UNIVERSITY OF
NEWCASTLE

University of Newcastle upon Tyne

COMPUTING
SCIENCE

Empirical and Analytical Evaluation of Systems with Multiple Unreliable Servers

J. Palmer and I. Mitrani

TECHNICAL REPORT SERIES

No. CS-TR-936 December, 2005

TECHNICAL REPORT SERIES

No. CS-TR-936 December, 2005

Empirical and Analytical Evaluation of Systems with Multiple Unreliable Servers

J. Palmer, I. Mitrani

Abstract

We construct, analyze and solve models of systems where a number of servers offer services to an incoming stream of demands. Each server goes through alternating periods of being operative and inoperative. The objective is to evaluate and optimize performance and cost metrics. A large real-life data set containing information about server breakdowns is analyzed first. The results indicate that the durations of the operative periods are not distributed exponentially. However, hyperexponential distributions are found to be a good fit for the observed data. A model based on these distributions is then formulated, and is solved exactly using the method of spectral expansion. A simple approximation which is accurate for heavily loaded systems is also proposed. The results of a number of numerical experiments are reported.

© 2005 University of Newcastle upon Tyne.

Printed and published by the University of Newcastle upon Tyne,
Computing Science, Claremont Tower, Claremont Road,
Newcastle upon Tyne, NE1 7RU, England.

Bibliographical details

PALMER, J., MITRANI, I.

Empirical and Analytical Evaluation of Systems with Multiple Unreliable Servers
[By] J. Palmer, I. Mitrani

Newcastle upon Tyne: University of Newcastle upon Tyne: Computing Science, 2005.

(University of Newcastle upon Tyne, Computing Science, Technical Report Series, No. CS-TR-936)

Added entries

UNIVERSITY OF NEWCASTLE UPON TYNE
Computing Science. Technical Report Series. CS-TR-936

Abstract

We construct, analyze and solve models of systems where a number of servers offer services to an incoming stream of demands. Each server goes through alternating periods of being operative and inoperative. The objective is to evaluate and optimize performance and cost metrics. A large real-life data set containing information about server breakdowns is analyzed first. The results indicate that the durations of the operative periods are not distributed exponentially. However, hyperexponential distributions are found to be a good fit for the observed data. A model based on these distributions is then formulated, and is solved exactly using the method of spectral expansion. A simple approximation which is accurate for heavily loaded systems is also proposed. The results of a number of numerical experiments are reported.

About the author

Jennie Palmer is a demonstrator at the School of Computing Science, Newcastle University.

Isi Mitrani studied Mathematics at the Universities of Sofia and Moscow (Diploma, 1965), and Operations Research at the Technion, Haifa (MSc, 1967). He joined the University of Newcastle in 1968 as a Programming Advisor, became a Lecturer in 1969 (PhD, 1973), Reader in 1986 and Professor in 1991. He has held several visiting positions, including sabbatical years at INRIA (Le Chesnay) and Bell Laboratories (Murray Hill). His research interests are in the areas of Probabilistic Modelling, Performance Evaluation and Optimization. Publications include 4 authored books, 3 edited books, more than 100 journal and conference papers and one patent.

Suggested keywords

MULTI-SERVER QUEUES,
BREAKDOWNS,
GRID COMPUTING,
SPECTRAL EXPANSION

Empirical and Analytical Evaluation of Systems with Multiple Unreliable Servers

J. Palmer I. Mitrani

School of Computing Science, University of Newcastle, NE1 7RU, UK
[jennie.palmer, isi.mitrani]@ncl.ac.uk
Submission category: Regular paper Contact author: J. Palmer

Abstract

We construct, analyze and solve models of systems where a number of servers offer services to an incoming stream of demands. Each server goes through alternating periods of being *operative* and *inoperative*. The objective is to evaluate and optimize performance and cost metrics. A large real-life data set containing information about server breakdowns is analyzed first. The results indicate that the durations of the operative periods are not distributed exponentially. However, hyperexponential distributions are found to be a good fit for the observed data. A model based on these distributions is then formulated, and is solved exactly using the method of spectral expansion. A simple approximation which is accurate for heavily loaded systems is also proposed. The results of a number of numerical experiments are reported.

Keywords: Multi-server queues, Breakdowns, Grid computing, Spectral Expansion.

1 Introduction

Service provisioning systems have been the subject of considerable interest in recent years. They come in different flavours, and may be described either in terms of web services, or as computing grids. However, the general idea is that a (possibly distributed) cluster of computers (to be referred to as 'servers') is made available for the execution of tasks submitted by users through a central dispatcher. In such systems, the quality of service, and the cost of providing it, is important both to the users and to the provider. A problem of particular interest is the effect that server breakdowns and other outages have on the performance of the system. That problem is the topic of the present paper.

The starting point of the study is a model where demands, or *jobs*, arrive in a Poisson stream into a common queue and are served by a number, N, of servers in parallel. Each server goes through alternating periods of being *operative* and *inoperative*, independently of the others; the events causing a change of server state will be referred to as *breakdowns* and *repairs*, although in practice they may have other causes (e.g., scheduled maintenance, or tasks of higher priority). A system of this type is illustrated in figure 1.

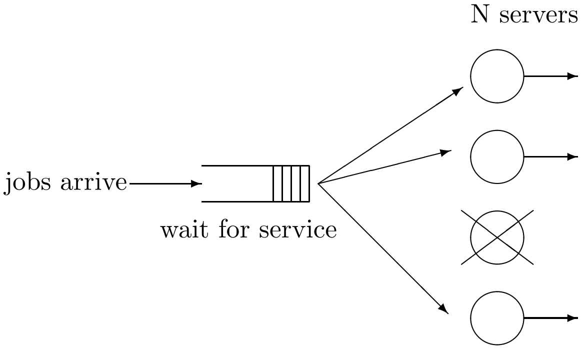

Figure 1: A multi-server system with breakdowns and repairs

Among the questions one may wish to ask in this context are:

1. How does the system perform? A common performance measure is the average response time or, equivalently, the average number of jobs present.

