Spectral Expansion Solutions for Markov-Modulated Queues

Isi Mitrani

Computing Science Department, University of Newcastle
isi.mitrani@ncl.ac.uk

1 Introduction

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 $0, 1, \ldots, N$; these are also called the environmental *phases*. The possible values of J_t are $0, 1, \ldots$. 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* (see, for example, Prabhu and Zhu [21]). The corresponding state space, $\{0, 1, \ldots, N\} \times \{0, 1, \ldots\}$ 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 sub-class of models which are sufficiently general to be useful, and yet can be solved efficiently. We shall introduce 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.

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, e.g. in Latouche et al. [12]). The state diagram for this common model, showing some transitions out of state (i, j), is illustrated in figure 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 we impose no limit on the magnitude of the threshold M, although it must be pointed out that

M.C. Calzarossa and S. Tucci (Eds.): Performance 2002, LNCS 2459, pp. 17–35, 2002.

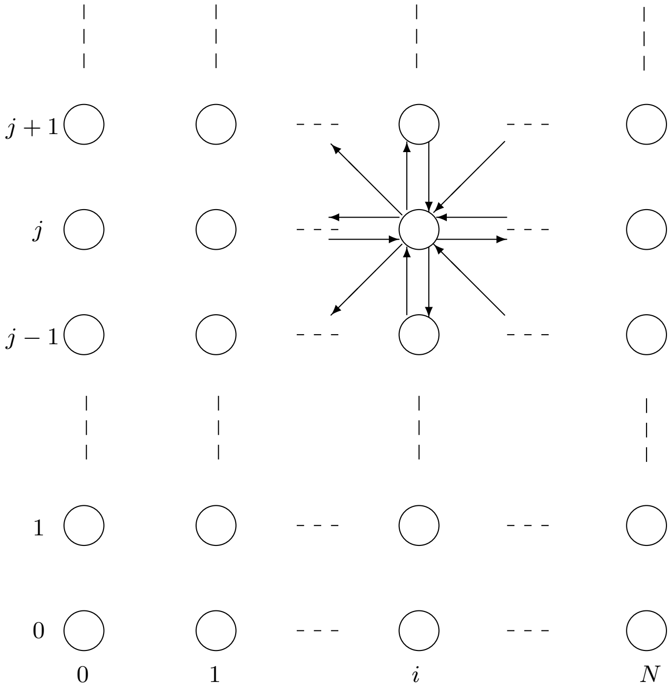

Fig. 1. State diagram of a QBD process

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:

$$p_{i,j} = \lim_{t\to\infty} P(I_t = i\,,\, J_t = j) \;\; ; \;\; i = 0, 1, \ldots, N \;\; ; \;\; j = 0, 1, \ldots \; . \tag{1}$$

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 normalized.

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

$$p_{\cdot,j} = \sum_{i=0}^{N} p_{i,j} \; . \tag{2}$$

$$p_{i,\cdot} = \sum_{j=0}^{\infty} p_{i,j} \ . \tag{3}$$

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

There are three ways of solving Markov-modulated queueing models exactly. Perhaps the most widely used one is the *matrix-geometric* method [18]. This approach relies on determining the minimal positive solution, R, of a non-linear matrix equation; the equilibrium distribution is then expressed in terms of powers of R.

The second method uses generating functions to solve the set of balance equations. A number of unknown probabilities which appear in the equations for those generating functions are determined by exploiting the singularities of the coefficient matrix. A comprehensive treatment of that approach, in the context of a discrete-time process with an M/G/1 structure, is presented in Gail et al. [5].

The third (and arguably best) method is the subject of this tutorial. It is called *spectral expansion*, and is based on expressing the equilibrium distribution of the process in terms of the eigenvalues and left eigenvectors of a certain matrix polynomial. The idea of the spectral expansion solution method has been known for some time (e.g., see Neuts [18]), but there are rather few examples of its application in the performance evaluation literature. Some instances where that solution has proved useful are reported in Elwalid et al. [3], and Mitrani and Mitra [17]; a more detailed treatment, including numerical results, is presented in Mitrani and Chakka [16]. More recently, Grassmann [7] has discussed models where the eigenvalues can be isolated and determined very efficiently. Some comparisons between the spectral expansion and the matrix-geometric solutions can be found in [16] and in Haverkort and Ost [8]. The available evidence suggests that, where both methods are applicable, spectral expansion is faster even if the matrix R is computed by the most efficient algorithm.

