A Manufacturing Production Line with Service Interruptions

Nigel Thomas[1] and Isi Mitrani[2]

[1] Department of Computer Science, University of Edinburgh
nat@dcs.ed.ac.uk
[2] Department of Computing Science, University of Newcastle upon Tyne
isi.mitrani@ncl.ac.uk

Abstract: Jobs from a single Poisson input stream receive K independent stages of service, one at each stage in the pipeline. At stage i jobs are routed through one of the N_i available nodes, modelled as $M/M/1$ queues. These nodes are subject to random failure and repairs which leave their corresponding queues intact, but may affect the routeing of jobs arriving at that stage during the subsequent repair period. Two possible approximate solutions for the marginal queue size distributions are obtained using Marie's method and spectral expansion. Approximations are compared with solutions obtained by simulation techniques. Two routeing strategies are considered, fixed and selective, and the relative accuracy of the approximate solutions and predicted optimal routeing vectors are discussed. This method is obviously applicable to other, more general, network models and it is therefore interesting to observe the accuracy of the approximations and predictions of an optimal routeing vector. Models such as this have traditionally been studied through simulation. However, an exceedingly long runtime is needed to obtain steady state results, especially when failures are rare and repairs are slow. The method presented here gives a very rapid response and as such is clearly of great practical benefit, especially when optimising the routing of jobs.

Keywords: queueing networks - breakdowns - spectral expansion

1 Introduction

The analysis of the performance of queueing systems which are subject to breakdowns has a long and interesting history. However very little work has been done involving more than one queue, notable exceptions being Mitrani [5], Mikou [3] and Idrissi-Kacemi et al [4]. Mitrani and Wright [8] analysed a system of nodes in parallel which suffered failures that caused all jobs to be lost, incoming jobs were then routed away from failed nodes, this resulted in an interesting trade off in performance between response time and job loss. Thomas and Mitrani [10] started with this basic model, but changed the nature of the failure so that queues were preserved during repair periods. Furthermore it was assumed that jobs could continue to join a queue even when the corresponding node was broken and that no jobs were lost. This gave rise to a number of possible routeing strategies which were contrasted and compared.

The structure of our model is such that jobs arrive at the start of the pipeline in a single Poisson stream and progress along the pipeline such that jobs departing from one stage constitute the input stream at the next stage. On arrival at stage i, jobs are directed to one of N_i alternative nodes, each of which has an associated independent unbounded queue. The service, repair and failure processes are independent of each other and, in general, possess differing parameters. When a node breaks down the queue is unaffected and may even continue to accept new jobs, although it may be desirable to redirect new jobs elsewhere, depending on the routeing strategy in operation. The choice of where to send a job on arrival at a stage is strictly Bernoulli, based purely on the operational states of the nodes at this stage and on the routeing strategy in operation, but independent of any past history and of the number of jobs present in each queue. No jobs are lost. The model and its parameters are specified in Section 2.

In order to determine the marginal queue size distribution for a given node at a given stage, it is necessary to consider the operational state of all the nodes in this and any preceding stages in the pipeline. Clearly this gives rise to a very complicated model for any non-trivial pipeline and so it is necessary to consider simplifications to the model in order to make it a more manageable size. Two such approximations are given in Section 4: one which treats each stage in isolation from its predecessors, and another which also takes into account the states of those nodes at the immediately preceding stage. Determining the exact solution of the relevant performance measures is achieved by simulation, thus providing the means to compare the accuracy of the two approximations (Section 6).

The solution method presented here is an example of Marie's method [2] combined with the matrix solution technique known as spectral expansion [6, 7]. However our method is more than simply a combination of these earlier results. Usually Marie's method is used to simplify a model to give a product form result. In this case the result of applying the approximation does not generally lead to product form (except in the trivial case where there is no rerouteing of jobs). Instead the approximation gives rise to expressions for the approximated marginal queue size distributions. These marginal probabilities are not, in general, independent and therefore do not give rise to expressions for the joint probability distribution. However, it is easily seen that average number of jobs in the system is the sum of the average number of jobs in each queue, hence many performance measures of interest can be derived. It is clear that this method is being applied in a situation where previously few, if any, interesting performance measures have been derived and so is obviously a significant contribution to the literature.

The motivation for studying this system comes from the field of manufacturing. The pipeline of parallel service centres relates to a production line. At each stage of the pipeline therefore a different process is undertaken. Each stage of production has to be completed before the next part of the manufacture can proceed. These processes are performed in a strict order and the result of the completion of all the processes is a finished manufactured product. In this interpretation the different nodes at a stage will represent different workshops offering the same process. Workshops suffer periodic mechanical disruption (breakdowns) and are unavailable for varying lengths of time. It is assumed that jobs already sent to a workshop will not be directed elsewhere if a failure occurs, but new arrivals may be rerouted. This naturally raises the question of how to set the routeing probabilities. Therefore an additional purpose of this study, as well as determining the accuracy of the approximations, is to determine how the optimal routeing probabilities differ between the approximations and the simulation for given parameter sets.

2 The Model

Jobs arrive into the system in a Poisson stream with rate λ. There are K stages in series and in stage i there are N_i nodes in parallel, each with an associated unbounded queue, to which incoming jobs may be directed. Server j at stage i goes through alternating independent operative and inoperative periods, distributed exponentially with means $1/\xi_{i,j}$ and $1/\eta_{i,j}$ respectively. While it is operative, the jobs in its queue receive service of an exponentially distributed duration with mean $1/\mu_{i,j}$, and leave the stage upon completion to proceed to the next (if any) stage of service. When a node becomes inoperative (breaks down), the corresponding queue, including the job in service (if any), remains in place. Services that are interrupted in this way are eventually resumed from the point of interruption. The system model is illustrated in Figure 1.

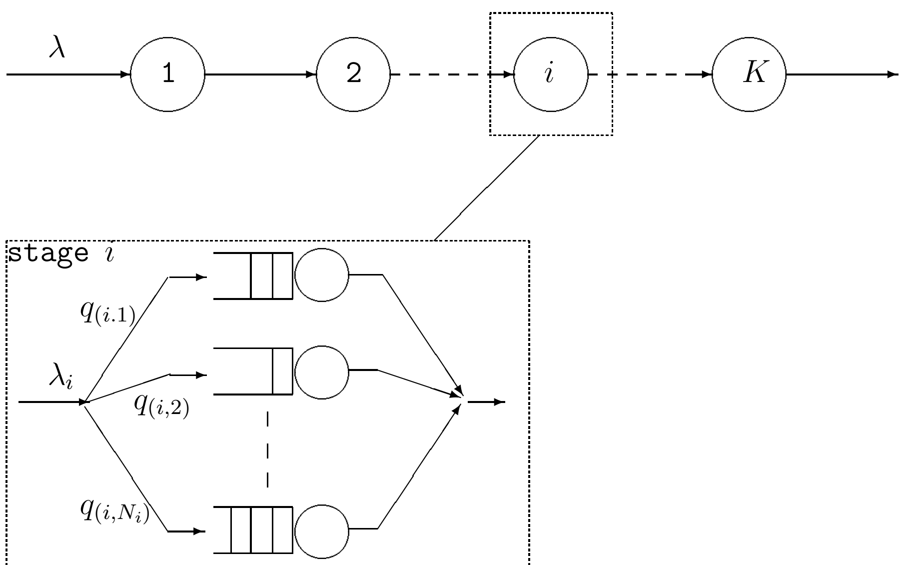

Figure 1: A single source to a pipeline of K stages, split between the nodes in each stage