2. What is the minimum number of servers that would ensure a desired level of performance?

3. If there is a trade-off between the cost of making jobs wait and that of providing servers, what is the optimal number of servers that should be used?

The answers to all those questions depend not only on rates of demand and service, but also on the nature of the operative and inoperative intervals. Moreover, that 'nature' encompasses not only the means, but also the distributions of those random variables. Exactly how one should model breakdowns and repairs is a point that has not received much attention in the literature. There are several papers on the subject of multi-server queues with service interruptions (e.g., see [1, 2, 5, 7, 9]). However, they all make the assumption that both the operative and inoperative intervals are distributed exponentially. The validity of that assumption has not been investigated.

The contribution of this paper is two-fold: First, we analyze a large real-life data set containing information about server breakdowns and repairs; this data has been collected and made available to us by Sun Microsystems. The results indicate that the distribution of repair times is reasonably close to an exponential, but that of the operative intervals is not. A good approximation for the latter is the hyperexponential distribution.

Second, we show how to obtain an exact solution for a model with non-exponentially distributed operative and/or inoperative intervals. This is done by reducing the problem to a Markov-modulated queue (with a suitably defined Markovian environment) and then solving it by spectral expansion. The solution can be computationally expensive and prone to numerical difficulties when the number of servers (and/or the number of phases in the hyperexponential distributions) is large. In those cases, one can apply a simple approximation whose accuracy improves when the offered load increases.

The statistical analysis of the Sun data set, including the fitting and testing of hyperexponential distributions, is described in section 2. The mathematical model based on these distributions, together with its exact and approximate solutions, is presented in section 3. Some numerical results illustrating the effects of different parameters on performance and costs are described in section 4. Section 5 contains a summary and conclusions, and mentions an open problem.

2 The data set

The Sun Microsystems data set contains 140,000 rows of data, each row giving details of a particular *event*, corresponding to a server breakdown. Of immediate interest were the fields representing the time a server was inoperative, referred to as *Outage Duration*, and the time between a server breakdown and its next breakdown, referred to as *Time Between Events*. The lengths of operative periods can then be calculated as illustrated in figure 2.

A small proportion of the data set (less than 4%) contained anomalous entries (Time Between Events was smaller than the Outage Duration). This data was ignored. Empirical probability density functions (histograms) were generated for both the operative and inoperative periods, by grouping observed period lengths into appropriate intervals.

Consider the operative periods. If the i^{th} observation interval has a midpoint x_i, and f_i of the observed operative periods fall into that interval, then the corresponding empirical density, d_i, is obtained by assuming that the operative periods take value x_i with probability $p_i = f_i/n$, where n is the total number of observations; then $d_i = p_i/\delta_i$, where δ_i is the length of the i^{th} interval. A similar procedure is followed for the inoperative periods.

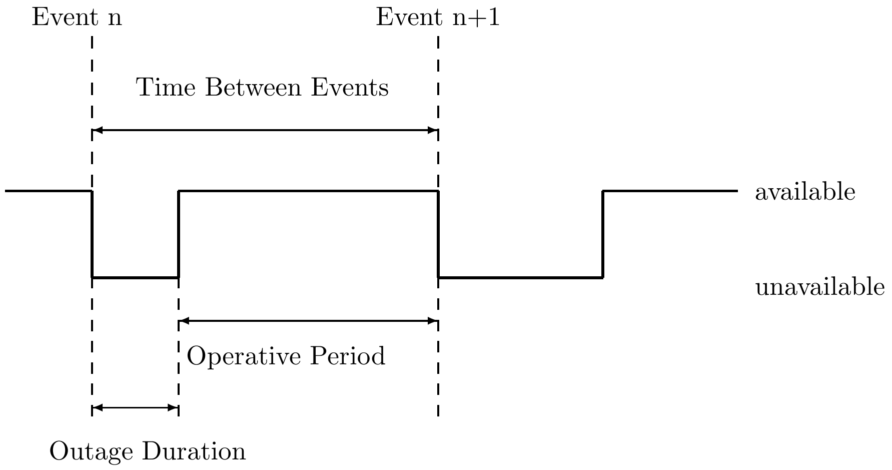

Figure 2: Alternating Periods of availability and unavailability

The empirical densities of the operative and inoperative periods are shown in figures 3 and 4, respectively (together with the fitted hypothetical distributions, to be described below). In each case, the observed range of values was divided into intervals of equal length. The time unit has been deliberately omitted, for reasons of confidentiality.

The two empirical densities were used to derive estimates for the moments of the corresponding distributions. The k^{th} estimated moment, $\tilde{M}_k$, is calculated as

$$\tilde{M}_k = \sum x_i^k p_i \;\; ; \;\; k = 1, 2, \dots \;, \tag{1}$$

where the sum extends over all empirical values.

The estimated variance, $\tilde{V}$, and coefficient of variation, $\tilde{C}^2$, are given by

$$\tilde{V} = \tilde{M}_2 - \tilde{M}_1^2 \;\; ; \;\; \tilde{C}^2 = \frac{\tilde{M}_2}{\tilde{M}_1^2} - 1 \; . \tag{2}$$

From the empirical densities one can also obtain the empirical cumulative distribution functions,

$$\tilde{F}(x_i) = \sum_{j=1}^{i} p_j \; . \tag{3}$$

The null hypothesis that an empirical cumulative distribution function, $\tilde{F}(x)$, is consistent with a given hypothetical one, $F(x)$, can be tested by means of the Kolmogorov-Smirnov *goodness-of-fit* test, [3]. The hypothesis is accepted if the value of the statistic D, calculated as

$$D = \max_{x_i} \left| F(x_i) - \tilde{F}(x_i) \right| , \tag{4}$$

is sufficiently small, for a given level of significance; otherwise it is rejected (the higher the level of significance, the more difficult it is to pass the test).