The presentation in this tutorial is largely based on the material in chapter 6 of [13] and chapter 13 of [14].

Before describing the details of the spectral expansion solution, it would be instructive to show some examples of systems which are modelled as Markov-modulated queues.

2 Examples of Markov-Modulated Queues

We shall start with a few models of the Quasi-Birth-and-Death type, where the queue size increases and decreases in steps of 1.

2.1 A Multiserver Queue with Breakdowns and Repairs

A single, unbounded queue is served by N identical parallel servers. 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 $\{0, 1, \ldots, N\}$. This is the environment in which the queue evolves: it is in phase i when there are i operative servers (see [15,20]).

Jobs arrive according to a Markov-Modulated Poisson Process controlled by I_t. When the phase is i, 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 figure 2.

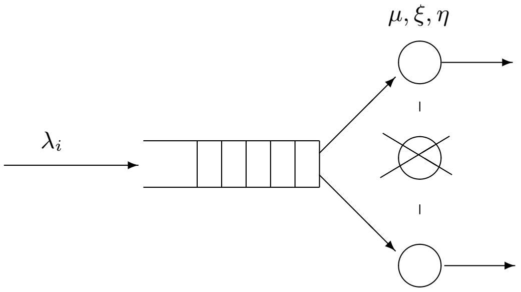

Fig. 2. A multiserver queue with breakdowns and repairs

The process $X = \{(I_t, J_t)\,;\, t \geq 0\}$ is QBD. 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 $\min(i, j)\mu$.

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

Remark. Even if the breakdown and repair processes were more complicated, e.g., if servers could break down and be repaired in batches, or if a server breakdown triggered a job departure, the queueing process would still be QBD. The environmental state transitions can be arbitrary, as long as the queue changes in steps of 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:

$$p_{i,\cdot} = \binom{N}{i} \left(\frac{\eta}{\xi+\eta}\right)^i \left(\frac{\xi}{\xi+\eta}\right)^{N-i} \quad ; \; i = 0, 1, \ldots, N \; . \tag{4}$$

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

$$E(X_t) = \frac{N\eta}{\xi+\eta} \; . \tag{5}$$

The overall average arrival rate is equal to

$$\lambda = \sum_{i=0}^{N} p_{i,\cdot}\lambda_i \; . \tag{6}$$

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

$$\frac{\lambda}{\mu} < \frac{N\eta}{\xi+\eta} \; . \tag{7}$$

2.2 Manufacturing Blocking

Consider a network of two nodes in tandem, such as the one in figure 3. 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 (see [1,19]).

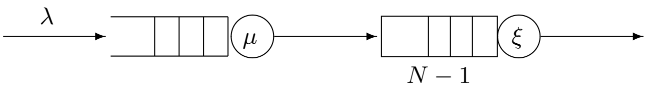

Fig. 3. Two nodes with a finite intermediate buffer

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 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 ($i = 0, 1, \ldots, N-1$). 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 $M = 1$.

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_t and J_t. Nor is the stability condition as simple as in the previous example.

2.3 Phase-Type Distributions

There is a large and useful family of distributions that can be incorporated into queueing models by means of Markovian environments. 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 $\{0, 1, \ldots, N\}$ and generator matrix $\tilde{A}$. States $0, 1, \ldots, N-1$ 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 ($i = 0, 1, \ldots, N-1$; $\alpha_1 + \alpha_2 + \ldots + \alpha_{N-1} = 1$). 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 (see [18]).

The exponential distribution is obviously phase-type ($N = 1$). So is the Erlang distribution—the convolution of N exponentials (exercise 5 in section 2.3). 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 $\alpha_0 = 1$, $\alpha_1 = \ldots = \alpha_{N-1} = 0$.

Another common PH distribution is the 'hyperexponential', where $I_0 = i$ with probability α_i, and absorbtion 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 $c_1X_1 + c_2X_2$ 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 $N_1 N_2^n$.

2.4 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) (see [11,8]).

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.

3 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. 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 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,

$$A_j = A \ ; \ B_j = B \ ; \ C_j = C \ , \ j \geq M \ . \tag{8}$$

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 some of the examples in the previous subsection.