The arrival rate at stage i is given in Figure 1 as λ_i, but since no jobs are lost the overall arrival rate at all stages will be the same as the external Poisson arrival rate λ. However, since the arrivals at stage i depend on the departures from stage $i-1$ then the arrival stream will, in general, cease to be Poisson. The *system configuration* at any moment is specified by the subset, σ, of nodes that are currently operative (that subset may be empty, or it may be the set of all nodes): $\sigma \subset \Omega_N$, where $\Omega_N = \{(1,1),(1,2),\ldots,(1,N_1),(2,1),\ldots,(K,N_K)\}$, where the pair $\{i,j\}$ represents node j at stage i. There are of course 2^N possible system configurations, where $N = \sum_{i=1}^{K} N_i$. In general it is more convenient to consider the subset σ_i whose elements are those nodes at stage i which are operative. The set of all nodes at stage i is denoted by Ω_{N_i}. Clearly $\sigma_i \subset \Omega_{N_i} \subset \Omega_N$ and $\sigma_i \subset \sigma$. The steady-state marginal probability, p_{σ_i}, of configuration σ_i at stage i is given by

$$p_{\sigma_i} = \prod_{j\in\sigma_i} \frac{\eta_{i,j}}{\xi_{i,j}+\eta_{i,j}} \prod_{j\in\overline{\sigma}_i} \frac{\xi_{i,j}}{\xi_{i,j}+\eta_{i,j}} \quad , \quad \sigma_i \subset \Omega_{N_i} \ , \tag{1}$$

And the steady-state marginal probability, p_σ, of configuration σ is given by

$$p_{\sigma} = \prod_{i,j\in\sigma} \frac{\eta_{i,j}}{\xi_{i,j}+\eta_{i,j}} \prod_{i,j\in\overline{\sigma}} \frac{\xi_{i,j}}{\xi_{i,j}+\eta_{i,j}} \quad , \quad \sigma \subset \Omega_{N} \ , \tag{2}$$

where $\overline{\sigma}_i$ is the complement of σ_i with respect to Ω_{N_i}, $\overline{\sigma}$ is the complement of σ with respect to Ω_N and an empty product is by definition equal to 1. These expressions follow from the fact that nodes break down and are repaired independently of each other.

If, at the time of arrival at stage i, a new job finds the stage in configuration σ_i, then it is directed to node j with probability $q_{i,j}(\sigma_i)$. These decisions are independent of each other, of past history, of the sizes of the various queues and of the state of any other stage in the pipeline. Thus, a routeing policy at stage i is defined by specifying 2^{N_i} vectors,

$$\mathbf{q}_i(\sigma_i) = [q_{i,1}(\sigma_i), q_{i,2}(\sigma_i), \ldots, q_{i,N_i}(\sigma_i)] \quad , \quad \sigma_i \subset \Omega_{N_i} \quad , \tag{3}$$

such that for every σ_i,

$$\sum_{j=1}^{N_i} q_{i,j}(\sigma_i) = 1 \quad .$$

The system state at time t is specified by the pair $[I(t), \mathbf{J}(t)]$, where $I(t)$ indicates the current configuration (the configurations can be numbered, so that $I(t)$ is an integer in the range $0, 1, \ldots, 2^N - 1$), and $\mathbf{J}(t)$ is an integer vector whose k'th element, $J_k(t)$, is the number of jobs in queue k ($k = 1, 2, \ldots, N$). The integer k is used here instead of the pair i,j for simplicity, the relationship between k and i,j is a simple 1 to 1 mapping such that

$$j + \sum_{x=1}^{i-1} N_x = k$$

Under the assumptions that have been made, $X = \{[I(t), \mathbf{J}(t)], t \geq 0\}$ is an irreducible Markov process. The condition for ergodicity of X is that, for every queue i, j, the overall arrival rate is lower than the overall service capacity:

$$\sum_{\forall \sigma_i} \lambda_i p_{\sigma_i} q_{i,j}(\sigma_i) \ < \ \mu_{i,j} \frac{\eta_{i,j}}{\xi_{i,j} + \eta_{i,j}} \ , \ i = 1, 2 \ldots, K, j = 1, 2, \ldots, N_i \ . \tag{4}$$

When the routeing probabilities at each stage depend on the system configuration, the process X is not separable (i.e., it does not have a product-form solution). Consequently, the problem of determining its equilibrium distribution is intractable in general. On the other hand, the quantities of principal interest are expressed in terms of averages only; they are the steady-state mean queue sizes, L_k, and the overall average response time, W, given by

$$W = \frac{1}{\lambda} \sum_{i=1}^{K} \sum_{j=1}^{N_i} L_{i,j} \ . \tag{5}$$

To determine those performance measures, it is not necessary to know the joint distribution of all queue sizes; the marginal distributions of the N queues in isolation are sufficient. Unfortunately, the isolated queue processes, $\{J_k(t) \ , \ t \geq 0\}$ $(k = 1, 2, \ldots, N)$, are not Markov. As mentioned earlier the arrival stream at stage i $(i \geq 2)$ is not Poisson since it depends on the activity of all the previous stages, this makes an exact solution of the marginal queue size distributions almost as intractable a problem as solving the joint distribution of all queue sizes. However, it is possible to obtain good approximate solutions for the marginal queue size distributions by assuming the arrival stream at stage i to be Markov-Modulated Poisson. Some discussion as to how best to form the approximated arrival streams is presented later.

Consider the stochastic processes $Y_{i,j}$,

$$Y_{i,j} = \{[I^*(t), J_{i,j}(t)] \ , \ t \geq 0\} \ , \ i = 1, 2, \ldots, K \ , \ j = 1, 2, \ldots, N_i$$

which model the joint behaviour of the configuration and the size of an individual queue i, j, where $I^*(t)$ indicates the current approximated system configuration. In general each possible approximated system configuration, $I^*(t)$, will represent a set of one or more of the exact system configurations, σ. The number of approximated system configurations considered, from now on referred to as I_{max}, will, in general, determine the accuracy of the solution and the amount of computation required. The value of I_{max} will therefore be limited at the upper bound by the amount of computational power available and the desired rapidity of the solution and at the lower bound by the desired accuracy of the solution.

