Application and Evaluation of the Spectral Expansion Solution Method

Isi Mitrani Ram Chakka*

Abstract

Many two-dimensional Markov models whose state space is a semi-infinite strip (i.e. finite in one dimension and infinite in the other) can be solved efficiently by means of spectral expansion. This method and its application are described in the context of an $M/M/N$ queue with general breakdowns and repairs. The results of experiments aimed at evaluating the relative merits of the spectral expansion and the matrix-geometric solutions are also presented.

1 Introduction

There is a large class of models which involve two-dimensional Markov processes on semi-infinite lattice strips. That is, the system state is described by two integer random variables, I and J; one of these has a finite range, and the other can take any non-negative value. Often these models are cast in the framework of a Markov-modulated queue: then the bounded variable, I, indicates the state of the Markovian environment, while the unbounded one, J, represents the number of jobs in the system (a recent survey can be found in Prabhu and Zhu [17]).

We are interested in a sub-class of the above processes, characterized by the following two properties:

(i) the instantaneous transition rates out of state (i, j) do not depend on j when the latter is sufficiently large;

(ii) the jumps of the random variable J are limited in size.

When the jumps of J are of size 1, the process is said to be of the *Quasi-Birth-and-Death* type (the term *skip-free* is also used, e.g. in Latouche et.al. [9]).

There are four known methods for solving such models. The first is to truncate the state space in the direction of the variable J, and exploit the structure of the resulting finite-state process in devising an efficient solution algorithm. Such an algorithm, based on aggregation-disaggregation, is developed in Seelen [15]. See also Keilson et. al. [8].

The second method is to reduce the infinite-state problem to a linear equation involving a vector generating function and some unknown probabilities. The latter are then determined with the aid of the singularities of the coefficient matrix. A comprehensive

*Computing Laboratory, University of Newcastle, Newcastle upon Tyne NE1 7RU, England. Email: ram.chakka@newcastle.ac.uk ; isi.mitrani@newcastle.ac.uk

treatment of that approach, in the context of a discrete-time process with a more general M/G/1 structure, is presented in Gail et. al. [4].

The third method, which is the subject of this paper, is called *spectral expansion*. It is based on expressing the invariant vector of the process in terms of the eigenvalues and left eigenvectors of a certain matrix polynomial. The idea of the spectral expansion method has been known for some time (e.g., see Neuts [13], Daigle and Lucantoni [2]), but there are very few examples of its application in the performance evaluation literature (some recent instances where that solution has proved useful are reported in Elwalid et. al. [3], and Mitra and Mitrani [11]; an earlier version of the present solution, without any comparisons, was presented in [1]). The generating function and the spectral expansion methods are closely related. However, the latter produces steady-state probabilities directly, whereas the former provides them through a transform.

The fourth way of solving these models is by the *matrix-geometric* method [13]. This widely used approach relies on determining the minimal positive solution, R, of a non-linear matrix equation; the invariant vector is then expressed in terms of powers of R.

The purpose of this paper is two-fold. First, we wish to popularize the spectral expansion method by describing its application in the context of a non-trivial problem. The model in question involves an M/M/N queue where job behaviour and processor availability are governed by a general Markovian environment. In that framework, multiple simultaneous breakdowns or repairs can be allowed, including ones triggered by job arrivals or departures. Moreover, the arrivals and departures may occur in batches of fixed or random size (in the latter case the batch sizes should be bounded). This model is interesting in its own right as a rather general example of a Markov-modulated queue (for some related work, see [10, 14, 16]).

The second objective is to compare the effectiveness of the spectral expansion and the matrix-geometric solutions. Such a comparison has not been carried out before. A number of numerical experiments are performed, where the same examples are solved by both methods. Issues of speed, numerical stability and accuracy are considered. The available results are strongly in favour of the spectral expansion method.

The model is described in section 2. To keep the presentation simple, only single arrivals and departures are considered initially, and all processors are assumed identical. The spectral expansion solution is presented in section 3. Section 4 shows how the approach generalises to models with batch arrivals and departures and with non-identical processors. Section 5 is devoted to the comparison between the spectral expansion and the matrix-geometric solutions.

2 The model

Consider a system where N identical parallel processors serve a common, unbounded job queue. The processors break down from time to time, and return to an operative state after being repaired. When operative, each processor serves jobs, one at a time, and each job is allowed to occupy at most one operative processor at a time. The policy concerning services interrupted by breakdowns can be thought of as either *resume*, or *repeat with re-sampling*; however, the precise specification is contained in the process evolution assumptions given below.