Multiserver Queue with Breakdowns and Repairs

Since the phase transitions are independent of the queue size, 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:

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

Denoting

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

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

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

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

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

Manufacturing Blocking

Remember that the environment changes phase without changing the queue size either when a service completes at node 2 and node 1 is not blocked, or when node 1 becomes blocked (if node 1 is already blocked, then a completion at node 2 changes both phase and queue size). Hence, when $j > 0$,

$$A_j = A = \begin{bmatrix} 0 & 0 & & & \\ \xi & 0 & 0 & & \\ & \ddots & \ddots & \ddots & \\ & & \xi & 0 & \mu \\ & & & 0 & 0 \end{bmatrix} \; ; \; j = 1, 2, \ldots \; .$$

When node 1 is empty ($j = 0$), it cannot become blocked; the state $(N, 0)$ does not exist and the matrix A_0 has only N rows and columns:

$$A_0 = \begin{bmatrix} 0 & & & \\ \xi & 0 & & \\ & \ddots & \ddots & \\ & & \xi & 0 \end{bmatrix} \; ;$$

Since the arrival rate into node 1 does not depend on either i or j, we have $B_j = B = \lambda I$, where I is the identity matrix of order $N + 1$. The departures from node 1 (which can occur when $i \neq N - 1$) are always accompanied by environmental changes: from state (i, j) the system moves to state $(i + 1, j - 1)$ with rate μ for $i < N - 1$; from state (N, j) to state $(N - 2, j - 1)$ with rate ξ. Hence, the departure rate matrices do not depend on j and are equal to

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

Balance Equations

Using the instantaneous transition rates defined at the beginning of this section, the balance equations of a general QBD process can be written as

$$p_{i,j}\sum_{k=0}^{N}[a_j(i,k)+b_j(i,k)+c_j(i,k)]$$

$$=\sum_{k=0}^{N}[p_{k,j}a_j(k,i)+p_{k,j-1}b_{j-1}(k,i)+p_{k,j+1}c_{j+1}(k,i)]\ , \tag{9}$$

where $p_{i,-1}=b_{-1}(k,i)=c_0(i,k)=0$ by definition. The left-hand side of (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:

$$\mathbf{v}_j=(p_{0,j},p_{1,j},\dots,p_{N,j})\ ;\ j=0,1,\dots\ . \tag{10}$$

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 (9), for $j=0,1,\dots$, can be written as:

$$\mathbf{v}_j[D_j^A+D_j^B+D_j^C]=\mathbf{v}_{j-1}B_{j-1}+\mathbf{v}_jA_j+\mathbf{v}_{j+1}C_{j+1}\ , \tag{11}$$

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 (11) cease to depend on j :

$$\mathbf{v}_j[D^A+D^B+D^C]=\mathbf{v}_{j-1}B+\mathbf{v}_jA+\mathbf{v}_{j+1}C\ , \tag{12}$$

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

In addition, all probabilities must sum up to 1:

$$\sum_{j=0}^{\infty}\mathbf{v}_j\mathbf{e}=1\ , \tag{13}$$

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

The first step of any solution method is to find the general solution of the infinite set of balance equations with constant coefficients, (12). The latter are normally written in the form of a homogeneous vector difference equation of order 2:

$$\mathbf{v}_jQ_0+\mathbf{v}_{j+1}Q_1+\mathbf{v}_{j+2}Q_2=\mathbf{0}\ ;\ j=M,M+1,\dots\ , \tag{14}$$

where $Q_0=B$, $Q_1=A-D^A-D^B-D^C$ and $Q_2=C$. There is more than one way of solving such equations.

Associated with equation (14) is the so-called 'characteristic matrix polynomial', $Q(x)$, defined as

$$Q(x) = Q_0 + Q_1 x + Q_2 x^2 \ . \tag{15}$$

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

$$\begin{aligned} det[Q(x_k)] &= 0 \ , \\ \mathbf{u}_k Q(x_k) &= \mathbf{0} \ ; \ k = 1, 2, \dots, d \ , \end{aligned} \tag{16}$$

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

$$\mathbf{v}_{k,j} = \mathbf{u}_k x_k^{j-M} \ ; \ j = M, M+1, \dots \ , \tag{17}$$

is a solution of equation (14). Indeed, substituting (17) into (14) we get

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

By combining any multiple eigenvalues with each of their independent eigenvectors, we thus obtain d linearly independent solutions of (14). On the other hand, it is known that there cannot be more than d linearly independent solutions. Therefore, any solution of (14) can be expressed as a linear combination of the d solutions (17):

$$\mathbf{v}_j = \sum_{k=1}^{d} \alpha_k \mathbf{u}_k x_k^{j-M} \ ; \ j = M, M+1, \dots \ , \tag{18}$$