The state space of $Y_{i,j}$ is infinite in one dimension only, which simplifies the solution considerably and makes it tractable for reasonably large values of I_{max}. The important observation here is that, with the assumption of a Markov-Modulated arrival process, $Y_{i,j}$ is an irreducible Markov process, for every i, j. This is because the arrivals into, and departures from queue i, j during a small interval $(t \, , t + \Delta t)$ depend only on the approximated system configuration and the size of queue i, j at time t, and not on the sizes of the other queues. As mentioned earlier, without the approximation of the arrival stream to a Markov-Modulated arrival process, this statement would not be true, since a job only arrives at stage $i+1$ after successfully completing service at stage i, therefore making the queue size at any stage dependent on all previous stages of service.

The next task, therefore, is to find the equilibrium distribution of $Y_{i,j}$:

$$p_{i,j}(x,y) = \lim_{t\to\infty} P[I^*(t) = x \, , \, J_{i,j}(t) = y] \quad , \quad x = 0, 1, \ldots, I_{max} - 1 \quad , \quad y = 0, 1, \ldots \tag{6}$$

Given the probabilities $p_{i,j}(x, y)$, the average size of queue i, j is obtained from

$$L_{i,j} = \sum_{y=1}^{\infty} y \sum_{x=0}^{I_{max}-1} p_{i,j}(x,y) \tag{7}$$

3 Marginal queue size distributions

The process $Y_{i,j}$ is of the *block tri-diagonal*, or *Quasi-Birth-and-Death* type. Its possible transitions are:

(a) from state (I^*, J) to state (I', J), where I' is an approximated configuration with either one more, or one fewer operative node in the relevant stage(s);

(b) from state (I^*, J) to state $(I^*, J+1)$, if the average routeing probability to queue i, j in approximated configuration I^*, $\overline{q}_{i,j}(I^*)$, is non-zero;

(c) from state (I^*, J) to state $(I^*, J-1)$, if $J > 0$ and node i, j is operative in approximated configuration I^*.

Thus, more simply, (a) represents failures and repairs, (b) represents arrivals and (c) represents services. The balance equations for $Y_{i,j}$ are best written in vector and matrix form. Define the (row) vector of equilibrium probabilities of all states with J jobs in queue i, j :

$$\mathbf{v}_{i,j}(J) = [p_{i,j}(0, J), p_{i,j}(1, J), \ldots, p_{i,j}(I_{max} - 1, J)] \quad , J = 0, 1, \ldots \quad . \tag{8}$$

It is assumed that approximated system configurations will be chosen such that node i, j will be either operative or inoperative in any approximated configuration, but not both. Let $A = (a_{I^*,I'})$ $(I^*, I' = 0, 1, \ldots, I_{max} - 1)$ be the matrix of instantaneous transition rates corresponding to transitions (a). If in approximated configuration I^* the subset of operative nodes can be σ, and in I' it can be $\sigma + \{\ell\}$, for some node ℓ, then $a_{I^*,I'} = \sum_{\forall \ell} \eta_\ell$; similarly, if in I' the approximated configuration can be $\sigma - \{\ell\}$, for some node ℓ, then $a_{I^*,I'} = \sum_{\forall \ell} \xi_\ell$. It is also useful to introduce the diagonal matrix, D_A, whose I^* 'th diagonal element is the I^* 'th row sum of A $(I^* = 0, 1, \ldots, I_{max})$.

Let $B_{i,j}$ be the diagonal matrix whose I^* 'th diagonal element is equal to $\overline{\lambda}_{i,j}(I^*)$; these elements are the instantaneous transition rates corresponding to transitions (b), where

$\overline{\lambda}_{i,j}(I^*)$ is the average arrival rate at queue j in stage i when the approximated configuration is I^*. Also, let $C_{i,j}$ be the diagonal matrix whose I^*'th diagonal element is equal to $\mu_{i,j}$ if node i,j is operative in approximated configuration I^*, and 0 otherwise; these are the instantaneous transition rates corresponding to transitions (c).

When $J > 0$, the vectors (8) satisfy the balance equations

$$\mathbf{v}_{i,j}(J)(D_A + B_{i,j} + C_{i,j}) = \mathbf{v}_{i,j}(J)A + \mathbf{v}_{i,j}(J-1)B_{i,j} + \mathbf{v}_{i,j}(J+1)C_{i,j}$$

$$, \quad , \quad J = 1, 2, \ldots \tag{9}$$

For $J = 0$, the equation is slightly different:

$$\mathbf{v}_{i,j}(0)(D_A + B_{i,j}) = \mathbf{v}_{i,j}(0)A + \mathbf{v}_{i,j}(1)C_{i,j} \ . \tag{10}$$

In addition, all probabilities must sum up to 1:

$$\sum_{J=0}^{\infty} \mathbf{v}_{i,j}(J)\mathbf{e} = 1 \ , \tag{11}$$

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

The above equations can be solved by several methods. Evidence presented in [6] suggests that the best approach for models such as this is to use *spectral expansion* (see [7] and [1]), in the same way as [10].

Rewrite (9) in the form

$$\mathbf{v}_{i,j}(J)Q_{i,j,0} + \mathbf{v}_{i,j}(J+1)Q_{i,j,1} + \mathbf{v}_{i,j}(J+2)Q_{i,j,2} = \mathbf{0} \ ,$$

$$J = 0, 1, \ldots \ , \tag{12}$$

where $Q_{i,j,0} = B_{i,j}$, $Q_{i,j,1} = A - D_A - B_{i,j} - C_{i,j}$ and $Q_{i,j,2} = C_{i,j}$. This is a homogeneous vector difference equation of order 2, with constant coefficients. Associated with it is the characteristic matrix polynomial, $Q_{i,j}(z)$, defined as

$$Q_{i,j}(z) = Q_{i,j,0} + Q_{i,j,1}z + Q_{i,j,2}z^2 \ . \tag{13}$$

Denote by $z_{i,j,\ell}$ and $\boldsymbol{\psi_{i,j,\ell}}$ the *generalised eigenvalues and left eigenvectors* of $Q_{i,j}(z)$. These quantities satisfy

$$\boldsymbol{\psi_{i,j,\ell}} Q_{i,j}(z_{i,j,\ell}) = \mathbf{0} \ , \quad \ell = 1, 2, \ldots, d \ , \tag{14}$$

where $d = degree\{det[Q_{i,j}(z)]\}$.

The eigenvalues do not have to be simple, but it is assumed that if $z_{i,j,\ell}$ has multiplicity r, then it has r linearly independent left eigenvectors. This is invariably observed to be the case in practice. Under that assumption, any solution of (12) is of the form

$$\mathbf{v_{i,j}}(J) = \sum_{\ell=1}^{d} x_{i,j,\ell} \boldsymbol{\psi_{i,j,\ell}} z_{i,j,\ell}^J \ , \quad J = 0, 1, \ldots \ , \tag{15}$$

where $x_{i,j,\ell}$ ($\ell = 1, 2, \ldots, d$), are arbitrary (complex) constants.

