. The idea is to use a “restricted” expansion, based on a single eigenvalue and its associated eigenvector. The eigenvalue provides a geometric approximation for the

queue size distribution, while the eigenvector approximates the distribution of the environmental phase.

An attractive feature of the geometric approximation is that its accuracy improves when the offered load increases. In the heavy-traffic limit, i.e. when the system approaches saturation, the approximation becomes asymptotically exact.

In order to keep the presentation simple, the discussion will be restricted to QBD Markov-modulated queues whose solution is given by Proposition 5.1, with simple eigenvalues. However, the applicability of the proposed approximation is much more general.

A central role in the approximation is played by the largest eigenvalue that appears in (5.20), and its left eigenvector. Assume, without loss of generality, that the eigenvalues are numbered in increasing order of modulus, so that the largest is x_{N+1} . When the queue is stable, x_{N+1} is real and positive. Moreover, it has a positive eigenvector. From now on, x_{N+1} will be referred to as the “dominant eigenvalue”, and will be denoted by γ .

The expression (5.20) implies that *the tail* of the joint distribution of the queue size and the environmental phase is approximately geometrically distributed, with parameter equal to the dominant eigenvalue, γ . To see that, divide both sides of (5.20) by γ^j and let $j \rightarrow \infty$. Since γ is strictly greater in modulus than all other eigenvalues, all terms in the summation vanish, except one:

$$\lim_{j \rightarrow \infty} \frac{\mathbf{v}_j}{\gamma^j} = \alpha_{N+1} \mathbf{u}_{N+1}. \quad (5.34)$$

In other words, when j is large,

$$\mathbf{v}_j \approx \alpha_{N+1} \mathbf{u}_{N+1} \gamma^j. \quad (5.35)$$

This product form implies that when the queue is large, its size is approximately independent of the environmental phase. The tail of the marginal distribution of the queue size is approximately geometric:

$$p_{\cdot,j} \approx \alpha_{N+1} (\mathbf{u}_{N+1} \cdot \mathbf{1}) \gamma^j, \quad (5.36)$$

where $\mathbf{1}$ is the column vector defined in (5.13).

These results suggest seeking an approximation of the form

$$\mathbf{v}_j = \alpha \mathbf{u}_{N+1} \gamma^j, \quad (5.37)$$

where α is some constant.

Note that γ and \mathbf{u}_{N+1} can be computed without having to find *all* eigenvalues and eigenvectors. There are techniques for determining the

eigenvalues that are near a given number. Here we are dealing with the eigenvalue that is nearest to but strictly less than 1.

If (5.37) is applied to all \mathbf{v}_j , for $j = 0, 1, \dots$, then the approximation depends on just one unknown constant, α . Its value is determined by (5.13) alone, and the expressions for \mathbf{v}_j become

$$\mathbf{v}_j = \frac{\mathbf{u}_{N+1}}{(\mathbf{u}_{N+1} \cdot \mathbf{1})} (1 - \gamma) \gamma^j; \quad j = 0, 1, \dots \quad (5.38)$$

This last approximation avoids completely the need to solve a set of linear equations. Hence, it also avoids all problems associated with ill-conditioned matrices. Moreover, it scales well. The complexity of computing γ and \mathbf{u}_{N+1} grows roughly linearly with N when the matrices A , B and C are sparse. The price paid for that convenience is that the balance equations for $j \leq M$ are no longer satisfied.

Despite its apparent over-simplicity, the geometric approximation (5.38) can be shown to be asymptotically exact when the offered load increases.

5.9. The heavy traffic limit

Consider the case where a parameter associated with arrivals or services changes so that system becomes heavily loaded and approaches saturation. The parameters governing the evolution of the environment are assumed to remain fixed. Then the dominant eigenvalue, γ , is known to approach 1 (Gail *et al.*, [3]). When $\gamma = 1$ (i.e. there is a double eigenvalue at 1), the process $X = \{(I, J)\}$ is recurrent-null; when γ leaves the unit disc, the process is transient. Hence, instead of taking a limit involving a particular parameter, e.g. $\lambda \rightarrow \lambda_{\max}$ (where λ_{\max} is the arrival rate that would saturate the system), we can equivalently treat the heavy-traffic regime in terms of the limit $\gamma \rightarrow 1$.