Let $I(t)$ and $J(t)$ be the random variables representing the number of operative processors at time t, and the number of jobs in the system at time t, respectively. We shall sometimes refer to $I(t)$ as the *operative state* of the set of processors. It is assumed that $X = \{[I(t), J(t)] \; ; \; t \geq 0\}$ is an irreducible Markov process with state space $\{0, 1, \ldots, N\} \times \{0, 1, \ldots\}$. The evolution of that process proceeds according to the following set of possible transitions:

(a) From state (i, j) to state (k, j) $(0 \leq i, k \leq N \; ; \; i \neq k)$;

(b) From state (i, j) to state $(k, j+1)$ $(0 \leq i, k \leq N)$;

(c) From state (i, j) to state $(k, j-1)$ $(0 \leq i, k \leq N)$.

We assume further that there is a threshold, M, $(M \geq 1)$ such that the instantaneous transition rates do not depend on j when $j \geq M$. In other words, if we denote the transition rate matrices associated with (a), (b) and (c) by A_j, B_j and C_j respectively (the main diagonal of A_j is zero by definition; also, $C_0 = 0$ by definition), then we have

$$A_j = A \; ; \; B_j = B \; ; \; C_j = C \; , \; j \geq M \; . \tag{1}$$

Transitions (a) correspond to changes in the number of operative processors. For example, if all processors break down and are repaired independently of each other and of the number of jobs present, with rates ξ and η per processor respectively, then the matrices A and A_j would be tri-diagonal, given by:

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

Clearly, a large class of different breakdown and repair patterns can be modelled by choosing A and A_j appropriately.

A transition of type (b) represents a job arrival coinciding with a change in the number of operative processors. If arrivals are not accompanied by such changes, then the matrices B and B_j are diagonal. For example, if jobs arrive according to an independent Poisson process with rate σ_i when there are i operative processors, then

$$B = B_j = diag[\sigma_0, \sigma_1, \ldots, \sigma_N] \; . \tag{3}$$

Similarly, a transition of type (c) represents a job departure coinciding with a change in the number of operative processors. Again, if such coincidences do not occur, then the matrices C and C_j are diagonal. In particular, if each operative processor serves jobs at rate μ, then

$$\begin{aligned} C_j &= diag[0, min(j,1)\mu, min(j,2)\mu, \ldots, min(j,N)\mu] \; ; \; j < N \; , \\ C &= diag[0, \mu, 2\mu, \ldots, N\mu] \; ; \; j \geq N \; . \end{aligned} \tag{4}$$

The requirement that all transition rates, and in particular those affecting the reliability of the system, cease to depend on the size of the job queue beyond a certain threshold, is not too restrictive. It enables, for example, the consideration of models where the the breakdown and repair rates depend on the number of busy processors. On the other hand, it is difficult to think of applications where those rates would depend on the number of jobs waiting to begin execution. Note that we impose no limit on the magnitude of the threshold M, although it must be pointed out that the larger M is, the greater may be the complexity of the solution. Of course, if the only dependencies of the various transition rates on j are the ones manifested in equations (4) and/or the ones mentioned in this paragraph, then the threshold is equal to the number of processors: $M = N$.

It is clear that, although we have been talking in terms of N servers subject to breakdowns and repairs, the above is in fact a general Quasi-Birth-and-Death process which can be used to model a large class of queues in Markovian environments.

As well as the matrices A_j, B_j and C_j, it will be convenient to define the diagonal matrices D_j^A, D_j^B and D_j^C, whose i^{th} diagonal element is the i^{th} row sum of A_j, B_j and C_j, respectively. Those row sums are the total rates at which the process X leaves state (i, j), due to (a) changes in the number of operative servers, (b) job arrivals (perhaps accompanied by such a change) and (c) job departures (ditto), respectively. The j-independent versions of these diagonal matrices are denoted by D^A, D^B and D^C, respectively.

The object of the analysis is to determine the joint steady-state distribution of the number of operative processors and the number of jobs in the system:

$$p_{i,j} = \lim_{t\to\infty} P(I(t) = i, J(t) = j) \; ; \; i = 0, 1, \dots, N \; ; \; j = 0, 1, \dots \; . \tag{5}$$

That distribution exists for an irreducible Markov process if, and only if, the corresponding set of balance equations has a unique normalisable solution.

Once the probabilities $p_{i,j}$ are known, various performance measures of interest can be obtained easily.

3 Spectral expansion solution

Rather than working with the two-dimensional distribution $\{p_{i,j}\}$, we shall introduce the row vectors,

$$\mathbf{v}_j = (p_{0,j}, p_{1,j}, \dots, p_{N,j}) \; ; \; j = 0, 1, \dots \; , \tag{6}$$

whose elements represent the states with j jobs in the system.