Moreover, since only solutions which can be normalised are acceptable, if $|z_{i,j,\ell}| \geq 1$ for some ℓ, then the corresponding coefficient $x_{i,j,\ell}$ must be set to 0. Numbering the eigenvalues of $Q_{i,j}(z)$ in increasing order of modulus, the spectral expansion solution of equation (12) can be written as

$$\mathbf{v_{i,j}}(J) = \sum_{\ell=1}^{c} x_{i,j,\ell} \boldsymbol{\psi_{i,j,\ell}} z_{i,j,\ell}^{J} \quad , \quad J = 0, 1, \ldots \quad , \tag{16}$$

where c is the number of eigenvalues strictly inside the unit disk (each counted according to its multiplicity).

In the numerical experiments carried out with this model, the eigenvalues and eigenvectors of $Q_{i,j}(z)$ have always been observed to be simple, real and positive.

Substituting (16), for $J = 0$ and $J = 1$, into (10), yields a set of homogeneous linear equations for the unknown coefficients $x_{i,j,\ell}$. There are $I_{max} - 1$ independent equations in this set (rather than I_{max}) because the generator matrix of the Markov process is singular. A further, non-homogeneous equation is provided by (11), which now becomes

$$\sum_{\ell=1}^{I_{max}} \frac{x_{i,j,\ell} \boldsymbol{\psi_{i,j,\ell}} \mathbf{e}}{1 - z_{i,j,\ell}} = 1 \quad .$$

These equations can be solved uniquely for the coefficients $x_{i,j,\ell}$, if $c = I_{max}$. This turns out to be the case when (4) is satisfied. Indeed, the ergodicity condition is equivalent to the requirement that $Q_{i,j}(z)$ has exactly I_{max} eigenvalues strictly inside the unit disk.

Having determined the coefficients $x_{i,j,\ell}$, the average number of jobs in queue i, j is obtained by substituting (16) into (7):

$$L_{i,j} = \sum_{\ell=1}^{I_{max}} \frac{x_{i,j,\ell} z_{i,j,\ell} \boldsymbol{\psi_{i,j,\ell}} \mathbf{e}}{(1 - z_{i,j,\ell})^2} \quad . \tag{17}$$

4 Approximated system configurations

In this model there are 2^N possible system configurations, which is clearly too large a number to solve for in any practical situation, hence the need for a reduced solution. In general, the arrivals at node i are dependent on all the preceding stages of service (or node configurations). However it is obvious that the nature of the arrivals at each node are most strongly linked to the configuration at the immediately preceding node. Thus one possible reduced solution method is clear, namely,

1. perform the solution described in Section 3 on the first node - this will be an exact solution since there are no preceding nodes to affect arrivals

2. extract from that solution the appropriate performance measures and the probabilities $p_{1,j}(\sigma_1), j = 0..N_1$, where $p_{i,j}(\sigma_i)$ is the probability that queue j (at node i) is non-empty given that the configuration of node i is σ_i.

3. perform the solution described above with the approximated system configurations merely the configuration of this node and that immediately preceding it, thus $I_{max} = 2^{N_i+N_{i-1}}$ and the arrival rate at node j (assumed Poisson) in configuration I^* is given by

$$q_j(\sigma_i) \sum_{k=0}^{N_i-1} (p_{i-1,k}(\sigma_{i-1})\mu_{i-1,k})$$

   where σ_{i-1} and σ_i represent the configurations at node $i-1$ and node i respectively at given approximated system configuration I^*.

4. extract from this solution the appropriate performance measures and the probabilities $p_{i,j}(\sigma_i), j = 0..N_i$

5. repeat steps 3 and 4 for the next node until all nodes have been solved.

Clearly this solution is only possible when N_i is relatively small for all i (if $N_i+N_{i-1} \geq 8$ then the matrices become very large) and so an alternative needs to be found. The simplest idea is to ignore all previous nodes in the solution of node i and take the arrival rate at that node to be Poisson rate λ, i.e. the same as the external arrival stream. This allows the solution of much larger parallel nodes, but at the expense of all consideration of the staged nature of service. A much better alternative would be for some halfway measure, allowing reasonably large systems to be solved with some knowledge of the preceding stage taken into account. In [9] and [10] some approximate methods for the solution of a single stage parallel system were presented, the best approximate solution being when the most significant arrival periods were treated independently and the remainder were amalgamated into logical groups. Applying the same technique here, one approach would be to have approximated system configurations based on the current server (i, j) either working or broken, with 0,1, or up to $N_i - 1$ other servers at stage i working, and 0,1, or up to N_{i-1} servers working at the previous stage, giving a total of $2N_i(N_{i-1} + 1)$ possible configurations. Another possibility is to consider all the possible configurations of stage i together with those arising from having 0,1 or up to N_{i-1} servers operative at stage $i-1$. These are just two examples, the best set of approximated system configurations will be determined by the server characteristics and the available computational resources. It is assumed that approximated system configurations will be chosen such that node i, j, the node whose queue is being evaluated, will be either operative or inoperative in any approximated configuration, but not both. This might appear in the first instance to be a restrictive assumption, however the effect of breakdowns on the performance of the system is the primary interest of this model. The behaviour of a node is significantly different when it is broken compared with when it is operative therefore this should be a major feature of any approximation.

The process $Y_{i,j}$ is of the *block tri-diagonal*, or *Quasi-Birth-and-Death* type: it can therefore be solved by spectral expansion to find the probabilities $p_{i,j}(x, y)$.

5 Scheduling strategies

As in [10] which considered a single stage parallel system, here strategies based on a single routeing vector, $\mathbf{q} = (q_1, q_2, \ldots, q_N)$, are evaluated and compared. In each case, the optimisation problem is to choose the elements of that vector so as to minimise the average response time.

1. *The fixed strategy.*

The most straightforward way of splitting the incoming stream at stage i is to send each job to queue j with probability q_j, regardless of the system configuration. In this simple case in the single stage model a simple equation could be used to determine the performance measures. However, with the introduction of several stages this is no longer true, as the arrival process at a given stage is affected by node failures at earlier stages.

2. *The selective strategy.*

Intuitively, it seems better not to send jobs to nodes where the server is inoperative, unless that is unavoidable. This suggests the following strategy: If the subset of operative nodes at stage i in the current system configuration is σ_i, and that subset is non-empty, send jobs to queue j only if $j \in \sigma_i$, with probability proportional to q_j :

$$q_j(\sigma_i) = \frac{q_j}{\sum_{\ell \in \sigma} q_\ell} \quad , \quad j \in \sigma \quad .$$

If σ is empty (i.e. all nodes are broken), send jobs to queue j with probability q_j ($j = 1, 2, \ldots, N_i$).