On the basis of the Sun data set, the hypothesis that the server operative periods are distributed exponentially with a mean obtained from the sample, is strongly rejected. The calculated value of the Kolmogorov-Smirnov statistic, using 50 points x_i, was $D = 0.4742$; it would have had to have been less than 0.19 to pass the test at 5% significance, and less than 0.23 at 1% significance.

The inoperative intervals are more likely to be exponentially distributed: that hypothesis also fails the Kolmogorov-Smirnov test, but not so badly. Moreover, we shall see later that an exponential distribution with a slightly different mean passes the test quite comfortably.

The next task is to find hypothetical distributions that do agree with the empirical densities for the operative and inoperative periods. An indication of where to look is provided by the values of the estimated coefficients of variation, which are both greater than 1 ($\tilde{C}^2 = 4.6$ for the operative periods). This suggests that the family of hyperexponential distributions, all of which have coefficients of variation greater than 1, may be a good place to start. An n-phase hyperexponential density function is a linear combination of n exponential densities with different parameters; it has the form

$$f(x) = \sum_{j=1}^{n} \alpha_j \xi_j e^{-\xi_j x} \;\; ; \;\; \alpha_j, \xi_j > 0 \;\; ; \;\; \sum_{j=1}^{n} \alpha_j = 1 \; . \tag{5}$$

Such a density is defined by $2n-1$ parameters: n 'rates' ξ_j and $n-1$ 'weights' α_j, (the last weight is given by the normalizing condition in (5)). Hence, an n-phase hyperexponential distribution is completely determined by its first $2n-1$ moments. Those moments are expressed in terms of the parameters as follows:

$$M_k = \sum_{j=1}^{n} \frac{k!\alpha_j}{\xi_j^k} \;\; ; \;\; k = 1, 2, \ldots 2n-1 \; . \tag{6}$$

Thus, a hyperexponential distribution can be fitted to a given empirical density by first choosing the number of phases, n, and then determining the parameters α_j and ξ_j so that

$$M_k = \tilde{M}_k \;\; ; \;\; k = 1, 2, \ldots 2n-1 \; . \tag{7}$$

The above procedure was carried out for the operative periods, with $n = 3$. The empirical density provided the first 5 estimated moments, $\tilde{M}_1, \ldots, \tilde{M}_5$. However, it turned out that the task of solving (7) is computationally difficult, because those equations are highly non-linear. Iterative methods such as Newton or Gauss-Seidel [8] failed to converge. Instead, the weights α_j were eliminated explicitly from the first two equations in (7), and a brute force search was used to find the rates ξ_j that minimize

$$\min_{\xi_1,\xi_2,\xi_3} \sum_{k=3}^{5} |M_k - \tilde{M}_k| \, . \tag{8}$$

It was observed that two of the rates thus calculated were almost equal. In other words, a 2-phase hyperexponential distribution fits the data as well as a 3-phase one (re-running the Gauss-Seidel iterations for $n = 2$ resulted in convergence).

The 2-phase hyperexponential distribution that provides the best fit to the empirical density has parameters $\alpha_1 = 0.7246$, $\alpha_2 = 0.2754$, $\xi_1 = 0.1663$ and $\xi_2 = 0.0091$. That is, approximately 72% of the operative periods are distributed exponentially with mean 6, and 28% of them are distributed exponentially with mean 110. That density is shown together with the empirical one in figure 3. It passes the Kolmogorov-Smirnov goodness-of-fit test at level of significance 5%, and also at 10% (the calculated statistic with 50 points x_i has value $D = 0.1412$, whereas the 5% and 10% critical values are 0.19 and 0.17 respectively).

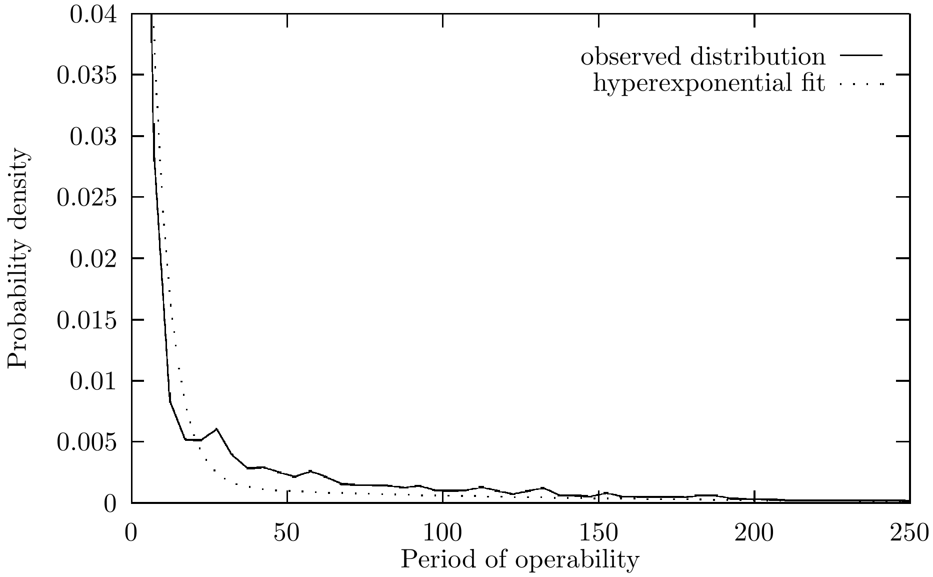

Figure 3: Densities of operative periods $(0 - 250)$

Similarly, for the inoperative periods, a best-fit 2-phase hyperexponential distribution was found with weights $\beta_1 = 0.9303$ and $\beta_2 = 0.0697$, and rates $\eta_1 = 25.0043$ and $\eta_2 = 1.6346$. This represents a mixture where approximately 93% of the inoperative periods are distributed exponentially with mean 0.04 and 7% are distributed exponentially with mean 0.61. The fitted density, together with the empirical one, is shown in figure 4. It passes the Kolmogorov-Smirnov test at both the 5% and the 10% level of significance (the calculated statistic with 40 points x_i is $D = 0.1832$; the 5% and 10% critical values are 0.21 and 0.19 respectively).