The balance equations satisfied by the probabilities $p_{i,j}$ can be written in terms of the vectors $\mathbf{v}_j$. They are:

$$\mathbf{v}_j[D_j^A + D_j^B + D_j^C] = \mathbf{v}_{j-1}B_{j-1} + \mathbf{v}_j A_j + \mathbf{v}_{j+1}C_{j+1} \; ; \; j = 0, 1, \dots M \tag{7}$$

(where $\mathbf{v}_{-1} = 0$ by definition), and

$$\mathbf{v}_j[D^A + D^B + D^C] = \mathbf{v}_{j-1}B + \mathbf{v}_j A + \mathbf{v}_{j+1}C \; ; \; j = M+1, M+2, \dots \; . \tag{8}$$

In addition, all probabilities must sum up to 1:

$$\sum_{j=0}^{\infty} \mathbf{v}_j \mathbf{e} = 1 \; , \tag{9}$$

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

We shall start by finding the general solution of equation (8). That equation has the nice property that its coefficients do not depend on j. It can be rewritten in the form

$$\mathbf{v}_j Q_0 + \mathbf{v}_{j+1} Q_1 + \mathbf{v}_{j+2} Q_2 = \mathbf{0} \; ; \; j = M, M+1, \dots \; , \tag{10}$$

where $Q_0 = B$, $Q_1 = A - D^A - D^B - D^C$ and $Q_2 = C$. This is a homogeneous vector difference equation of order 2, with constant coefficients. Associated with it is the characteristic matrix polynomial, $Q(\lambda)$, defined as

$$Q(\lambda) = Q_0 + Q_1\lambda + Q_2\lambda^2 \; . \tag{11}$$

Denote by λ_k and $\boldsymbol{\psi}_k$ the eigenvalues and corresponding left eigenvectors of $Q(\lambda)$. In other words, these are quantities which satisfy

$$\boldsymbol{\psi}_k Q(\lambda_k) = \mathbf{0} \; ; \; k = 1, 2, \ldots, d \; , \tag{12}$$

where $d = degree\{det[Q(\lambda)]\}$.

The eigenvalues do not have to be simple, but we shall assume that if λ_k has multiplicity m, then it also has m linearly independent left eigenvectors. This is invariably observed to be the case in practice. Some experiments in that direction are mentioned in section 5.

It is readily seen, by direct substitution, that for every eigenvalue λ_k, and corresponding left eigenvector $\boldsymbol{\psi}_k$, the sequence

$$\{\boldsymbol{\psi}_k \lambda_k^j \; ; \; j = M, M+1, \ldots \} \; , \tag{13}$$

is a solution of equation (10). By combining multiple eigenvalues with each of their independent eigenvectors, we thus obtain a total of d linearly independent solutions. On the other hand, it is known (e.g., see [5]), that the dimensionality of the solution space of equation (10) is exactly d. Therefore, any solution of (10) can be expressed as a linear combination of the d solutions (13):

$$\mathbf{v}_j = \sum_{k=1}^{d} x_k \boldsymbol{\psi}_k \lambda_k^j \;\; ; \;\; j = M, M+1, \ldots \; , \tag{14}$$

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

However, we are only interested in solutions which are, or can be normalised to become, probability distributions. Hence, we wish to select from the set (14), those sequences for which the series $\sum \mathbf{v}_j \mathbf{e}$ converges. Moreover, the component series

$$\sum_{j=M}^{\infty} \mathbf{v}_{j,i} = \sum_{j=M}^{\infty} \sum_{k=1}^{d} x_k \boldsymbol{\psi}_{k,i} \lambda_k^j \; ,$$

must converge for every $i = 0, 1, \ldots, N$, in order to ensure the existence of the marginal probability that i processors are operative. These requirements imply that if $|\lambda_k| \geq 1$, then the corresponding coefficient x_k must be 0. We have thus arrived at the following

Proposition 1 *Suppose that c of the eigenvalues of $Q(\lambda)$ are strictly inside the unit disk (each counted according to its multiplicity), while the others are on the circumference or outside. Let the numbering be such that $|\lambda_k| < 1$ for $k = 1, 2, \ldots, c$. The corresponding independent eigenvectors are $\boldsymbol{\psi}_1$, $\boldsymbol{\psi}_2$, ..., $\boldsymbol{\psi}_c$. Then any solution of equation (8) which can be normalised to a probability distribution is of the form*

$$\mathbf{v}_j = \sum_{k=1}^{c} x_k \boldsymbol{\psi}_k \lambda_k^j \;\; ; \;\; j = M, M+1, \ldots \; , \tag{15}$$

where x_k $(k = 1, 2, \ldots, c)$, are arbitrary constants.

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 x_k, in order that the right-hand side of (15) be real. To ensure that it is also positive, it seems that the real parts of λ_k, $\boldsymbol{\psi}_k$ and x_k should be positive. Indeed, that is always found to be the case.

Clearly, the powers λ_k^j in (15) can be replaced by λ_k^{j+m}, for any fixed m, without altering the validity of the proposition. Such a change has to be made, for instance, when one or more of the eigenvalues are zero. Then the expansion should be of the form

$$\mathbf{v}_j = \sum_{k=1}^{c} x_k \boldsymbol{\psi}_k \lambda_k^{j-M} \quad ; \quad j = M, M+1, \dots ,$$

in order to make use of the corresponding eigenvectors.