Note that neither of these strategies take account of the states of nodes at other stages in the system. However the existence of other stages may have an effect on the optimal routeing vector for a given strategy. In [10] two further scheduling strategies were considered. These strategies restricted the behaviour of some of the queues in a stage such that they could only accept jobs if the server was active; the remaining queues operated a fixed or selective strategy. Whilst these strategies provided interesting intermediate behaviours to the fixed and selective strategies they did not give an improved performance. Obviously many further strategies are possible, but these would either be difficult to compare numerically, difficult to implement or not give any improvement over the two cases considered. It would, however, be worth considering the case where jobs are routed differently for every operational state of the stage, whilst maintaining the 'no loss' condition. Such a study would be a large undertaking.

6 Numerical results

Numerical experiments were carried out in order to determine both the accuracy of the approximations suggested and the characteristics of the behaviour of the pipeline system. In most practical situations it is normal to find nodes with a high degree of reliability, however, as is the case with most models involving node breakdowns, systems of such nodes may behave much like nodes without breakdowns. It has been necessary, therefore, to consider here nodes with somewhat extreme characteristics in order to highlight the strengths and weaknesses of the approximations and to show the limiting behaviour of such a system of nodes. However, it is also true to say that even nodes with a high degree of reliability may suffer rare, but prolonged, breakdowns which can have a significant effect on performance measures.

If few arrivals occur during a period of breakdown (i.e. $\eta \sim \lambda$) then the effect of a failure on the sizes of the queues at a stage will be minimal, assuming the node is reasonably reliable, just as for the single stage parallel node models considered previously. Also if the service rate is of a similar order ($\mu \approx \lambda$) then the departures will not be unduly interrupted by failures, so the arrivals at the following stage may be assumed to be nearly Poisson, hence the single stage approximation will work well for either routeing strategy.

However, if the repair rate is small compared to the arrival rate, then many arrivals will occur during a breakdown period. Under the selective routeing strategy this will cause the other nodes at that stage to be more heavily loaded, causing the queues at those nodes to grow. With the fixed routeing strategy the queue of a broken node will grow larger during a period of breakdown, leading to a large backlog of jobs if the load is sufficiently high. The solution of the model for the stage where this behaviour occurs is still exact, but the arrivals at the next stage are now distinctly 'bursty', rather than nearly Poisson and so the accuracy of the approximations is in question.

If N_i is large then the effect of an individual failure at stage i will be reduced, since one node failing out of N_i identical nodes will mean a reduction of at most $1/N_i$ in the overall service at stage i. In fact the reduction could be considerably less than $1/N_i$ if the load at stage i is not excessively high and the selective routeing strategy is used, since the remaining nodes will be less likely to be idle. Since the arrivals at stage $i+1$ are in fact the departures from stage i then any reduction in the effect of failures at stage i will result in an improvement in the approximation of the arrivals at stage $i+1$ as a Poisson stream. Thus, although the 2-stage approximation becomes too costly to use when N_i is large, the accuracy of the simple approximation can be seen to improve in general (assuming the repair rates are sufficiently great).

Figures 2 and 3 show the average response time of a 2 stage pipeline where there are 2 nodes at each stage and all the nodes are identical. In Figure 2 the routeing strategy is fixed and in Figure 3 it is selective, in both cases the routeing vectors are simple and identical for each node, i.e. $(\frac{1}{2}, \frac{1}{2})$. Results are given for the simple (Poisson) approximation, the 2-stage (full Markov modulated) approximation and simulation.

In both figures as the arrival rate increases the response time increases as expected and the average response time is higher under the fixed strategy. When the load is light all three methods give very similar results (for both strategies), but as the load increases the simple approximation becomes somewhat less accurate than the 2-stage approximation. In Figure 4 the differences between the two approximate methods are highlighted further.

Here the structure of the pipeline is the same, but the nodes are not reliable. As mentioned earlier the simple approximation becomes much less accurate when the duration of the periods of inoperation is increased. This is shown in Figure 5, where once again there are 4 identical nodes in a 2 stage pipeline, showing results for the selective strategy. The

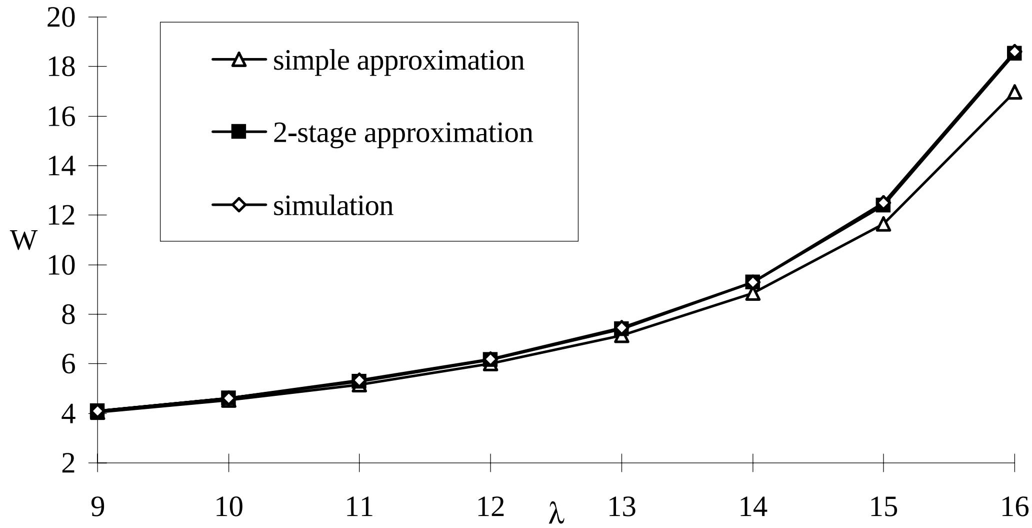

Figure 2: Average response time as a function of arrival rate for a 2 stage service where each stage has 2 identical servers and a fixed routeing strategy
$K = 2,\ N_i = 2,\ \mu_{i,j} = 10,\ \xi_{i,j} = 0.01,\ \eta_{i,j} = 0.1,\ i = 1, 2,\ j = 1, 2$

