A novel approximated joint activity transition structure in a tandem feedback unreliable server queue

Dave Thornley[†], Harf Zatschler[†], Nigel Thomas[#]

[†]Department of Computing, Imperial College London, UK
{djt|hz3}@doc.ic.ac.uk
[#]Department of Computer Science, Durham University, UK
nigel.thomas@durham.ac.uk

Abstract. Unreliable servers are an important performance analysis component. In this paper we construct a novel approximation for the steady state joint solution of a tandem feedback unreliable server queue, and compare its behaviour to other approximation techniques and simulation. The queue transition structures are interesting, as they incorporate off-diagonal terms which create a MAP-like arrival process. The system is solved using spectral expansion in preference to matrix geometric methods, as this provides better stability and accuracy in finite queues. The results suggest that this novel approach of approximating joint behaviour in this manner can be of value in providing computationally inexpensive approximations to network solutions of this type.

1 Introduction

An important theme in performance analysis research is the solution for performance measures of networks. This can take place at a wide range of levels of abstraction, and here we focus on a particular issue when dealing with unreliable servers. An issue we seek to address in a largely empirical manner is the form of the departure process from an unreliable M/M/1 server, i.e. one in which the server breaks down as a Poisson process of rate ξ, and is repaired from this inactive state as a Poisson process of rate η. It is well known that the output of a reliable M/M/1 queue is *long term* Poisson, but when the processor is unreliable, we do not have this property. Networks of queues display correlations of activity which are most strongly exhibited when traffic is interrupted and/or batched. Here, we explore the significance of this effect in a novel formulation of the joint activity of the simplest possible open network. The results do not explicitly capture correlation in a manner which can be traced through a larger network, but model the effects of correlation between the two queues at steady state. The material of this paper is an in-depth analysis and extension of an important aspect of the work in [1].

2 Example queue

We solve for the steady state of a tandem unreliable server queue with feedback. There are Poisson arrivals at queue A at rate λ. The output of queue A is the sole arrival process to queue B. The output of queue B is routed back to the input of queue A at probability p. The processor in each queue is subject to Poisson breakdowns at rate ξ_X and repairs at rate η_X, where $X = A|B$ giving the identity of the server.

Arrivals at B occur only when A executes a processing completion, which requires that A be occupied. Arrivals occur at A from the external Poisson process, and from processing completions at B, which require that B be occupied. Tracking this state of occupation of each queue's "partner" is the subject of the approximation. A complete joint solution for the queues would occupy an infinite 2D lattice. We collapse one of the axes to a finite number of states by approximating the states of "occupied" and "unoccupied."

We label the states of activity and inactivity of queue X as a_X and i_X, and the states indicating that queue X is occupied and unoccupied o_X and u_X.

The solution mechanism is iterative using the following procedure:

1. Establish an initial estimate of steady state of queue B based on traffic equations.
2. Solve for the steady state of queue A using the joint state approximation and use this solution in subsequent calculations.
3. Solve for the steady state of queue B using the joint state approximation and use this solution in subsequent calculations.
4. If not converged (change in mean queue lengths less than 10^-6), repeat from step 1.

3 Initialization from traffic equations

The solution for queue B is initialized by solving for its steady state using traffic equations. If we label the link traffic from queue A to queue B as T, then the mean rate λ_T of T is found as follows:

$$\lambda_T = \lambda + \lambda_T \Rightarrow \lambda_T = \frac{\lambda}{1-p}$$

Queue B is then solved on the assumption of Poisson arrival traffic of rate λ_T. This is achieved in the normal manner for a single unreliable server queue (see *e.g.* [2]).

4 Joint state approximation for solving queue A

We use a Markov modulated queue with 8 modulation states. These track the total states of queue A, and the activity and approximation occupation states

of the "partner" queue B. These states are $i_A i_B u_B$, $i_A i_B o_B$, $a_A i_B u_B$, $a_A i_B o_B$, $i_A a_B u_B$, $i_A a_B o_B$, $a_A a_B u_B$, $a_A a_B o_B$ to be labelled 1 through 8 respectively.