In view of the fact that the second component of the fitted hyperexponential distribution contributes very little to the mixture, it is not unreasonable to model the inoperative periods as being distributed exponentially. Indeed, the first component on its own, i.e. the exponential distribution with mean 0.04, passes the Kolmogorov-Smirnov test at level 5% (fails, but not badly, at 10%).

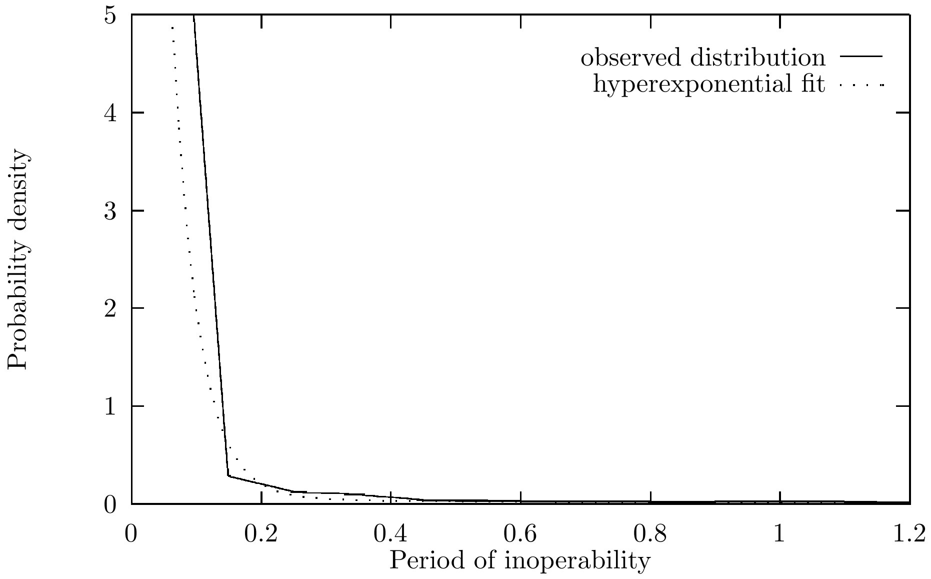

Figure 4: Densities of Inoperative periods $(0 - 1.2)$

3 The model and its solution

The preceding section offers evidence that, in a realistic model of a multi-server system with breakdowns and repairs, it is appropriate to assume that the distribution of the operative periods is hyperexponential, with n phases and suitably chosen weight and rate parameters α_j and ξ_j $(j = 1, 2, \ldots, n)$. Similarly, the inoperative periods can be assumed to have a hyperexponential distribution with m phases and weight and rate parameters β_k and η_k $(k = 1, 2, \ldots, m)$.

Assume further that jobs arrive according to a Poisson process with rate λ, and their required service times are distributed exponentially with mean

$1/\mu$. The queue is unbounded. All arrival, service, breakdown and repair events are mutually independent. 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 above assumptions ensure that the system is modelled by a Markov process whose state at any moment in time is described by a triple $S = (\mathbf{X}, \mathbf{Y}, Z)$. Here, $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ is a vector whose j^{th} element, x_j, indicates how many servers are in phase j of an operative period; the number of operative servers is $x = x_1 + x_2 + \ldots + x_n$. Similarly, $\mathbf{Y}$ is a vector whose k^{th} element, y_k, indicates how many servers are in phase k of an inoperative period; the number of inoperative servers is $y = y_1 + y_2 + \ldots + y_m$. Finally, Z is the number of jobs present. The valid states must of course satisfy $x + y = N$, where N is the total number of servers.

The instantaneous transition rates from state $S = (\mathbf{X}, \mathbf{Y}, Z)$ to state $S' = (\mathbf{X}', \mathbf{Y}', Z')$ are equal to

$$r(S, S') = \begin{cases} \lambda & \text{if } Z' = Z + 1 \\ \min(Z, x)\mu & \text{if } Z' = Z - 1 \\ x_j \xi_j \beta_k & \text{if } x'_j = x_j - 1,\ y'_k = y_k + 1 \\ y_k \eta_k \alpha_j & \text{if } x'_j = x_j + 1,\ y'_k = y_k - 1 \\ 0 & \text{otherwise} \end{cases} , \tag{9}$$

where all state variables that have not been mentioned have the same values in S and S'.

The stability condition for this queue has a simple form. Denote the average lengths of the operative and inoperative periods by $1/\xi$ and $1/\eta$, respectively. They are given by

$$\frac{1}{\xi} = \sum_{j=1}^{n} \frac{\alpha_j}{\xi_j} \ ; \ \frac{1}{\eta} = \sum_{k=1}^{m} \frac{\beta_k}{\eta_k} . \tag{10}$$

The long-term fraction of time that a given server is operative is equal to $\eta/(\xi + \eta)$. Hence, the steady state average number of operative servers is $N\eta/(\xi + \eta)$; this is independent of the queue of jobs. That queue is stable iff the offered load is less then the average number of operative servers, i.e.

$$\frac{\lambda}{\mu} < \frac{N\eta}{\xi + \eta} . \tag{11}$$

Note that this condition depends only on the averages of the operative and inoperative periods; not on their distribution. However, the queue size distribution, and hence the measures of performance, depend very much on the distributions of operative and inoperative periods. Computing those performance measures is our next task.