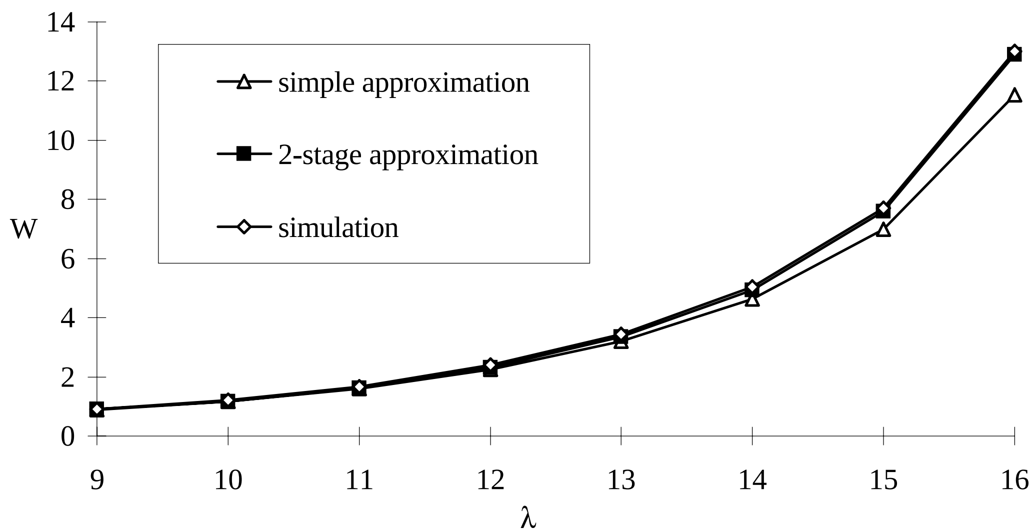

Figure 3: Average response time as a function of arrival rate for a 2 stage service where each stage has 2 identical servers and a selective routeing strategy
$K = 2,\ N_i = 2,\ \mu_{i,j} = 10,\ \xi_{i,j} = 0.01,\ \eta_{i,j} = 0.1,\ i = 1, 2,\ j = 1, 2$

overall reliability of the nodes $(\eta/(\eta+\xi))$ remains constant, but the durations of the periods of operation and in-operation are increased exponentially. When the failure and repair rates are relatively large the effect of failures is minimal and so both approximations work well, however as the repair and failure rates decrease the simple approximation become highly inaccurate as the arrivals become more and more 'bursty'.

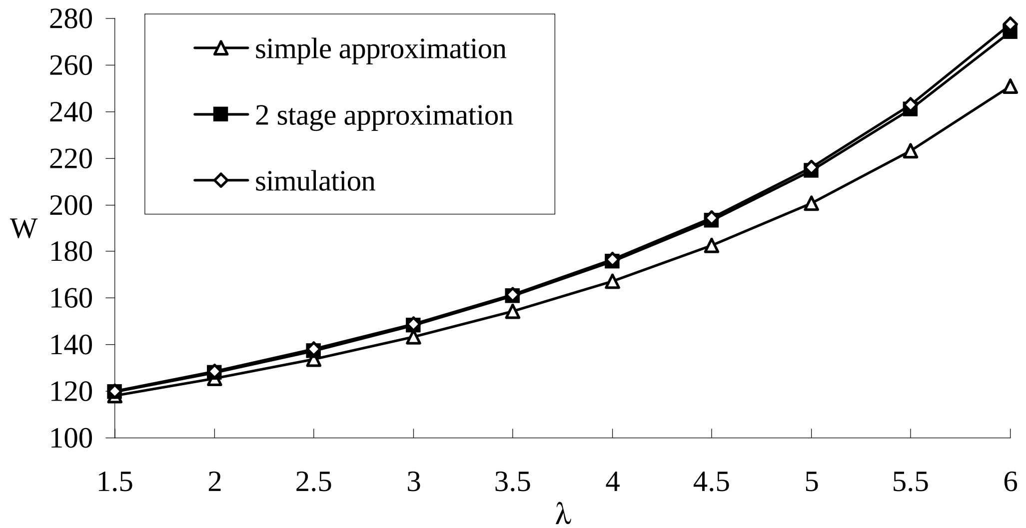

Figure 4: Average response time as a function of arrival rate for a 2 stage service where each stage has 2 identical servers and a fixed routeing strategy
$K = 2,\ N_i = 2,\ \mu_{i,j} = 10,\ \xi_{i,j} = 0.01,\ \eta_{i,j} = 0.01,\ i = 1, 2,\ j = 1, 2$

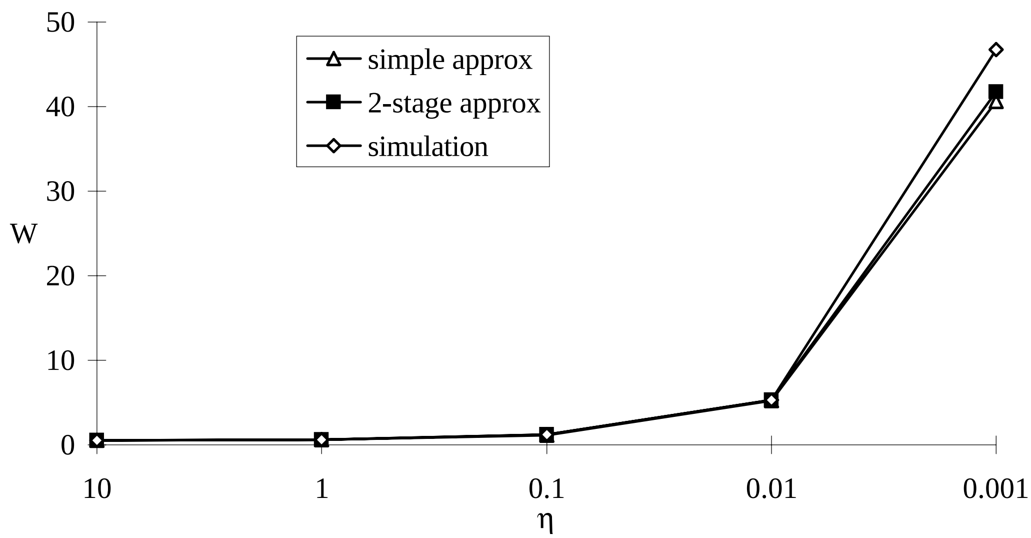

Figure 5: Average response time as a function of repair rate for a 2 stage service where the proportion of time operative is a constant
$K = 2,\ N_i = 2,\ \mu_{i,j} = 10,\ \xi_{i,j} = \eta_{i,j}/10,\ \lambda = 2,\ i = 1, 2,\ j = 1, 2$

The 2-stage approximation does not always give such accurate results as those shown above. With the fixed routeing strategy in particular a large backlog of jobs may build up during a period of failure, thus the probability of the queue being non-empty may be significantly less for sometime immediately following a failure than after a long period of operation. However, such node characteristics would be somewhat extreme: the average number of jobs in the queue would have to be small during operation, but large during inoperation, thus λ would have to be significantly less than μ ($\lambda = \mu/2$, say) and the period of inoperation would have to be very long ($\eta \ll \lambda, \eta < \lambda/10^4$, say). A simulation of such a pipeline would take an exceedingly long time to produce an accurate result.

In general the optimal routeing weights are not greatly affected by the presence of preceding stages, but an unbalanced system will perform significantly worse as a result of increased 'burstiness'. This is illustrated in the following 4 graphs, each of which shows the performance at the final stage of a pipeline only. Figure 6 shows the average response time at a single stage of 2 nodes as a function of the proportion of jobs sent to node 1 when both are available (q), thus this is a selective routeing strategy with routeing vector $(q, 1-q)$. The 2-stage approximation takes account of the behaviour at a preceding stage which has long periods of inoperation whereas the simple approximation considers the same stage in isolation; the arrival rate is identical in both cases.