So far, we have obtained expressions for the vectors $\mathbf{v}_M, \mathbf{v}_{M+1}, \dots$, which contain c unknown constants. Now it is time to consider equations (7), for $j = 0, 1, \dots, M$. This is a set of $(M+1)\times(N+1)$ linear equations with $M\times(N+1)$ unknown probabilities (the vectors $\mathbf{v}_j$ for $j = 0, 1, \dots, M-1$), plus the c constants x_k. However, only $(M+1)\times(N+1)-1$ 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 (9).

Clearly, this set of $(M+1)\times(N+1)$ equations with $M\times(N+1)+c$ unknowns will have a unique solution if, and only if, $c = N+1$. This observation, together with the fact that an irreducible Markov process has a steady-state distribution if, and only if, its balance and normalisation equations have a unique solution, implies

Proposition 2 *The condition $c = N+1$ (the number of eigenvalues of $Q(\lambda)$ strictly inside the unit disk is equal to the number of operative states of the set of processors), is necessary and sufficient for the ergodicity of the Markov process X.*

If job arrivals and departures do not trigger processor breakdowns or repairs (i.e., the matrices B, B_j, C and C_j are diagonal), and if the breakdown and repair rates do not depend on the number of jobs in the system (i.e., $A_j = A$; $j = 0, 1, \dots$), then there is an intuitively appealing probabilistic condition for ergodicity. Denote by $\mathbf{v}$ the marginal distribution of the number of operative processors:

$$\mathbf{v} = (p_{0\cdot}, p_{1\cdot}, \dots, p_{N\cdot}) = \sum_{j=0}^{\infty} \mathbf{v}_j \, .$$

This is the stationary probability vector of the matrix $A - D^A$, and can be obtained by solving the equations

$$\mathbf{v}(A - D^A) = 0 \quad ; \quad \mathbf{ve} = 1 \, . \tag{16}$$

It is then possible to assert that the system is stable if, and only if, when the queue is long, the average incoming load is less than the average service capacity [13]:

$$\mathbf{v}B\mathbf{e} < \mathbf{v}C\mathbf{e} \, . \tag{17}$$

(the matrices B_j and C_j, which are relevant only for queue sizes below the threshold M, play no role here).

Condition (17) is thus equivalent to the requirement in Proposition 2.

Numerical experiments have shown that, with a general matrix A, an ergodic process may exhibit complex conjugate eigenvalues inside the unit disk. However, the eigenvalue with the largest modulus less than 1 is always real (it is the Perron-Frobenius eigenvalue of Neuts' matrix R, [13]). It is observed that if the parameters of the model are varied in such a way that the inequality (17) is violated, passing through an equality, then that eigenvalue leaves the unit disk, passing through point 1.

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

(1) Compute the eigenvalues, λ_k, and the corresponding left eigenvectors, ψ_k, of $Q(\lambda)$. If $c < N+1$, stop; a steady-state distribution does not exist.

(2) Solve the finite set of linear equations (7) and (9), with $\mathbf{v}_M$ and $\mathbf{v}_{M+1}$ given by (15), to determine the constants x_k and the vectors $\mathbf{v}_j$ for $j < M$.

(3) Use the obtained solution for the purpose of determining 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 $det[Q(\lambda)]$, then finding its roots, may be rather time-consuming for large N and is therefore not recommended. An alternative which is preferable in most cases is to reduce the quadratic eigenvalue-eigenvector problem

$$\psi[Q_0 + Q_1\lambda + Q_2\lambda^2] = 0 , \tag{18}$$

to a linear one of the form $\mathbf{y}Q = \lambda\mathbf{y}$, 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 when at least one of the matrices B and C is non-singular. Indeed, suppose that $C^{-1} = Q_2^{-1}$ exists. After multiplying (18) on the right by C^{-1}, it becomes

$$\psi[R_0 + R_1\lambda + I\lambda^2] = 0 , \tag{19}$$

where $R_0 = Q_0C^{-1}$ and $R_1 = Q_1C^{-1}$. By introducing the vector $\varphi = \lambda\psi$, equation (19) can be rewritten in the equivalent linear form

$$[\psi \quad \varphi]\begin{bmatrix} 0 & -R_0 \\ I & -R_1 \end{bmatrix} = \lambda[\psi \quad \varphi] . \tag{20}$$

If C is singular but B is not, a similar linearisation is achieved by a multiplication of (18) on the right by B^{-1}, and a change of variable $\lambda \to 1/\lambda$. The relevant eigenvalues are then those outside the unit disk.

If both B and C are singular, then the desired result is achieved by an alternative change of variable: $\lambda \to (\theta+\lambda)/(\theta-\lambda)$ (see [7]). The parameter θ is chosen so that the matrix $S = \theta^2Q_2 + \theta Q_1 + Q_0$ is non-singular.