3.1 Spectral expansion solution

The model defined here is a special case of a *Markov-modulated queue*, i.e., a queue whose arrival and/or service parameters depend on the state of a Markovian environment. In our case, the state of the environment is described by the vectors $\mathbf{X}$ and $\mathbf{Y}$, specifying the numbers of servers in each of the possible operative/inoperative states. The environment affects the queue via the number of operative servers, which determines the departure rate (second line in the right-hand side of (9)).

Markov-modulated queues can be solved by the method of 'Spectral Expansion' (e.g., see [6]).

The number, s, of different environment states, is equal to the number of ways that the integer N can be partitioned into a sum of $n+m$ components. That number is

$$s = \left(\begin{array}{c} N+n+m-1 \\ n+m-1 \end{array} \right) . \tag{12}$$

One can therefore enumerate the states of the environment using a single integer, i, called 'operational mode': $i = 0, 1, \ldots, s-1$. For example, in a system where $N = 2$, $n = 2$ and $m = 1$, the operational mode i may take 6 different values:

i=0: 2 inoperative servers
i=1: 1 operative in phase 1 and 1 inoperative
i=2: 1 operative in phase 2 and 1 inoperative
i=3: 2 operative in phase 1
i=4: 1 operative in phase 1 and 1 operative in phase 2
i=5: 2 operative in phase 2

The system is then said to be in state (i, j) if the operational mode is i and there are j jobs present ($i = 0, 1, \ldots, s-1$; $j = 0, 1, \ldots$). The transition rates (9) can be expressed in terms of the following $s \times s$ matrices:

(a) Matrix A contains transition rates that change the operative mode but not the number of jobs: from state (i, j) to state (k, j) ($0 \leq i, k < s,\ i \neq k$). The main diagonal of A is zero by definition. In the above example, with $s = 6$, the matrix A has the form

$$A = \begin{bmatrix} 0 & 2\eta\alpha_1 & 2\eta\alpha_2 & 0 & 0 & 0 \\ \xi_1 & 0 & 0 & \eta\alpha_1 & \eta\alpha_2 & 0 \\ \xi_2 & 0 & 0 & 0 & \eta\alpha_1 & \eta\alpha_2 \\ 0 & 2\xi_1 & 0 & 0 & 0 & 0 \\ 0 & \xi_2 & \xi_1 & 0 & 0 & 0 \\ 0 & 0 & 2\xi_2 & 0 & 0 & 0 \end{bmatrix}$$

(b) Matrix B contains transitions that increase the number of jobs in the system by 1: from state (i, j) to state $(k, j+1)$. Since in our model arrivals do not change the operational mode, B is diagonal:

$$B = \lambda I \ ,$$

where I is the identity matrix of order s.

(c) Matrices C_j contain transitions that decrease the number of jobs in the system by 1: from state (i, j) to state $(k, j-1)$. Since departures do not change the operational mode, C_j is diagonal. For the same example, C_j has the form

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

where $\mu_{i,j} = min(i, j)\mu$; Note that these matrices depend on j if $j < N$, but cease to do so when $j \geq N$; then the index j may be dropped and we can write $C_j = C$. Also, $C_0 = 0$ by definition.

Let $p_{i,j}$ be the steady state probability that the system is in state (i, j). Define also the row vectors of probabilities corresponding to states with j jobs in the system:

$$\mathbf{v}_j = (p_{0,j}, p_{1,j}, \ldots, p_{s-1,j}) \ ; \ j = 0, 1, \ldots \ . \qquad (13)$$

Then, the balance equations for the equilibrium probabilities can be written as:

$$\mathbf{v}_j[D^A + B + C_j] = \mathbf{v}_{j-1}B + \mathbf{v}_j A + \mathbf{v}_{j+1}C_{j+1} \ , \ \ j = 0, 1, \ldots \ , \qquad (14)$$

where D^A is the diagonal matrix whose i th diagonal element is equal to the i th row sum of A.

When $j > N$, the matrices in equations (14) do not depend on j. The equations can be re-written in the form of a homogeneous vector difference equation of order 2:

$$\mathbf{v}_j Q_0 + \mathbf{v}_{j+1} Q_1 + \mathbf{v}_{j+2} Q_2 = \mathbf{0} \ ; \ j = N, N+1, \ldots \ , \qquad (15)$$

where $Q_0 = B$, $Q_1 = A - D^A - B - C$ and $Q_2 = C$. Associated with equation (15) is the so-called ‘characteristic matrix polynomial’, $Q(z)$, defined as

$$Q(z) = Q_0 + Q_1 z + Q_2 z^2 \ . \qquad (16)$$

Let z_k be the ‘generalized eigenvalues’, of $Q(z)$ in the interior of the unit disk, and d be their number. Denote by $\mathbf{u}_k$ the corresponding ‘generalized left eigenvectors’. These eigenvalues and eigenvectors satisfy

$$det[Q(z_k)] = 0 \ \ ; \ \ |z_k| < 1 \ \ ; \ \ k = 1, 2, \ldots, d \ , \qquad (17)$$

where $det[Q(z)]$ is the determinant of $Q(z)$. Also,

$$\mathbf{u}_k Q(z_k) = \mathbf{0} \ ; \ \ k = 1, 2, \ldots, d \ . \tag{18}$$

In what follows, the qualification *generalized* will be omitted.