We derive matrix Kolmogorov balance equations of the following form for an infinite queue A:

$$\mathbf{v}_{j-1}.(M_B + \Lambda) + \mathbf{v}_j.(Q + Q_B - D(M_B) - D(M_A) - \Lambda) + \mathbf{v}_{j+1}.(M_A), \text{ for } j > 0$$
$$\mathbf{v}_j.(Q + Q_B - \Lambda - D(M_B)) + \mathbf{v}_{j+1}.(M_A), \text{ for } j = 0$$

The vectors $\mathbf{v}_j$ are the vectors of state occupation probabilities of the modulation states at queue length j. The function $D(A)$ returns a diagonal matrix of the row sums of A. The matrix terms are as follows:

Q provides the independent modulation of the active and inactive states of queues A and B.

$$Q = \begin{pmatrix} -\Sigma & 0 & \eta_A & 0 & \eta_B & 0 & 0 & 0 \\ 0 & -\Sigma & 0 & \eta_A & 0 & \eta_B & 0 & 0 \\ \xi_A & 0 & -\Sigma & 0 & 0 & 0 & \eta_B & 0 \\ 0 & \xi_A & 0 & -\Sigma & 0 & 0 & 0 & \eta_B \\ \xi_B & 0 & 0 & 0 & -\Sigma & 0 & \eta_A & 0 \\ 0 & \xi_B & 0 & 0 & 0 & -\Sigma & 0 & \eta_A \\ 0 & 0 & \xi_B & 0 & \xi_A & 0 & -\Sigma & 0 \\ 0 & 0 & 0 & \xi_B & 0 & \xi_A & 0 & -\Sigma \end{pmatrix}$$

In the following, $\pi_{X,>1}$ is the probability that queue X contains more than one job given that it is occupied, given by the following calculated using $\pi_{X,j}$ is the marginal probability of queue length j in queue X:

$$\pi_{X,>1} = \frac{1 - \pi_{X,0} - \pi_{X,1}}{1 - \pi_{X,0}}$$

Q_B provides the horizontal transitions due to processing completions from queue B. d_B is the (approximated) rate of jobs leaving queue B when it has a single job, which are not routed to queue A:

$$Q_B = \begin{pmatrix} -\Sigma & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\Sigma & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\Sigma & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\Sigma & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\Sigma & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & d_B & -\Sigma & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\Sigma & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & d_B & -\Sigma \end{pmatrix}$$
$$d_B = (1-p)\mu_B(1 - \pi_{B,>1})$$

M_B is the matrix of transitions due to processing completions from queue B. e_B is the rate of jobs leaving queue B containing a single job, which are routed to queue A. f_B is the rate of jobs leaving queue B with more than one job, which are routed to queue A:

$$M_B = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & e_B & f_B & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & e_B & f_B \end{pmatrix}$$
$$e_B = p\mu_B(1 - \pi_{B,>1})$$
$$f_B = p\mu_B\pi_{B,>1}$$

M_A is the matrix of transitions due to processing completions in queue A.

$$M_A = \begin{pmatrix} 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&\mu_A&0&0&0&0 \\ 0&0&0&\mu_A&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&\mu_A \\ 0&0&0&0&0&0&0&\mu_A \end{pmatrix}$$

$$\Lambda = \text{diagonal}(\lambda, \lambda, \lambda, \lambda, \lambda, \lambda, \lambda, \lambda,)$$

5 Joint state approximation for solving queue B

In solving for the steady state of queue B we approximate the occupation of its partner queue A in a similar manner, and use the states $i_A i_B u_A$, $i_A i_B o_A$, $a_A i_B u_A$, $a_A i_B o_A$, $i_A a_B u_A$, $i_A a_B o_A$, $a_A a_B u_A$, $a_A a_B o_A$ to be labelled 1 through 8 respectively. Because the two queues have the same reliability characteristics (identical values of η and ξ), this gives us the same matrix Q as before in the following matrix Kolmogorov balance equations:

$$\mathbf{v}_{j-1}.(M_A) + \mathbf{v}_j.(Q + Q_\Lambda - D(M_B) - D(M_A)) + \mathbf{v}_{j+1}.(M_B), \text{ for } j > 0$$
$$\mathbf{v}_j.(Q + Q_\Lambda - D(M_A)) + \mathbf{v}_{j+1}.(M_B), \text{ for } j = 0$$