Step 2 may be computationally expensive if the threshold M is large. In the present case, because of the block tridiagonal structure of the set of equations, they can be solved efficiently by a block iterative method. Also, if the matrices B_j ($j = 0, 1, \ldots, M-1$) are non-singular (which is usually the case in pactice), 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 (7) for $j = M, M-1, \ldots, 1$. One is then left with equations (7) for $j = 0$, plus (9) (a total of $N+1$ independent linear equations), for the $N+1$ unknowns x_k. This is in fact the method adopted in all our implementations.

Before leaving this section, it is worth pointing out that the spectral expansion method can be applied also to processes where the variable J (i.e. the queue size), is bounded [8, 15]. Suppose that there is a threshold, $T > M$, such that $A_j = C_j = 0$ for $j > T$, and $B_j = 0$ for $j \geq T$. The vector difference equation (10) holds for $j = M, M+1, \ldots, T-2$, and a modified, two-term equation holds for $j = T-1$. This means that there are $N+1$ additional detailed balance equations to be satisfied. On the other hand, it is no longer necessary to restrict the expansion to eigenvalues inside the unit disk, since the state space is finite. One can therefore use the general solution (14), which has d terms, for ($j = M, M+1, \ldots, T$). If $d = 2N+2$, i.e. if the matrix Q_2 in (11) does not have zero rows, then the number of extra coefficients to be determined exactly matches the number of extra equations. If Q_2 has zero rows but Q_0 does not (which is the case in the present model) then the 'dual' characteristic polynomial,

$$\hat{Q}(\lambda) = Q_2 + Q_1\lambda + Q_0\lambda^2 \ ,$$

has a determinant of degree $d = 2N+2$ and one can use the expansion

$$\mathbf{v}_j = \sum_{k=1}^{d} x_k \boldsymbol{\psi}_k \lambda_k^{T-j} \ \ ; \ \ j = M, M+1, \ldots T \ .$$

If both Q_0 and Q_2 have zero rows, the method still applies, but a more complicated development is involved.

4 Generalisations

The model in section 2 can be generalised in different directions, without affecting the applicability of the spectral expansion solution. First, one may wish to model batch arrivals and/or departures. This leads to a Markov process $X = \{[I(t), J(t)]\,;\, t \geq 0\}$, on the state space $\{0, 1, \ldots, N\} \times \{0, 1, \ldots\}$, where the variable J may jump by arbitrary, but bounded amounts in either direction. In other words, we wish to allow transitions

(a) From state (i, j) to state (k, j) $(0 \leq i, k \leq N \ ; \ i \neq k)$;

(b) From state (i, j) to state $(k, j+s)$ $(0 \leq i, k \leq N \ ; \ 1 \leq s \leq r_1 \ ; \ r_1 \geq 1)$;

(c) From state (i, j) to state $(k, j-s)$ $(0 \leq i, k \leq N \ ; \ 1 \leq s \leq r_2 \ ; \ r_2 \geq 1)$.

Denote by A_j, $B_j(s)$ and $C_j(s)$ the transition rate matrices associated with (a), (b) and (c), respectively. There is a threshold M, such that

$$A_j = A \ \ ; \ \ B_j(s) = B(s) \ \ ; \ \ C_j(s) = C(s) \ , \ j \geq M \ . \tag{21}$$

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, we can write the balance equations for $j > M + r_1$ in a form analogous to (8):

$$\mathbf{v}_j[D^A + \sum_{s=1}^{r_1} D^{B(s)} + \sum_{s=1}^{r_2} D^{C(s)}] = \sum_{s=1}^{r_1} \mathbf{v}_{j-s}B(s) + \mathbf{v}_j A + \sum_{s=1}^{r_2} \mathbf{v}_{j+s}C(s) . \qquad (22)$$

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, we note that (22) can be rewritten as a vector difference equation, this time of order $r = r_1 + r_2$, with constant coefficients:

$$\sum_{\ell=0}^{r} \mathbf{v}_{j+\ell} Q_\ell = 0 \quad . \qquad (23)$$

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 corresponding matrix polynomial is

$$Q(\lambda) = \sum_{\ell=0}^{r} Q_\ell \lambda^\ell . \qquad (24)$$

¿From here on, the development is similar to the one in section 3. The normalizeable solution of equation (23) is of the form

$$\mathbf{v}_j = \sum_{k=1}^{c} x_k \boldsymbol{\psi}_k \lambda_k^j \;\; ; \;\; j = M, M+1, \ldots , \qquad (25)$$

where λ_k are the eigenvalues of $Q(\lambda)$ in the interior of the unit disk, $\boldsymbol{\psi}_k$ are the corresponding independent left eigenvectors, and x_k are arbitrary constants ($k = 1, 2, \ldots, c$). The latter constants are determined with the aid of the state-dependent balance equations and the normalising equation. Propositions 1 and 2 continue to hold. The degree of $Q(\lambda)$ is now higher, but the extra eigenvalues are on, or outside the unit disk. The number of those inside is still $N + 1$.