The theory of spectral expansion shows that, when the queue is ergodic, the number of eigenvalues in the interior of the unit disk is equal to the number of states of the Markovian environment. In our case, $d = s$. Moreover, experience indicates that they are simple. Then, the the solution of (15) has the form

$$\mathbf{v}_j = \sum_{k=1}^{s} \gamma_k \mathbf{u}_k z_k^j \ ; \ j = N, N+1, \ldots \ , \tag{19}$$

where γ_1, …, γ_s are some (possibly complex) constants. Those coefficients, and the 'boundary' probabilities, $p_{i,j}$, for $j < N$, are determined from the balance equations (14), for $j = 0, 1, \ldots, N$. This is a set of $(N+1)s$ linear equations with Ns unknown probabilities (the vectors $\mathbf{v}_j$ for $j = 0, 1, \ldots, N-1$), plus the s constants γ_k. However, only $(N+1)s - 1$ 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 the requirement that all probabilities must sum up to 1:

$$\sum_{j=0}^{\infty} (\mathbf{v}_j \cdot \mathbf{1}) = 1 \ , \tag{20}$$

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

3.2 A simple approximation

The exact solution is computationally intensive, and for systems with many operational modes (large number of servers and/or large number of phases in the hyperexponential distributions), that solution may be intractable or prone to numerical problems. In that case, one may accept an approximate solution which is numerically robust and very simple to implement [4].

The approximation consists of discarding all terms in the spectral expansion solution (19), except the one corresponding to the eigenvalue with the largest modulus, z_s (which is always real and positive). That amounts to assuming that the queue size is distributed geometrically with parameter z_s, and is independent of the operational mode. The approximate solution has the form

$$\mathbf{v}_j = \frac{\mathbf{u}_s}{(\mathbf{u}_s \cdot \mathbf{1})} (1 - z_s) z_s^j \ ; \ \ j = 0, 1, \ldots \ . \tag{21}$$

It requires the computation of only one eigenvalue and its corresponding left eigenvector.

It has been shown (see [4]) that the geometric approximation is asymptotically exact when the system is heavily loaded.

4 Numerical results

The solution described in the previous section yields performance metrics which may be used to answer the questions raised in the Introduction. In particular, we can compute the average number of jobs present in the system, L, and hence the average response time, $W = L/\lambda$ (by Little's theorem). Moreover, if it costs c_1 per unit time to hold a job in the system, and c_2 per unit time to provide a server, then the steady state total cost, C, associated with hosting a service cluster may be expressed as

$$C = c_1 L + c_2 N \ . \tag{22}$$

Such a cost function implies that there is a trade-off between the 'user' costs (measured by $c_1 L$), which decrease with N, and the 'provider' costs (measured by $c_2 N$), which increase with N. It can be expected that, for each set of parameters, there will be an optimal number of serves.

Several numerical experiments were carried out in the context of a system where the operative periods have a 2-phase hyperexponential distribution, while the inoperative periods are distributed exponentially (i.e., $n = 2$, $m = 1$). That is, the queue is modulated by an environment which, when there are N servers, has $s = (N + 2)(N + 1)/2$ operational modes. In all cases, the average required service time is $1/\mu = 1$.

In the first experiment, the parameters of the operative and inoperative periods are fixed as for the fitted distributions ($\alpha_1 = 0.7246, \xi_1 = 0.1663, \alpha_2 = 0.2754, \xi_2 = 0.0091, \eta = 25$), and N is varied.

Figure 5 shows how the cost function (22) changes with N, for three different values of the arrival rate. The values of the cost coefficients, $c_1 = 4, c_2 = 1$, reflect a situation where waiting is quite strongly discouraged. As expected, for each λ there is an optimal value of N that minimizes C. Moreover, the heavier the load, the larger the optimal N (the latter is 11 for $\lambda = 7$, 12 for $\lambda = 8$ and 13 for $\lambda = 8.5$).

The next experiment aims to evaluate the effect of operative period variability on performance. The average length of the operative period, $1/\xi = \alpha_1/\xi_1 + \alpha_2/\xi_2$, is kept fixed at 34.62, but the coefficient of variation is varied by changing α_1, α_2 and ξ_1 (the operative periods in phase 1 become larger and less likely). In figure 6, L is plotted against C^2 for two different arrival rates, λ. The average inoperative period is fixed at $1/\eta = 5$. The value $C^2 = 1$ corresponds to the exponential distribution. The first point on each curve, where $C^2 = 0$ (i.e., constant operative periods) was obtained by simulation.

In all cases, the average queue size grows with the coefficient of variation. The effect is weak when the system is lightly loaded, but becomes more pronounced as the load increases. At heavy loads, an assumption of exponentially distributed operative periods can seriously underestimate the

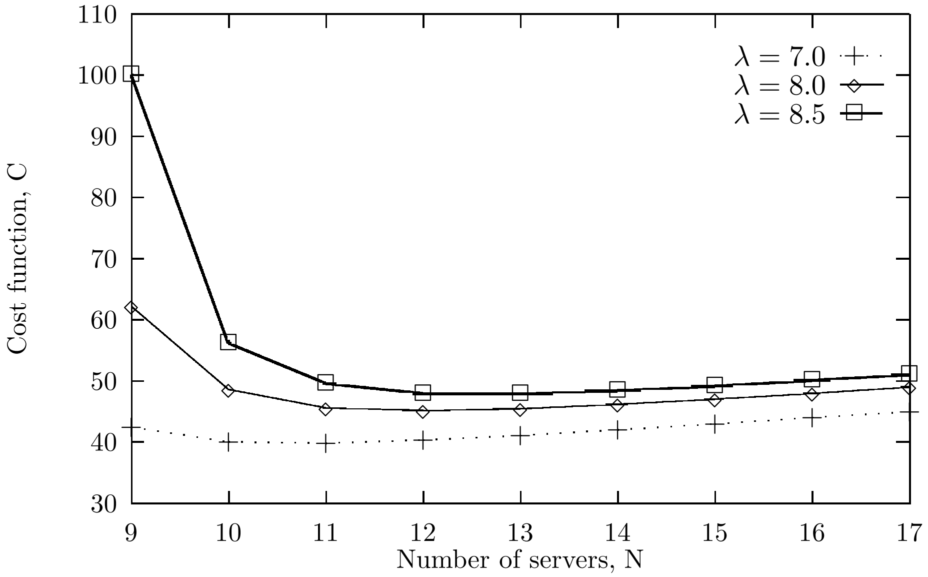

Figure 5: Cost as a function of N

$\alpha_1 = 0.7246, \xi_1 = 0.1663, \xi_2 = 0.0091, \eta = 25, c_1 = 4, c_2 = 1$

average queue size and hence the number of servers that are required in order to ensure a target quality of service.