Since there is no equilibrium distribution when X is recurrent-null, we must have

$$\lim_{\gamma \rightarrow 1} \mathbf{v}_j = \mathbf{0}; \quad j = 0, 1, \dots \quad (5.39)$$

Hence, in order to talk sensibly about the “limiting distribution”, some kind of normalisation must be applied. Multiply the queue size by $1 - \gamma$ and consider the process $Y = \{[I, J(1 - \gamma)]\}$. The limiting joint distribution of Y will be determined by means of the vector Laplace transform

$$\mathbf{h}(s) = [h_0(s), h_1(s), \dots, h_N(s)], \quad (5.40)$$

where

$$h_i(s) = \lim_{\gamma \rightarrow 1} E[\delta(I = i)e^{-s(1-\gamma)J}]; \quad i = 0, 1, \dots, N, \quad (5.41)$$

and $\delta(B)$ is the indicator of the boolean B : it is equal to 1 if B is true, 0 otherwise. In terms of the vectors \mathbf{v}_j , (5.40) is expressed as

$$\mathbf{h}(s) = \lim_{\gamma \rightarrow 1} \sum_{j=0}^{\infty} \mathbf{v}_j e^{-s(1-\gamma)j}. \quad (5.42)$$

The objective will be to show that both the exact distribution, where the vectors \mathbf{v}_j are given by (5.20), and the geometric approximation, where they are given by (5.38), have the same limiting distribution.

Consider first the exact distribution. When all eigenvalues are simple, the equations (5.20) and (5.39) imply that

$$\lim_{\gamma \rightarrow 1} \alpha_k \mathbf{u}_k = \mathbf{0}; \quad k = 1, 2, \dots, N + 1. \quad (5.43)$$

This can be seen by taking $N + 1$ consecutive equations (5.20) and setting their left-hand sides to 0; the Vandermonde matrix involving powers of different eigenvalues is non-singular, and so the only solution is $\alpha_k \mathbf{u}_k = \mathbf{0}$.

On the other hand, since the environmental process has a finite number of states, and since the corresponding transition rates are fixed, the stationary marginal distribution of the environmental phase always exists and has a non-zero limit when $\gamma \rightarrow 1$. Denote that limit by the vector \mathbf{q} . This is the limiting eigenvector corresponding to the eigenvalue 1; it satisfies the equations

$$\mathbf{q}G = \mathbf{0}; \quad (\mathbf{q} \cdot \mathbf{1}) = 1, \quad (5.44)$$

where G is the generator matrix of the environmental process. In terms of the matrix polynomial (5.15), G is the limiting matrix $Q(1) = Q_0 + Q_1 + Q_2$, obtained by replacing the changing traffic parameter with its limit. In particular, if the matrices B and C are diagonal, then $G = A - D^A$.

Hence, we can write

$$\lim_{\gamma \rightarrow 1} \sum_{j=0}^{\infty} \mathbf{v}_j = \mathbf{q}. \quad (5.45)$$

Moreover, in view of (5.39), equation (5.45) holds if the lower index of the summation is $j = M$ (or any other non-negative integer), instead of $j = 0$.