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 normalized to become probability distributions. Hence, it is necessary to select from the set (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 $\mathbf{u}_1$, $\mathbf{u}_2$, …, $\mathbf{u}_c$. Then any normalizable solution of equation (14) can be expressed as

$$\mathbf{v}_j = \sum_{k=1}^{c} \alpha_k \mathbf{u}_k x_k^{j-M} \ ; \ j = M, M+1, \dots \ , \tag{19}$$

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

Expression (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 (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 $\mathbf{v}_M$, $\mathbf{v}_{M+1}$, ...; these contain c unknown constants. Now it is time to consider the balance equations (11), for $j = 0, 1, \ldots, M$. This is a set of $(M+1)(N+1)$ linear equations with $M(N+1)$ unknown probabilities (the vectors $\mathbf{v}_j$ for $j = 0, 1, \ldots, M-1$), 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 (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

Proposition 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 (12) has the form*

$$\mathbf{v}_j = \sum_{k=1}^{N+1} \alpha_k \mathbf{u}_k x_k^{j-M} \; ; \; j = M, M+1, \ldots \; . \tag{20}$$

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 (11), for $j = 0, 1, \ldots, M$, and (13), with $\mathbf{v}_M$ and $\mathbf{v}_{M+1}$ given by (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

$$\mathbf{u}[Q_0 + Q_1 x + Q_2 x^2] = \mathbf{0} \; , \tag{21}$$

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 linearization can be achieved quite easily if the matrix $C = Q_2$ is non-singular. Indeed, after multiplying (21) on the right by Q_2^{-1}, it becomes

$$\mathbf{u}[H_0 + H_1 x + I x^2] = \mathbf{0} , \tag{22}$$

where $H_0 = Q_0 C^{-1}$, $H_1 = Q_1 C^{-1}$, and I is the identity matrix. By introducing the vector $\mathbf{y} = x\mathbf{u}$, equation (22) can be rewritten in the equivalent linear form

$$[\mathbf{u}, \mathbf{y}] \begin{bmatrix} 0 & -H_0 \\ I & -H_1 \end{bmatrix} = x[\mathbf{u}, \mathbf{y}] . \tag{23}$$

If C is singular but B is not, a similar linearization is achieved by multiplying (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 $S = \gamma^2 Q_2 + \gamma Q_1 + Q_0$ 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 (22).

The computational demands of step 2 may be high if the threshold M is large. However, if the matrices B_j ($j = 0, 1, \ldots, M-1$) are non-singular (which is often the case in practice), then the vectors $\mathbf{v}_{M-1}, \mathbf{v}_{M-2}, \ldots, \mathbf{v}_0$ can be expressed in terms of $\mathbf{v}_M$ and $\mathbf{v}_{M+1}$, with the aid of equations (11) for $j = M, M-1, \ldots, 1$. One is then left with equations (11) for $j = 0$, plus (13) (a total of $N+1$ independent linear equations), for the $N+1$ unknowns x_k.

Having determined the coefficients in the expansion (19) 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

$$p_{i,\cdot} = \sum_{j=0}^{M-1} p_{i,j} + \sum_{k=1}^{N+1} \alpha_k u_{k,i} \frac{1}{1 - x_k} , \tag{24}$$

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

$$L_i = \frac{1}{p_{i,\cdot}} \left[\sum_{j=1}^{M-1} j p_{i,j} + \sum_{k=1}^{N+1} \alpha_k u_{k,i} \frac{M - (M-1)x_k}{(1 - x_k)^2} \right] . \tag{25}$$

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

$$L = \sum_{i=0}^{N} p_{i,\cdot} L_i . \tag{26}$$

The spectral expansion solution can also be used to provide simple estimates of performance when the system is heavily loaded. The important observation in this connection is that when the system approaches instability, the expansion (19) is dominated by the eigenvalue with the largest modulus inside the unit disk, x_{N+1}. That eigenvalue is always real. It can be shown that when the offered load is high, the average number of jobs in the system is approximately equal to $x_{N+1}/(1 - x_{N+1})$.