A similar effect is displayed in figure 7, where the distribution of the operative periods is kept fixed, while the availability of the servers is reduced by increasing the average inoperative period. The figure shows the average queue sizes under exponentially and hyperexponentially distributed operative periods with the same mean. The predictions corresponding to the exponential distribution are seen to become more and more over-optimistic as the average repair time increases.

The accuracy of the geometric approximation (21) is illustrated in figure 8. The average queue size is plotted against the arrival rate, for a system with 10 servers; the other parameters are the same as in figure 5. The figure confirms the theory that as the load increases, the approximation becomes more accurate.

The last experiment demonstrates how the model and its solution can be used to answer questions of the type "what is the minimum number of servers that would ensure a certain level of performance?". In figure 9, the average response is plotted against the number of servers. The characteristics of the operative and inoperative periods are the same as in figure 5. Both the exact and the approximate solutions were evaluated. As an example of an application of such a figure, suppose that the objective was to ensure

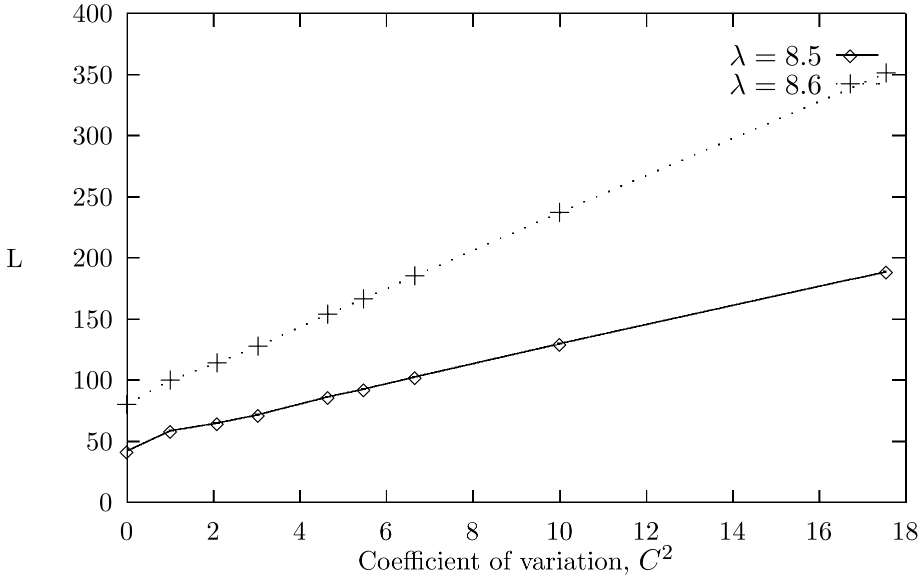

Figure 6: Average queue size against coefficient of variation

$N = 10, \eta = 0.2, \xi = 0.0289$

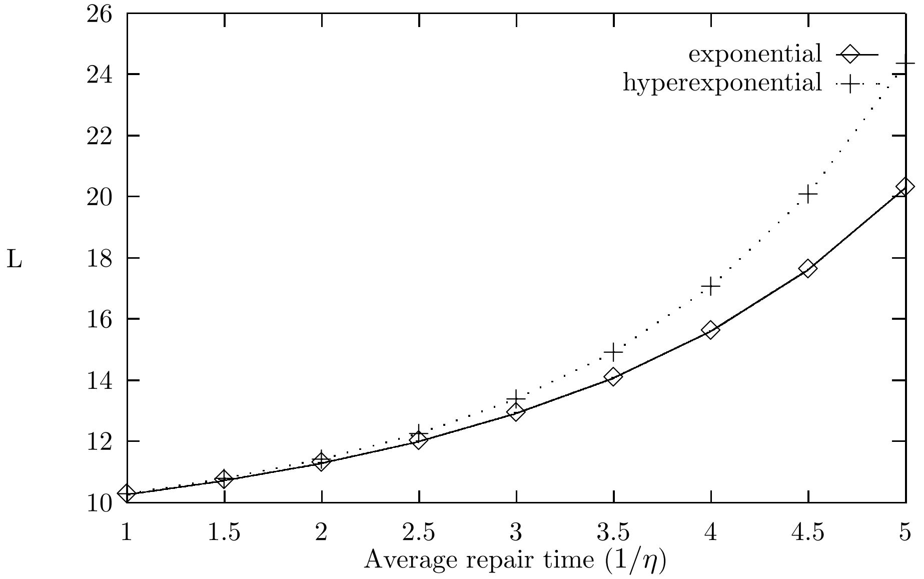

Figure 7: Average queue size against average repair time

$N = 10, \lambda = 8, \xi = 0.0289$

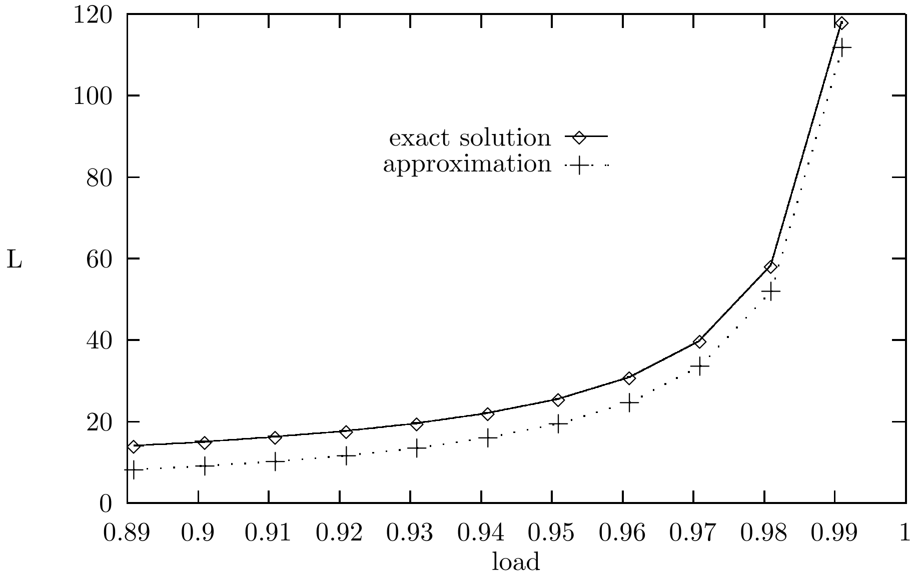

Figure 8: Exact and approximate solutions: increasing load

$N = 10, \alpha_1 = 0.7246, \xi_1 = 0.1663, \xi_2 = 0.0091, \eta = 25$

that the average response time does not exceed 1.5. The results would then indicate that at least 9 servers should be deployed. On this occasion the approximate solution underestimates the average response times; in other cases it overestimates them.

When N becomes large (greater than about 24), the exact solution begins to warn of possible numerical problems due to ill-conditioned matrices. The approximation does not display such problems.

5 Conclusions

An attempt has been made to improve the realism of models used to evaluate and optimize multi-server systems subject to breakdowns and repairs. Statistical analysis of a large volume of data concerning real servers has shown that their operative periods are not distributed exponentially, but that a good fit can be obtained with a hyperexponential distribution. The inoperative periods may reasonably be assumed to have an exponential distribution, although a hyperexponential distribution would be more accurate for them too.