Substituting (5.20) into (5.45) and changing the lower summation index to $j = M$ yields

$$\lim_{\gamma \rightarrow 1} \sum_{k=1}^{N+1} \alpha_k \mathbf{u}_k \frac{x_k^M}{1 - x_k} = \mathbf{q}. \quad (5.46)$$

However, the first N eigenvalues do not approach 1, while the last one, $x_{N+1} = \gamma$, does. Hence, according to (5.43), the first N terms in (5.46) vanish and leave

$$\lim_{\gamma \rightarrow 1} \frac{\alpha_{N+1} \mathbf{u}_{N+1}}{1 - \gamma} = \mathbf{q}. \quad (5.47)$$

Now, substituting (5.20) into (5.42), and arguing as for (5.47), we see that only the term involving the dominant eigenvalue survives:

$$\begin{aligned} \mathbf{h}(s) &= \lim_{\gamma \rightarrow 1} \sum_{j=M}^{\infty} e^{-s(1-\gamma)j} \sum_{k=1}^{N+1} \alpha_k \mathbf{u}_k x_k^j \\ &= \lim_{\gamma \rightarrow 1} \sum_{k=1}^{N+1} \alpha_k \mathbf{u}_k \sum_{j=M}^{\infty} x_k^j e^{-s(1-\gamma)j} \\ &= \lim_{\gamma \rightarrow 1} \sum_{k=1}^{N+1} \alpha_k \mathbf{u}_k \frac{x_k^M e^{-s(1-\gamma)M}}{1 - x_k e^{-s(1-\gamma)}} \\ &= \lim_{\gamma \rightarrow 1} \frac{\alpha_{N+1} \mathbf{u}_{N+1}}{1 - \gamma e^{-s(1-\gamma)}}. \end{aligned} \quad (5.48)$$

Combining this with (5.47) leads to

$$\mathbf{h}(s) = \mathbf{q} \lim_{\gamma \rightarrow 1} \frac{1 - \gamma}{1 - \gamma e^{-s(1-\gamma)}} = \mathbf{q} \frac{1}{1 + s}. \quad (5.49)$$

The last limit follows from L'Hospital's rule. The Laplace transform appearing in the right-hand side of (5.49) is that of the exponential distribution with mean 1. Thus we have established the following rather general result:

Proposition 5.3. *In any Markov-modulated queue, in the heavy-traffic limit $\gamma \rightarrow 1$, the environmental state I and the normalised queue size*

$(1 - \gamma)J$ are independent of each other. The first has distribution \mathbf{q} , while the second is distributed exponentially with mean 1.

It now remains to compare the limit (5.49) with the corresponding one for the geometric approximation, (5.38). Denote the approximate limiting vector Laplace transform by $\hat{\mathbf{h}}(s)$; it is given by (5.42), with \mathbf{v}_j replaced by the approximations (5.38):

$$\begin{aligned}\hat{\mathbf{h}}(s) &= \lim_{\gamma \rightarrow 1} \frac{\mathbf{u}_{N+1}}{(\mathbf{u}_{N+1} \cdot \mathbf{1})} \sum_{j=0}^{\infty} (1 - \gamma) \gamma^j e^{-s(1-\gamma)j} \\ &= \lim_{\gamma \rightarrow 1} \frac{\mathbf{u}_{N+1}}{(\mathbf{u}_{N+1} \cdot \mathbf{1})} \lim_{\gamma \rightarrow 1} \frac{1 - \gamma}{1 - \gamma e^{-s(1-\gamma)}} \\ &= \frac{1}{1 + s} \lim_{\gamma \rightarrow 1} \frac{\mathbf{u}_{N+1}}{(\mathbf{u}_{N+1} \cdot \mathbf{1})},\end{aligned}\tag{5.50}$$

again using L'Hospital's rule.

The last limit in the right-hand side of (5.50) is simply the vector \mathbf{q} . This can be seen by arguing that the normalised left eigenvector of the eigenvalue γ must approach the normalised left eigenvector of the eigenvalue 1. Alternatively, multiply both sides of (5.47) by the column vector $\mathbf{1}$:

$$\lim_{\gamma \rightarrow 1} \frac{\alpha_{N+1}(\mathbf{u}_{N+1} \cdot \mathbf{1})}{1 - \gamma} = 1.\tag{5.51}$$

Hence rewrite (5.47) as

$$\lim_{\gamma \rightarrow 1} \frac{\mathbf{u}_{N+1}}{(\mathbf{u}_{N+1} \cdot \mathbf{1})} = \mathbf{q}.\tag{5.52}$$

Thus we have

$$\hat{\mathbf{h}}(s) = \mathbf{q} \frac{1}{1 + s} = \mathbf{h}(s).\tag{5.53}$$

So, in heavy traffic, the geometric approximation is asymptotically exact, in the sense that it yields the same limiting normalised distribution of environmental phase and queue size as the exact solution.