The external arrivals to queue A cause a transition from u_A to o_A.

$$Q_\Lambda = \begin{pmatrix} -\Sigma & \lambda & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\Sigma & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\Sigma & \lambda & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\Sigma & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\Sigma & \lambda & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\Sigma & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\Sigma & \lambda \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\Sigma \end{pmatrix}$$

A processing completion in queue B which is routed back at probability p will cause a transition from u_A to o_A if queue A is empty beforehand.

$$M_B = \begin{pmatrix} 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&0&g_B&h_B&0&0 \\ 0&0&0&0&0&\mu_B&0&0 \\ 0&0&0&0&0&0&g_B&h_B \\ 0&0&0&0&0&0&0&\mu_B \end{pmatrix}$$

$$g_B = (1-p)\mu_B$$
$$h_B = p\mu_B$$

Processing completions in queue A provide queue B's only arrivals, and these cause a transition from u_B to o_B.

$$M_A = \begin{pmatrix} 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&e_A&f_A&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&e_A&f_A \end{pmatrix}$$

$$e_A = \mu_A(1 - \pi_{A,>1})$$
$$f_A = \mu_A \pi_{A,>1}$$

6 Results

We compare the results of using this approximated joint activity approach to the use of two alternative approximations. One, which we label "Poisson" in the graphs, solves the queues as if they were M/M/1 with mean processing rates calculated using the breakdown and repair rates.

The other begins to take account of the possible correlations between inactivity of a source queue and the state of its target by modelling the departure process of a queue with an interrupted Poisson process transitioning between the traffic on and off states according to the breakdown and repair rates.

Figure 1 shows the results of the three approximations when there is no feedback. In this case, the solution for queue A does not generate an approximation for the state of queue B, as it does not interact with it.

Figure 2 shows the results of the three approximations when half the departures from queue B are fed back to queue A.

In general, the joint activity approximation performs best at predicting mean response time at low loads and high rates of repair.

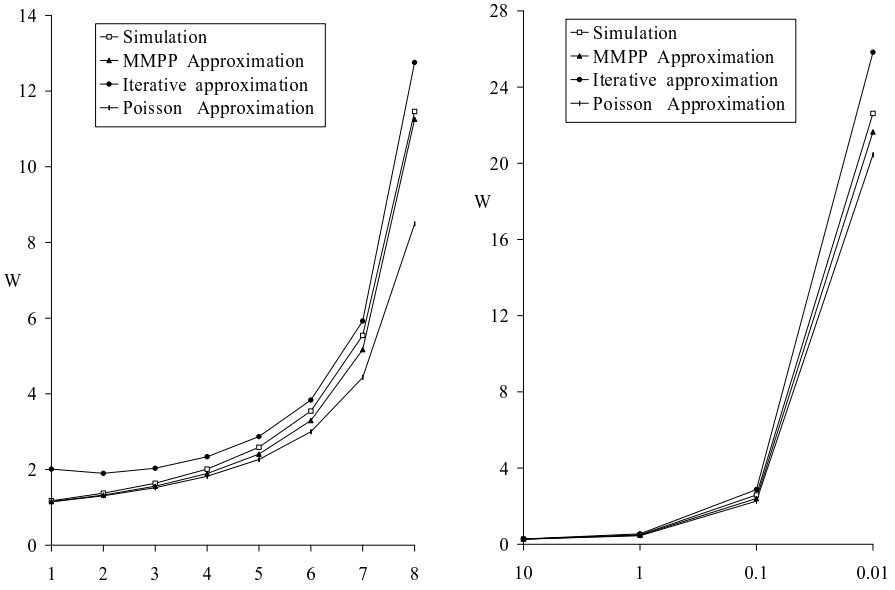

Fig. 1. Mean response time at node 2 against arrival rate λ ($\mu_i = 10$,$\eta_i = 0.1$,$\xi_i = 0.01$,$p = 0$), and against repair rate η ($\mu_i = 10$,$\lambda = 5$,$\xi_i = \eta_i/10$,$p = 0$)