As before, the solution is applicable to processes where the random variable J is bounded. In that case, the spectral expansion must include all the eigenvalues and eigenvectors of $Q(\lambda)$. There are more unknown coefficients to determine, but also more boundary equations.

For computational purposes, the polynomial eigenvalue-eigenvector problem of degree r can be transformed into a linear one in much the same way as described in section 3. For example, suppose that Q_r is non-singular and multiply (23) on the right by Q_r^{-1}. This leads to the problem

$$\psi\,[\sum_{\ell=0}^{r-1} R_\ell \lambda^\ell + I\lambda^r] = 0 , \qquad (26)$$

where $R_\ell = Q_\ell Q_r^{-1}$. Introducing the vectors $\varphi_\ell = \lambda^\ell \psi,\ \ell = 1, 2, \ldots, r-1$, we obtain the equivalent linear form

$$[\psi\ \varphi_1\ \ldots\ \varphi_{r-1}] \begin{bmatrix} 0 & & & -R_0 \\ I & 0 & & -R_1 \\ & \ddots & \ddots & \\ & & I & -R_{r-1} \end{bmatrix} = \lambda[\psi\ \varphi_1\ \ldots\ \varphi_{r-1}] .$$

Other transformations to linear form are available when Q_r is singular and Q_0 is not, or when both Q_0 and Q_r are singular.

Another direction in which the model can be generalised is by allowing more complex operative states for the set of processors. For instance, the processors need not be identical, but may have different breakdown, repair and service characteristics. Then an operative state would be defined by specifying the subset of processors that are operative; the variable $I(t)$ would take 2^N possible values, instead of $N+1$. Also, non-exponential (Phase-type or Coxian) distributions of various random variables may be handled by the method of stages, again at the price of increasing the number of operative states. Note that as the number of operative states increases, so does the number of eigenvalues, eigenvectors and unknown constants that have to be determined.

5 Comparison with the matrix-geometric solution

The models described in this paper can be solved also by the matrix-geometric method [13]. For the Quasi-Birth-and-Death processes considered in section 3, the vectors $\mathbf{v}_j$, $j \geq M$, are expressed in the form

$$\mathbf{v}_j = \mathbf{v}_M R^{j-M} \quad , \tag{27}$$

where R is the minimal non-negative solution of the quadratic matrix equation corresponding to the difference equation (10):

$$Q_0 + RQ_1 + R^2 Q_2 = 0 \quad . \tag{28}$$

The remaining unknown probabilities are determined from the balance equations for $j \leq M$, together with the normalizing equation. This part of the solution is very similar to step 2 of the spectral expansion procedure (summary in section 3), and has the same order of complexity.

There is a close relationship between the matrix equation (28) and the eigenvalue-eigenvector equation (12). If a matrix R satisfies (28), and $\boldsymbol{\psi} R = \lambda \boldsymbol{\psi}$ for some $\boldsymbol{\psi}$ and λ, then those $\boldsymbol{\psi}$ and λ satisfy (12). Conversely, any $N+1$ numbers λ_k and independent vectors $\boldsymbol{\psi}_k$ that satisfy (12), determine a matrix R which has them as eigenvalues and eigenvectors, and satisfies (28). For an ergodic system, the minimal solution of (28) is the one whose eigenvalues are the $N+1$ solutions of (12) that are inside the unit disk.

In the more general case of section 4, the matrix-geometric solution is still of the form (27), but the matrix R is obtained as the minimal non-negative solution of a matrix equation of a higher order.

Thus the main operational difference between the two approaches is that spectral expansion involves the computation of eigenvalues and eigenvectors of a given matrix, whereas the matrix-geometric method requires the determination of an unknown matrix by solving a matrix equation. To evaluate the impact of that difference, we have performed a series of experiments where the same models were solved by both methods. The performance measure that is computed is $E(J)$, the average number of jobs in the system. The aim of these experiments is to examine the issues of efficiency, accuracy, numerical stability and limits of application.

There are several iterative procedures for solving equation (28). We have used the so-called *modified SS method* [6, 13]:

$$R_{n+1} = [-Q_0 - R_n^2 Q_2]Q_1^{-1} \ , \tag{29}$$

starting with $R_0 = 0$. The number of iterations depends on the model parameters and on the desired accuracy.

The tasks that are similar in both methods are, whenever possible, carried out by the same piece of software. For instance, the solution of a set of simultaneous linear equations is performed by routine F04ATF from the NAG numerical algorithms package. That program produces diagnostic messages if the input matrix is singular or ill-conditioned. Other NAG routines are used for matrix inversion and (in the case of spectral expansion), to find eigenvalues and eigenvectors. No attempt is made, in either method, to optimize the code or to exploit particular model features. Double precision arithmetic is used throughout (except inside the NAG routines, which sometimes employ extended precision).