5.10. Applications and comparisons

It is instructive to present some numerical experiments aimed at evaluating the accuracy of the geometric approximation in the context of two different models of Markov-modulated queues. In all cases, the exact values of the

performance measures are computed by applying the full spectral expansion solution (5.20).

The first system examined is the network of two nodes in tandem, with manufacturing blocking at node 1. The model is illustrated in Fig. 5.3. The parameters are λ (external arrival rate), μ (service rate at node 1), ξ (service rate at node 2) and N (the storage capacity at node 2 is $N - 1$).

In this system, the unbounded queue at node 1 is modulated by a finite-state environment defined by node 2. The environment, I , is in state i if there are i jobs at node 2 and server 1 is not blocked ($i = 0, 1, \dots, N - 1$). An extra state, $I = N$, is needed to describe the situation where there are $N - 1$ jobs at node 2 and server 1 is blocked.

The pair $X = \{(I, J)\}$, where J is the number of jobs at node 1, is a QBD process. The transitions out of state (i, j) were given earlier.

Because the environmental process is coupled with the queueing process, the marginal distribution of the former (i.e. the number of jobs at node 2), cannot be determined without finding the joint distribution of I and J . There is no simple expression for the stability condition.

Figure 5.4 illustrates the close agreement between the exact solution of this model and the geometric approximation (5.38), when the system is heavily loaded. The performance measure is the average size of the unbounded queue; it is plotted against the arrival rate, λ . The service rates

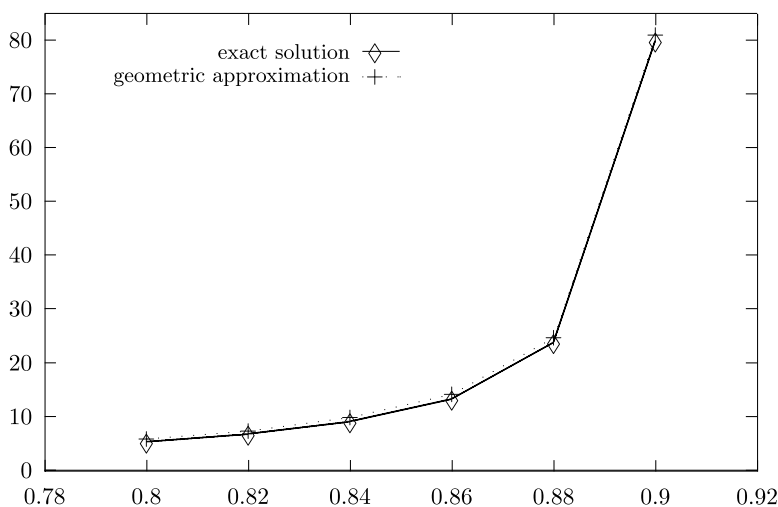


Fig. 5.4. Manufacturing blocking: Average node 1 queue size against arrival rate, $N = 10$, $\mu = 1$, $\xi = 1$.

at nodes 1 and 2 are the same. Hence, the busier node 1, the higher the likelihood that the buffer will fill up and cause blocking. Because of that, the saturation point is not at $\lambda = 1$ (as it would be if node 1 was isolated), but at approximately $\lambda = 0.909$.