3.1 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) $(0 \le i, k \le N\ ;\ i \ne k)$, with rate $a_j(i, k)$;
(b) Transitions incrementing the queue by s: from state (i, j) to state $(k, j + s)$ $(0 \le i, k \le N\ ;\ 1 \le s \le r_1\ ;\ r_1 \ge 1)$, with rate $b_{j,s}(i, k)$;
(c) Transitions decrementing the queue by s: from state (i, j) to state $(k, j - s)$ $(0 \le i, k \le N\ ;\ 1 \le s \le r_2\ ;\ r_2 \ge 1)$, 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

$$A_j = A\ ;\ B_{j,s} = B_s\ ;\ C_{j,s} = C_s\ ;\ j \ge M\ . \tag{27}$$

Defining again the diagonal matrices D^A, D^{B_s} and D^{C_s}, whose ith diagonal element is equal to the ith 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 (12):

$$\mathbf{v}_j[D^A + \sum_{s=1}^{r_1} D^{B_s} + \sum_{s=1}^{r_2} D^{C_s}] = \sum_{s=1}^{r_1} \mathbf{v}_{j-s} B_s + \mathbf{v}_j A + \sum_{s=1}^{r_2} \mathbf{v}_{j+s} C_s\ . \tag{28}$$

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

As before, (28) can be rewritten as a vector difference equation, this time of order $r = r_1 + r_2$, with constant coefficients:

$$\sum_{\ell=0}^{r} \mathbf{v}_{j+\ell} Q_\ell = \mathbf{0}\ ;\ j \ge M\ . \tag{29}$$

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

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

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

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

$$Q(x) = \sum_{\ell=0}^{r} Q_\ell x^\ell . \tag{30}$$

The solution is of the form

$$\mathbf{v}_j = \sum_{k=1}^{c} \alpha_k \mathbf{u}_k x_k^{j-M} \ ; \ j = M, M+1, \ldots , \tag{31}$$

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 $(k = 1, 2, \ldots, c)$. These constants, together with the the probability vectors $\mathbf{v}_j$ for $j < M$, are determined with the aid of the state-dependent balance equations and the normalizing 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 normalizing equation. The number of unknowns is $M(N+1)+c$ (the vectors $\mathbf{v}_j$ for $j = 0, 1, \ldots, M-1$), plus the c constants α_k. Hence, there is a unique solution when $c = r_1(N + 1)$.

Proposition 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 (28) has the form*

$$\mathbf{v}_j = \sum_{k=1}^{r_1*(N+1)} \alpha_k \mathbf{u}_k x_k^{j-M} \ ; \ j = M, M+1, \ldots . \tag{32}$$

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 (29) on the right by Q_r^{-1}. This leads to the problem

$$\mathbf{u} \left[\sum_{\ell=0}^{r-1} H_\ell x^\ell + I x^r \right] = \mathbf{0} , \tag{33}$$

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.

Example: Checkpointing and Recovery

Consider the transaction processing system described in section 2.4. Here $r_1 = N$ and $r_2 = 1$ (the queue size is incremented by 1 when jobs arrive and by $1, 2, \dots, N$ when faults occur; it is decremented by 1 when a transaction completes service. The threshold M is equal to 0. The matrices A, B_s and C_s are given by:

$$A_j = A = \begin{bmatrix} 0 & & & \\ 0 & 0 & & \\ \vdots & & & \\ \beta & 0 & \dots & 0 \end{bmatrix} \quad ; \ j = 0, 1, \dots \ .$$

The only transition which changes the environment, but not the queue, is the establishment of a checkpoint in state (N, j).

$$B_{j,1} = B_1 = \begin{bmatrix} \lambda & & & & \\ \xi & \lambda & & & \\ 0 & 0 & \lambda & & \\ & & & \ddots & \\ & & & & \lambda \end{bmatrix} \quad ; \ j = 0, 1, \dots \ .$$

The queue size increases by 1 when a job arrives, causing a transition from (i, j) to $(i, j+1)$, and also when a fault occurs in state $(1, j)$; then the new state is $(0, j+1)$.

$$B_{j,2} = B_2 = \begin{bmatrix} 0 & & & \\ 0 & 0 & & \\ \xi & 0 & 0 & \\ \vdots & & & \\ 0 & 0 & \dots & 0 \end{bmatrix} \quad ; \ j = 0, 1, \dots \ .$$