It should be emphasized that both solution methods are exact in principle. That is, if the eigenvalues and eigenvectors on the one hand, and the matrix R on the other, are determined exactly, and if the respective sets of linear equations are solved exactly, then in both cases the values of the resulting probabilities and performance measures would be correct. In practice, the accuracy of the results depends on many factors and varies between the methods.

The routine that solves the linear eigenvalue-eigenvector problem proceeds as follows: (i) the matrix is transformed into upper Hessenberg form; (ii) the QR algorithm is applied to bring it into real Schur form, with 1×1 and/or 2×2 blocks on the main diagonal; (iii) the real eigenvalues (and corresponding eigenvectors) are obtained from the 1×1 blocks, and the pairs of complex-conjugate ones from the 2×2 blocks. Step (ii) involves iterations, but their number is small (an average of 2 per diagonal element). In fact, if the QR algorithm, which works to machine precision, does not complete after a total of $30(N+1)$ iterations, the routine quits, producing a diagnostic message. This has never happened in our experiments.

Thus, the overall computational complexity of determining all eigenvalues and eigenvectors is $O(N^3)$. Moreover, that complexity is more or less independent of the parameters of the model. Even when two eigenvalues coincide, or almost coincide, the algorithm finds linearly independent eigenvectors if the latter exist. The accuracy of the spectral expansion solution is limited only by the precision of the numerical operations. When there is no diagnostic about an ill-conditioned set of linear equations, the value of $E(J)$ computed by this method is observed to be correct to about 10 signifficant digits, i.e. it is exact for all practical purposes.

The matrix-geometric solution relies on the matrix R, which is computed approximately by the iterative procedure described above. The criterion for terminating that procedure would normally include some trade-off between computation time and accuracy.

The situation is illustrated in table 1, where the 10-server model from the example at the end of section 3 is solved by both methods. The iterative procedure (29) is terminated when

$$\max_{i,j} |R_{n+1}(i,j) - R_n(i,j)| < \epsilon ,$$

for a given value of ϵ. The performance measure that is computed is the average queue size, $E(J)$, for three different values of the arrival rate, σ. The spectral expansion solution is of course independent of ϵ. Column 4 gives the percentage relative difference $100 * [E(J)_{(spect.exp.)} - E(J)_{(mat.geom.)}]/E(J)_{(spect.exp.)}$

Table 1: Trade-off between accuracy and complexity

$\sigma = 3.0$; $E(J)_{(spect.exp.)} = 5.199710420277$			
ϵ	iterations	$E(J)_{(mat.geom.)}$	% difference
10^{-3}	29	5.175072879090	0.4738252555
10^{-6}	93	5.199686527032	0.0004595110
10^{-9}	158	5.199710396767	0.0000004521
10^{-12}	223	5.199710420254	0.0000000004
$\sigma = 6.0$; $E(J)_{(spect.exp.)} = 50.4058205572$			
ϵ	iterations	$E(J)_{(mat.geom.)}$	% difference
10^{-3}	77	35.0382555318	30.4876795091
10^{-6}	670	50.3726795846	0.0657483051
10^{-9}	1334	50.4057872349	0.0000661080
10^{-12}	1999	50.4058205241	0.0000000657
$\sigma = 6.6$; $E(J)_{(spect.exp.)} = 540.46702456$			
ϵ	iterations	$E(J)_{(mat.geom.)}$	% difference
10^{-3}	77	58.19477818	89.23250160
10^{-6}	2636	506.34712584	6.31303986
10^{-9}	9174	540.42821366	0.00718099
10^{-12}	15836	540.46698572	0.00000719

The table confirms that when ϵ decreases, the matrix-geometric solution approaches the spectral expansion one. However, it is important to observe that the accuracy of R is not related in any obvious way to the accuracy of $E(J)$. Thus, taking $\epsilon = 10^{-6}$ yields an answer whose relative error is 0.0004% when $\sigma = 3$, 0.06% when $\sigma = 6$, and 6.3% when $\sigma = 6.6$. Another important aspect of the table is that, for a given ϵ, the number of iterations required to computate R increases with σ.

Extensive comparisons of the run times of the spectral expansion and the matrix-geometric solutions have shown that the latter increases without bound when the offered load increases. For a heavily loaded system, the matrix-geometric run time may be many orders of magnitude larger than the spectral expansion one. The extra time is almost entirely taken up by the procedure that calculates the matrix R. It seems clear that, even when the accuracy requirements are modest, the computational demands of the iterative procedure have a vertical asymptote at the saturation point.

It can be argued that a different procedure for computing R may yield better results. A number of such procedures have been compared in [6], and improvements have indeed been observed. However, the results reported in [6] indicate that all iterative approaches suffer from the drawbacks exhibited in table 1 and figure 2, namely that their accuracy and efficiency decrease when the load increases. Perhaps the fastest existing method for computing R is not by iterations, but via its eigenvalues and eigenvectors (this was done in [2] for a model with real eigenvalues and a non-singular matrix C). If that method is adopted, then one might as well use spectral expansion directly, and avoid the matrix R altogether.