The geometric approximation for the marginal distribution of the environmental variable, I , indicating the number of jobs at node 2 and whether or not node 1 is blocked, is given by (5.38) as $\mathbf{q} \approx \mathbf{u}_{N+1}/(\mathbf{u}_{N+1} \cdot \mathbf{1})$. Since there are two environmental states, $I = N-1$ and $I = N$, representing $N-1$ jobs at node 2, the average length of the node 2 queue, L_2 , is given by

$$L_2 = \sum_{i=1}^{N-1} i q_i + (N-1) q_N,$$

where q_i is the $i+1$ st element of the vector \mathbf{q} . Figure 5.5 compares the exact value of L_2 with that provided by the geometric approximation, for the same parameters as in Fig. 5.4. It can be seen that this time the approximation is relatively less accurate, and converges to the exact solution more slowly. Intuitively, this is due to the fact that, in order to obtain an accurate value for L_2 , *all* elements of \mathbf{q} need to be accurate. Whereas, in a heavily loaded unbounded queue, only the tail of the distribution is important.

In Fig. 5.6, the average unbounded queue size is plotted against N . Increasing the size of the finite buffer enlarges the environmental state

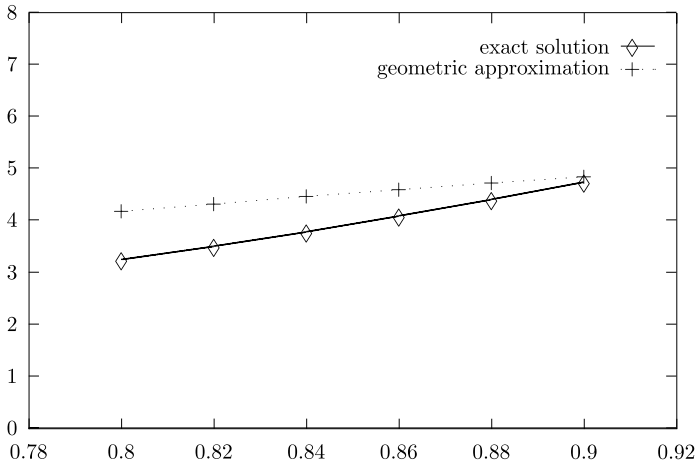


Fig. 5.5. Manufacturing blocking: Average node 2 queue size against arrival rate, $N = 10$, $\mu = 1$, $\xi = 1$.

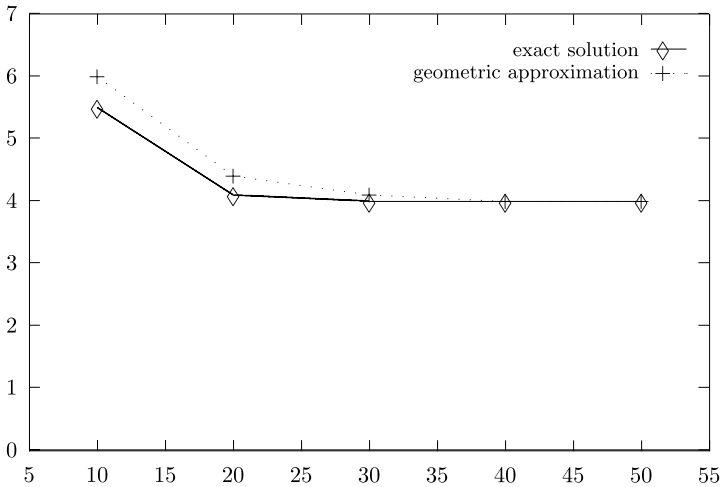


Fig. 5.6. Manufacturing blocking: Average node 1 queue size against N , $\lambda = 0.8$, $\mu = 1$, $\xi = 1$.

space. Consequently, the exact solution needs to compute more eigenvalues and eigenvectors, and solve larger sets of linear equations.

The accuracy of the geometric approximation is seen to increase with N . This is not really surprising, because enlarging the intermediate buffer reduces the coupling between the two nodes, making them behave more like independent queues. Nevertheless, the exact solution begins to experience numerical difficulties when $N > 35$. The software (Matlab) starts issuing warnings to the effect that the matrix is ill-conditioned, and the results may not be reliable (as it happens, the results returned seem fine). Of course the approximation displays no such symptoms, since it has no equations to solve.