Figure 7 shows the same system operating the fixed routeing strategy. Clearly in this (extreme) case the optimal routeing vector is slightly altered by considering a preceding stage, but perhaps more significant is the much greater steepness exhibited by the curve of the 2-stage approximation. Thus a routeing vector which gives a near optimal average response time when the stage is considered in isolation could give a very poor response time when the preceding stage is taken into account.

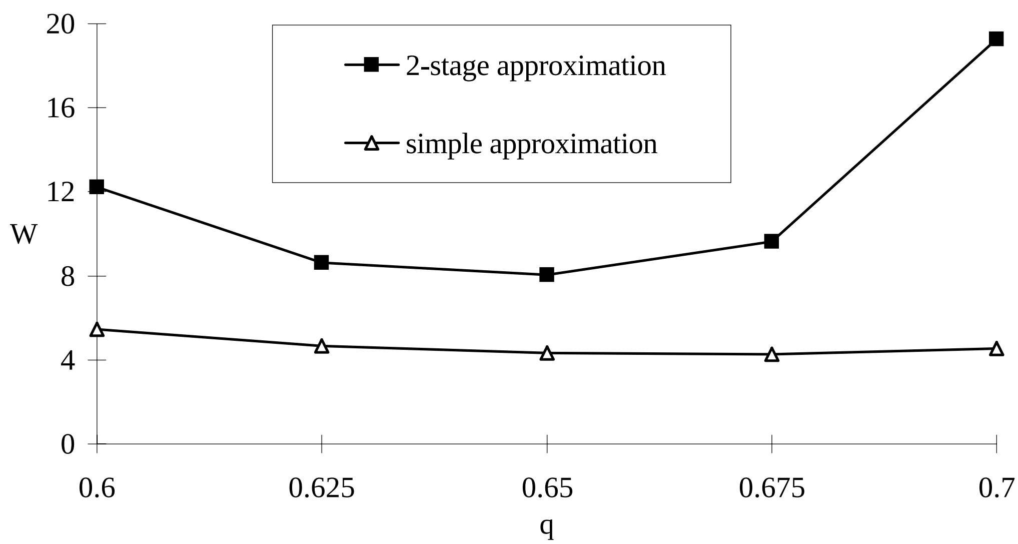

Figure 6: Average response time as a function of job share q
at the 2nd stage of a 2 stage pipeline with a selective routeing strategy
$K = 2,\ N_i = 2,\ \lambda = 17,\ \mu_{1,j} = 10,\ \mu_{2,1} = 14,\ \mu_{2,2} = 9,$
$\xi_{1,j} = 0.0001,\ \eta_{1,j} = 0.001,\ \xi_{2,j} = 0.01,\ \eta_{2,1} = 0.1,\ \eta_{2,2} = 0.07,\ i = 1, 2,\ j = 1, 2$

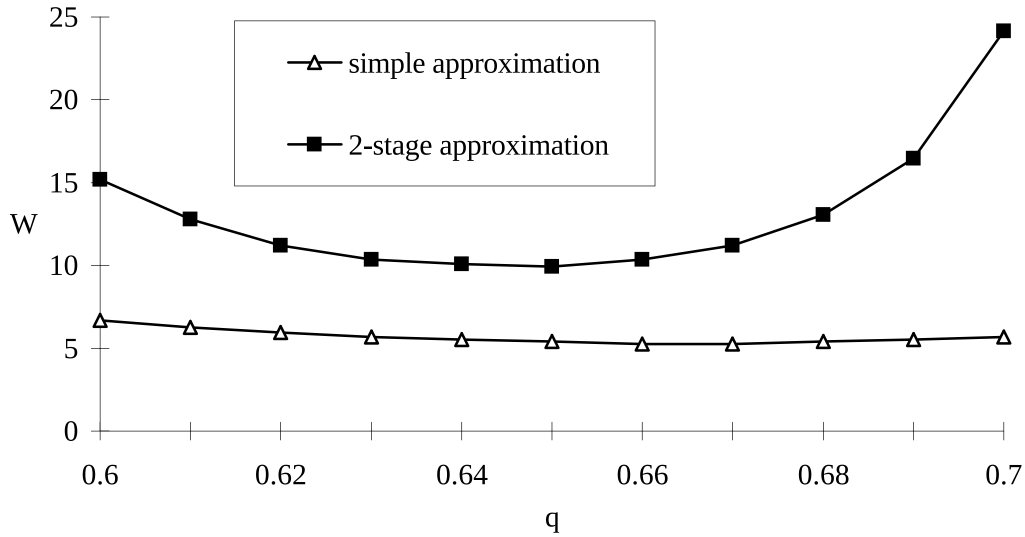

Figure 7: Average response time as a function of job share q
at the 2nd stage of a 2 stage pipeline with a fixed routeing strategy
$K = 2$, $N_i = 2$, $\lambda = 17$, $\mu_{1,j} = 10$, $\mu_{2,1} = 14$, $\mu_{2,2} = 9$, $\xi_{1,j} = 0.0001$,
$\eta_{1,j} = 0.001$, $\xi_{2,j} = 0.01$, $\eta_{2,1} = 0.1$, $\eta_{2,2} = 0.07$, $i = 1, 2$, $j = 1, 2$

In Figure 8 a 3-stage pipeline is illustrated. The performance measure displayed is the average time a job spends at the final stage of the pipeline. Results obtained by simulation are compared with the simple (stage in isolation) approximation and two versions of the two stage approximation, isolated and progressive: the first where the approximation is applied to the final 2 stages of the pipeline in isolation (referred to as 2-stage approximation); and the second where the approximation is applied to each stage of the pipeline in turn (as described above, referred to as 3-stage approximation). In addition results obtained by simulation are shown for the time spent at the final stage of a 2 stage pipeline with identical parameters for comparison. It is interesting to note that there is a significant increase in average response time calculated by simulation for the 3rd stage of a 3-stage pipeline as opposed to the 2nd stage of a 2-stage pipeline with the same parameters. Unfortunately the same cannot be said for calculations made by approximation where there is only a slight difference between the 2 and 3 stage results. Clearly therefore the earlier assertion that the performance of one stage of a pipeline is heavily dominated by its preceding stage is not an altogether accurate one. All the approximations accurately ape the curve of the simulations, albeit with some displacement. There is some deviation from this as one moves away from the optimum routeing vector, although this is much more marked in the simple approximation. Again there is a slight difference in the optimal routeing vector between the approximations, but the 2-stage approximations are fairly accurate when compared to the optimal routeing vector found by simulation. As would be expected in this case the progressive 2-stage approximation gives a much more accurate fit to the simulated models than does the simple approximation, although there is still an appreciable error.