In our exparience, using different starting points for the iterations (rather than $R_0 = 0$), does not bring signifficant benefits and may be dangerous. The procedure could diverge or it could, conceivably, converge to a non-minimal solution of (28).

The second step in both algorithms (solving a set of linear equations), is where emerging numerical problems begin to manifest themselves. The observable symptom of such problems is usually an ill-conditioned matrix. It should be emphasized that neither the spectral expansion nor the matrix-geometric algorithm is immune to ill-conditioned matrices. We have observed them generally when the number of processors is large and the system is lightly loaded. For example, taking $N = 50$, $\sigma < 15$ (other parameters as in the figures), leads to ill-conditioned matrices in both algorithms.

In many cases, the problem of ill-conditioning can be alleviated by appropriate scaling (e.g. multiplying certain vectors by suitable constants). We have tried a few scaling techniques in connection with the spectral expansion algorithm, and have had some success. However, it is a reasonable guess that, within the constraint of double precision arithmetic, neither method can handle accurately models with more than about 100 processors. Of course, that range can be extended considerably by using extended precision, but then the choice of programming language and/or compiler would be restricted.

6 Conclusion

We have considered a class of Markov-modulated models, exemplified by the multiprocessor queue subject to general breakdowns and repairs. In that context, spectral expansion is an efficient and readily implementable solution method. It enables various queue length statistics to be computed easily and accurately, over a wide range of parameter values.

When comparing the spectral expansion and the matrix-geometric solutions, a strong case can be made for the former. Spectral expansion offers considerable advantages in efficiency, without undue penalties in terms of numerical stability. The speed gains are especially pronounced, and indeed appear to be unbounded, when the system approaches saturation. However, further comparisons, perhaps with different methods for computing R, or with other methods such as aggregation-disaggregation, would be desirable. It would also be interesting to compare the different methods on problems where the jumps of J are greater than 1.

It should be pointed out that the matrix-geometric approach can be applied to models with block-M/G/1 or block-G/M/1 structure, which are not of the type discussed here. We believe that it would be possible to extend the spectral expansion method to those models too (see [4]).

Another aspect of the spectral expansion method which deserves at least a mention, is that it provides a good basis for obtaining asymptotic and approximate results. For example, the tail of the steady-state distribution (for large j), is dominated by the term in the expansion containing the eigenvalue with the largest modulus less than 1 [18, 12].

Acknowledgements

We wish to thank P.-J. Courtois and P. Semal for some pertinent remarks concerning Proposition 1, and for other helpful comments. Thanks are also due to the anonymous referees, whose remarks resulted in many improvements. This work was carried out in connection with the ESPRIT Basic Research projects PDCS and QMIPS, supported by the European Community.

References

[1] R. Chakka and I. Mitrani, *A Numerical Solution Method for Multiprocessor Systems with General Breakdowns and Repairs*, Procs., 6th Int. Conf. on Tools and Techniques, Edinburg, 1992.

[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] H.R. Gail, S.L. Hantler and B.A. Taylor, *Spectral Analysis of M/G/1 Type Markov Chains*, RC17765, IBM Research Division, 1992.

[5] I. Gohberg, P. Lancaster and L. Rodman, *Matrix Polynomials*, Academic Press, 1982.

[6] L. Gun, *Experimental Results on Matrix-Analytical Solution Techniques - Extensions and Comparisons*, Stochastic Models, 5, 4, pp. 669-682, 1989.

[7] A. Jennings, *Matrix Computations for Engineers and Scientists*, Wiley, 1977.

[8] J. Keilson, U. Sumita and M. Zachmann, *Row-Continuous Finite Markov Chains: Structure and Algorithms*, Journal of the Operations Research Society of Japan, 30, 3, 1987.

[9] 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.

[10] I. Mitrani and B. Avi-Itzhak, *A Many-Server Queue with Service Interruptions*, Operations Research, 16, 3, pp.628-638, 1968.

[11] 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.

[12] I. Mitrani and A. Puhalskii, *Limiting Results for Multiprocessor Systems with Breakdowns and Repairs*, to appear, QUESTA.

[13] M.F. Neuts, *Matrix Geometric Solutions in Stochastic Models*, John Hopkins Press, 1981.

[14] 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.

[15] L.P. Seelen *An Algorithm for Ph/Ph/c Queues*, European J. of Oper. Res., 23, pp. 118-127, 1986.

[16] B. Sengupta, *A Queue with Service Interruptions in an Alternating Markovian Environment*, Operations Research, 38, pp.308-318, 1990.

[17] N.U. Prabhu and Y. Zhu, *Markov-Modulated Queueing Systems*, QUESTA, 5, pp. 215-246, 1989.

[18] H.C. Tijms and M.C.T. van der Coevering, *A Simple Numerical Approach for Infinite-State Markov Chains*, Probability in Eng. and Inf. Sci., 5, pp. 285-295, 1991.