A model with hyperexponentially distributed operative and inoperative periods has been formulated and solved exactly and approximately. These

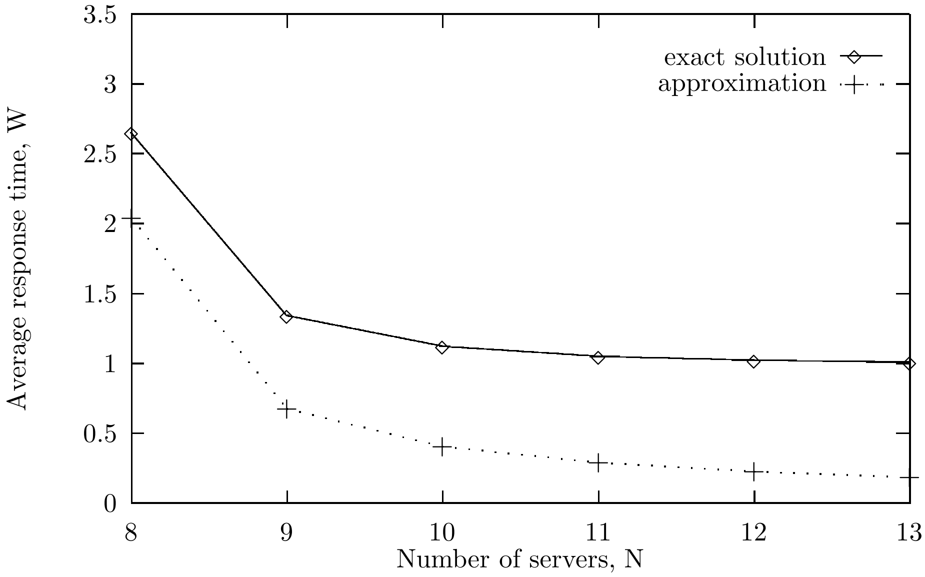

Figure 9: Average response time as a function of N

$\alpha_1 = 0.7246, \xi_1 = 0.1663, \xi_2 = 0.0091, \eta = 25, \lambda = 7.5$

solutions have been used in numerical experiments, addressing a number of cost and quality of service problems.

An open problem in this area concerns the distribution of response times. The solutions presented here can be used to determine the distribution of the queue size, hence the average queue size and the average response time. However, they do not provide the distribution (e.g., the 90% percentile) of the response time. That would be an interesting topic for future research.

Acknowledgement

This work was carried out as part of the collaborative project GridSHED (Grid Scheduling and Hosting Environment Development), funded by British Telecom and the North-East Regional e-Science center. We also acknowledge the help of Sun Microsystems, who provided the data set containing their observations of server breakdowns and repairs.

References

[1] R. Chakka and I. Mitrani, "Heterogeneous Multiprocessor Systems with Breakdowns: Performance and Optimal Repair Strategies", Theoretical

Computer Science, 125, 1, 91-109, 1994.

[2] J.-F. Dantzer, I. Mitrani, and Ph. Robert, “Large Scale and Heavy Traffic Asymptotics for Systems with Unreliable Servers”, *Queueing Systems*, **38**, 5-24, 2001.

[3] F.J. Massey Jr., “The Kolmogorov-Smirnov test of goodness of fit”, *J. Amer. Stat. Ass.*, 46, 68-78, 1951.

[4] I. Mitrani, “Approximate solutions for heavily loaded Markov-modulated queues”, *Perf. Eval.*, 62, 117-131, 2005.

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

[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**, 241-260, 1995.

[7] M.F. Neuts and D.M. Lucantoni, “A Markovian Queue with N Servers Subject to Breakdowns and Repairs”, *Management Science*, **25**, 849-861, 1979.

[8] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, 1970.

[9] B. Vinod and T. Altiok, “Approximating Unreliable Queueing Networks Under the Assumption of Exponentiality”, *J. Oper. Res. Soc.*, **37**, 309-316, 1986.