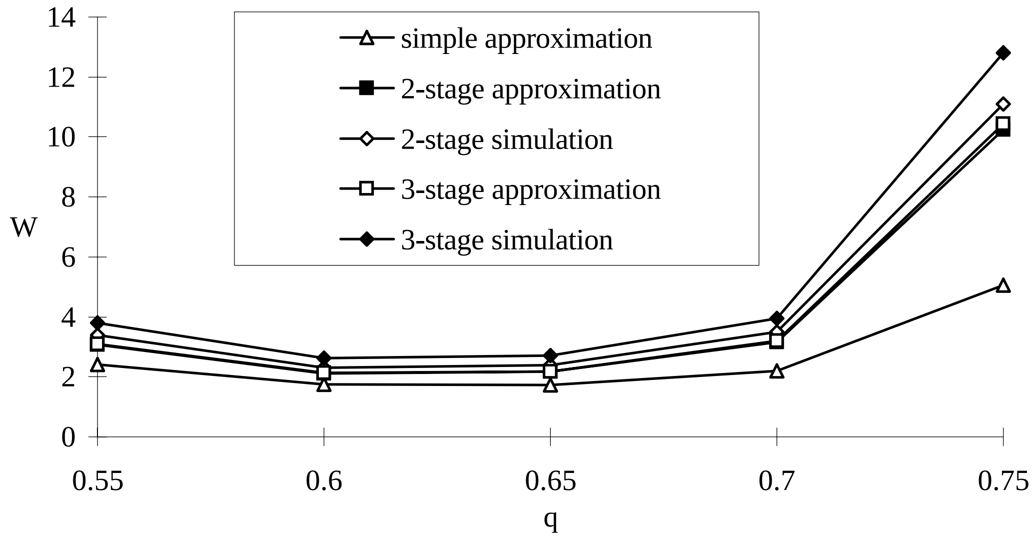

Figure 8: Average response time as a function of job share q
at the final stage of a pipeline with a selective routeing strategy
$N_i = 2,\ \lambda = 15,\ \xi_{i,j} = 0.01,\ \mu_{1,j} = 10,\ \eta_{1,j} = 0.2,\ i = 1..3,\ j = 1,2$
2-stage pipeline: $\mu_{2,1} = 12,\ \mu_{2,2} = 8,\ \eta_{2,1} = 0.2,\ \eta_{2,2} = 0.1$
3-stage pipeline: $\mu_{2,j} = 10,\ \mu_{3,1} = 12,\ \mu_{3,2} = 8,\ \eta_{2,j} = 0.2,\ \eta_{3,1} = 0.2,\ \eta_{3,2} = 0.1$

7 Conclusions

Under many common practical situations a fairly good approximation to this pipeline model can be made by considering each of the stages in isolation. This is particularly true when the nodes are highly reliable, periods of in-operation are relatively short and the number of nodes at a stage is relatively large. When these conditions do not apply it is necessary to use a more involved Markov-modulated approximation such as the 2-stage approximation suggested here. In certain circumstances it may be advantageous to look for alternative approximations, more detailed than the simple approximation, but less costly than the 2-stage approximation. The exact choice of what approximation to consider will depend on many variables (the node characteristics, available computational power, desired accuracy, etc) which are out of the scope of this paper, but are worthwhile directions of research none the less. Also it may be worth considering other heuristics to predict the optimal routeing vectors in light of the increased penalties to an unbalanced system when previous stages of service are involved.

For models with characteristics like some of those illustrated here, i.e. N_i small and fairly long periods of inoperation, simulations need to be run for a very long time before producing a steady-state result. Typically simulations were run overnight, but in several cases run times in excess of 24 hours were necessary. In contrast the analytical results were much faster to obtain. In the worst case it took around 3 minutes to obtain a single result running on the same platform, typically it took less than 1 minute. Admittedly our simulation program did not employ any optimisation for rare events and the hardware on which it ran could have been substantially faster. However these are still excessively long run times and the analytical method is far superior in that respect. This in itself is clear justification for attempting to find suitable approximations for these models. The inaccuracy in the predictions will, in most circumstances, be sufficiently small for the speed of the calculation to be a definite bonus. In every case the approximation has been observed to underestimate the average response time of the pipeline. This is due to the smoothing out of bursty behaviour. As such, the approximation may be viewed as being an estimation of lower bound of the average response time, although this has only been shown numerically and not proved.

It was stated in the introduction that the method presented here enables interesting performance measures to be derived for a class of models for which this was previously not possible. This method is also obviously applicable to other, more general, network models and it is therefore interesting to observe the accuracy of the approximations and predictions of an optimal routeing vector. In particular it would be interesting to apply this method to models where the progressive condition is relaxed, thus a job completing service at stage i may proceed to any stage j such that $j > i$. A further consideration would be to allow jobs to feedback in the network so that a a job completing service at stage i may proceed to any stage in the network (including i). In order to solve such a model it would be necessary to iterate the method many times until stability is reached. It would be a major task to find convergence conditions for such an iterative method.

This paper serves as a good introduction to this problem area as well as presenting and evaluating an important application of approximation theory.

References

[1] R. Chakka and I. Mitrani, A Numerical Solution Method for Multiprocessor Systems with General Breakdowns and Repairs, *Proceedings of 6th International Conference on Modelling Techniques*, 1991.

[2] R.A. Marie, An Approximate Analytical Method for General Queueing Networks, *IEEE Transactions on Software Engineering*, SE-5(5), pp. 530-538, 1979.

[3] N. Mikou, A Two-Node Jackson Network Subject to Breakdowns, *Stochastic Models* **4**, pp. 523–552, 1988.

[4] N. Mikou, O. Idrissi-Kacemi and S. Saadi, Two Processors Only Interacting During Breakdown: The Case Where the Load is Not Lost, *Queueing Systems*, 19, pp. 301-317, 1995.

[5] I. Mitrani Networks of Unreliable Computers, in: E. Gelenbe and R. Mahl, eds., *Computer Architectures and Networks*, North-Holland, 1974.

[6] I. Mitrani and R. Chakka, Spectral Expansion Solution for a Class of Markov Models: Application and Comparison with the Matrix-Geometric Method, *Performance Evaluation*, 23, pp. 241-260, 1995.

[7] I. Mitrani and D. Mitra, A Spectral Expansion Method for Random Walks on Semi-Infinite Strips, in: R. Beauwens and P. de Groen, eds., *Iterative Methods in Linear Algebra*, North-Holland, 1992.

[8] I. Mitrani and P.E. Wright, Routing in the Presence of Breakdowns, *Performance Evaluation* **20**, pp. 151-164, 1994.

[9] N. Thomas and I. Mitrani, Routing among different stages, in: F. Baccelli, A. Jean-Marie and I. Mitrani, eds., *Quantitative Methods in Parallel Systems*, Springer-Verlag, 1995.

[10] N. Thomas and I. Mitrani, Routing Among Different Nodes Where Servers Break Down Without Losing Jobs, *Proceedings of IEEE International Computer Performance and Dependability Symposium*, pp. 246-255, 1995.