The second model to be evaluated is that of the multiserver queue with breakdowns and repairs, described at the beginning of the chapter (Fig. 5.2). The parameters are λ (arrival rate; it will be assumed independent of the operative state of the servers), μ (service rate), ξ (breakdown rate), η (repair rate) and N (number of servers. The queue evolves in a Markovian environment which is in phase i ($i = 0, 1, \dots, N$) when there are i operative servers.

In applying the geometric approximation to this model, there is a choice of approaches. One could use (5.37) for $j \geq N$, together with the balance equations for $j < N$. This will be referred to as the “partial geometric”

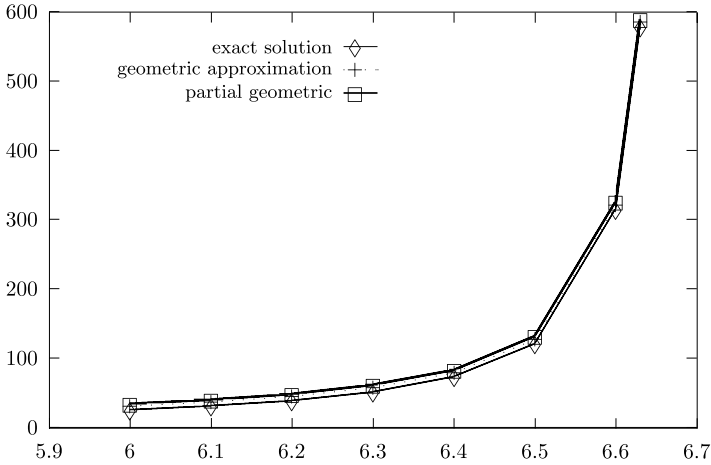


Fig. 5.7. Breakdowns and repairs: Average queue size against arrival rate, $N = 10$, $\mu = 1$, $\xi = 0.05$, $\eta = 0.1$.

approximation. Alternatively, the geometric approximation (5.38) can be used for all $j \geq 0$.

Intuitively, the partial geometric approximation can be expected to be more accurate, since it satisfies more of the balance equations. In fact, the results in Fig. 5.7 suggest that the opposite is true. The average queue size is plotted against the arrival rate, with parameters chosen so that the system is heavily loaded (the saturation point is $\lambda = 6.666\dots$). It turns out that the simple geometric approximation is more accurate than the more complex partial geometric one. There seem to be two opposing effects here. On the one hand, relying only on the dominant eigenvalue tends to overestimate the average queue size; on the other hand, the additional approximation introduced by ignoring the boundary balance equations reduces that overestimation.

Since the marginal distribution of the environmental variable I is known to be given by (5.4), there is not much point in trying to approximate it. However, if the geometric approximation is nevertheless applied, e.g. to compute the average number of operative servers, then a similar picture to Fig. 5.5 emerges. The approximation improves when λ increases, even though the exact value of the average does not depend on λ .

In Fig. 5.8, the average queue size is evaluated for increasing number of servers, and hence decreasing load. This experiment disproves the conjecture that the geometric approximation always overestimates the exact

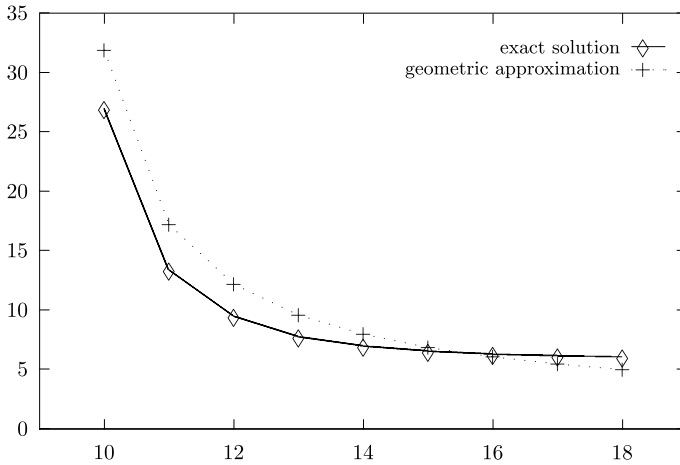


Fig. 5.8. Breakdowns and repairs: Average queue size against number of servers, $\lambda = 6$, $\mu = 1$, $\xi = 0.05$, $\eta = 0.1$.