The queue size increases by 2 when a fault occurs in state $(2, j)$, causing a transition to state $(0, j+2)$. The other B_s matrices have a similar form, until

$$B_{j,N} = B_N = \begin{bmatrix} 0 & & & \\ 0 & 0 & & \\ \vdots & & & \\ \xi & 0 & \dots & 0 \end{bmatrix} \quad ; \ j = 0, 1, \dots \ .$$

There is only one matrix corresponding to decrementing queue:

$$C_{j,1} = C_1 = \begin{bmatrix} 0 & \mu & & & \\ & 0 & \mu & & \\ & & \ddots & \ddots & \\ & & & 0 & \mu \\ & & & & 0 \end{bmatrix} ; \; j = 1, 2, \ldots \; .$$

The matrix polynomial $Q(x)$ is of degree $N + 1$. According to Proposition 2, the condition for stability is that the number of eigenvalues in the interior of the unit disk is $N(N + 1)$.

References

1. J.A. Buzacott and J.G. Shanthikumar, *Stochastic Models of Manufacturing Systems*, Prentice-Hall, 1993.
2. J.N. Daigle and D.M. Lucantoni, Queueing systems having phase-dependent arrival and service rates, in *Numerical Solutions of Markov Chains*, (ed. W.J. Stewart), Marcel Dekker, 1991.
3. A.I. Elwalid, D. Mitra and T.E. Stern, Statistical multiplexing of Markov modulated sources: Theory and computational algorithms, *Int. Teletraffic Congress*, 1991.
4. M. Ettl and I. Mitrani, Applying spectral expansion in evaluating the performance of multiprocessor systems, *CWI Tracts* (ed. O. Boxma and G. Koole), 1994.
5. H.R. Gail, S.L. Hantler and B.A. Taylor, Spectral analysis of M/G/1 type Markov chains, *RC17765, IBM Research Division*, 1992.
6. I. Gohberg, P. Lancaster and L. Rodman, *Matrix Polynomials*, Academic Press, 1982.
7. W.K. Grassmann and S. Drekic, An analytical solution for a tandem queue with blocking, *Queueing Systems*, 36, pp. 221–235, 2000.
8. B.R. Haverkort and A. Ost, Steady-State Analysis of Infinite Stochastic Petri Nets: Comparing the Spectral Expansion and the Matrix-Geometric Method, *Procs., 7th Int. Workshop on Petri Nets and Performance Models*, San Malo, 1997.
9. A. Jennings, *Matrix Computations for Engineers and Scientists*, Wiley, 1977.
10. A.G. Konheim and M. Reiser, A queueing model with finite waiting room and blocking, *JACM*, 23, 2, pp. 328–341, 1976.
11. L. Kumar, M. Misra and I. Mitrani, Analysis of a Transaction System with Checkpointing, Failures and Rollback, *Computer Performance Evaluation* (Eds T. Field, P.G. Harrison and U. Harder), LNCS 2324, Springer, 2002.
12. G. Latouche, P.A. Jacobs and D.P. Gaver, Finite Markov chain models skip-free in one direction, *Naval Res. Log. Quart.*, 31, pp. 571–588, 1984.
13. I. Mitrani, *Probabilistic Modelling*, Cambridge University Press, 1998.
14. I. Mitrani, The Spectral Expansion Solution Method for Markov Processes on Lattice Strips, Chapter 13 in *Advances in Queueing*, (Ed. J.H. Dshalalow), CRC Press, 1995.
15. I. Mitrani and B. Avi-Itzhak, A many-server queue with service interruptions, *Operations Research*, 16, 3, pp.628-638, 1968.

16. I. Mitrani and R. Chakka, Spectral expansion solution for a class of Markov models: Application and comparison with the matrix-geometric method, to appear in *Performance Evaluation*, 1995.
17. I. Mitrani and D. Mitra, A spectral expansion method for random walks on semi-infinite strips, *IMACS Symposium on Iterative Methods in Linear Algebra*, Brussels, 1991.
18. M.F. Neuts, *Matrix Geometric Solutions in Stochastic Models*, John Hopkins Press, 1981.
19. M.F. Neuts, Two queues in series with a finite intermediate waiting room, *J. Appl. Prob.*, 5, pp. 123–142, 1968.
20. M.F. Neuts and D.M. Lucantoni, A Markovian queue with N servers subject to breakdowns and repairs, *Management Science*, 25, pp. 849–861, 1979.
21. N.U. Prabhu and Y. Zhu, Markov-modulated queueing systems, *QUESTA*, 5, pp. 215–246, 1989.