values. Here the approximation starts off as an overestimate, but as N increases, it becomes an underestimate.

As in the previous model, when N becomes large (greater than about 30), the exact solution begins to warn of possible numerical problems due to ill-conditioned matrices; the geometric approximation does not display such symptoms.

5.11. Remarks

The presentation in this chapter is based on material from [8, 10, 11]. It is perhaps worth mentioning that there are two other solution techniques that can be used in the context of Markov-modulated queues. These are the matrix-geometric method (Neuts, [12]) and the generating functions method (as applied, for example, in [9]). However, we have chosen to concentrate on the spectral expansion solution method because it is versatile, readily implementable and efficient. A strong case can be made for using it, whenever possible, in preference to the other methods [10]. An additional point in its favour is that it provides the basis for a simple approximate solution.

The geometric approximation is valid for a large class of heavily loaded systems. The arguments presented here do not rely on any particular model structure. One could relax the QBD assumption and allow batch arrivals

and departures. As long as there is a spectral expansion solution with finitely many eigenvalues, there would be a single dominant eigenvalue and therefore the geometric approximation would be asymptotically exact in heavy traffic. Moreover, it *may* also be reasonable for moderate and light loads, as the examples in Figs. 5.5 and 5.8 illustrate.

References

1. Buzacott, J. A. and Shanthikumar, J. G. (1993). *Stochastic Models of Manufacturing Systems*, Prentice-Hall.
2. Daigle, J. N. and Lucantoni, D. M. (1991). Queueing systems having phase-dependent arrival and service rates, in *Numerical Solutions of Markov Chains*, (ed. W. J. Stewart), Marcel Dekker.
3. Gail, H. R., Hantler, S. L. and Taylor, B. A. (1996). Spectral analysis of M/G/1 and G/M/1 type Markov chains, *Adv. in Appl. Prob.*, **28**, 114–165.
4. Gohberg, I., Lancaster, P. and Rodman, L. (1982). *Matrix Polynomials*, Academic Press.
5. Jennings, A. (1977). *Matrix Computations for Engineers and Scientists*, Wiley.
6. Konheim, A. G. and Reiser, M. (1976). A queueing model with finite waiting room and blocking, *JACM*, **23**(2), 328–341.
7. Latouche, G., Jacobs, P. A. and Gaver, D. P. (1984). Finite Markov chain models skip-free in one direction, *Naval Res. Log. Quart.*, **31**, 571–588.
8. Mitrani, I. (2005). Approximate Solutions for Heavily Loaded Markov Modulated Queues, *Performance Evaluation*, **62**, 117–131.
9. Mitrani, I. and Avi-Itzhak, B. (1968). A many-server queue with service interruptions, *Operations Research*, **16**(3), 628–638.
10. Mitrani, I. and Chakka, R. (1995). Spectral expansion solution for a class of Markov models: Application and comparison with the matrix-geometric method, *Performance Evaluation*.
11. Mitrani, I. and Mitra, D. (1991). A spectral expansion method for random walks on semi-infinite strips, *IMACS Symposium on Iterative Methods in Linear Algebra*, Brussels.
12. Neuts, M. F. (1981). *Matrix Geometric Solutions in Stochastic Models*, John Hopkins Press.
13. Neuts, M. F. and Lucantoni, D. M. (1979). A Markovian queue with N servers subject to breakdowns and repairs, *Management Science*, **25**, 849–861.