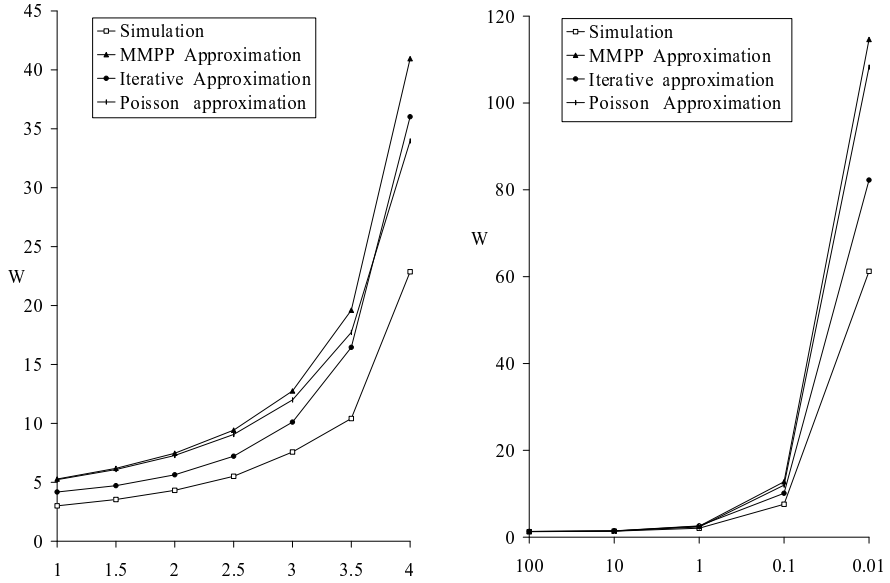

Fig. 2. Mean system response time against arrival rate λ ($\mu_i = 10$,$\eta_i = 0.1$,$\xi_i = 0.01$,$p = 0.5$), and against repair rate η ($\mu_i = 10$, $\lambda = 3$,$\xi_i = \eta_i/10$,$p = 0.5$)

7 Further work

Since the process of transitioning from the occupied region to the unoccupied state is in fact a random walk, we might expect an improved match to arise from enriching our approximation from Poisson to somewhere closer to the actual walk. As part of selecting the approximation, we might also expect it to be more important that the queue's approximated behaviour be more accurate in the more commonly occupied states. To this end, we can group the states of a queue into blocks of increasing coarseness as the queue length increases.

We base our formulation of this approximation on the measurements and definitions we already have. Beginning with queue B, which has the simplest arrival process, we know that its approximation for queue A must transition from the unoccupied state (u_A) to an occupied state (o_A) at rate λ when queue B is producing departures. The mean of this rate is calculated as the product of the processing rate μ and the probability of queue B being occupied and active, which we will call $\pi_B oa$. Since the breakdown and repair process is independent, this is easily calculated as the probability of being in the active state – which is known *a priori* – minus the probability of being unoccupied in the active state, π_{Bua} which is calculated as part of the iteration. Since the behaviour of queue B is affected by the behaviour of queue A, this means that the value of π_{Bua} is necessarily taken from a previous iteration. This also means that it must be

initialized. We have a choice of means of achieving this, of which the simplest is to take the probability of being unoccupied calculated using traffic equations.

The notation for the approximation of the states of the approximation processes must be enriched to take into account the additional states to be used. We use states a_{Ai}, where a_{A0} is identical to u_A. The state representing the first level of occupancy of queue A is labelled a_{A1}. We introduce representative transition rates between approximating states a_{Ai} and $a_{A(i+1)}$ for all $i > 0$. Arrivals cause a transition from a_{Ai} to $a_A(i+1)$ at a representative rate λ_i. Departures cause transitions from $a_{A(i+1)}$ to a_{Ai} at a representative rate μ_{i+1}.

If state a_A1 is taken to represent the presence of a single job in queue A, then $\mu_1 = \mu$. If, however we generalize this to represent queue lengths $t_0 + 1 = 1$ through t_1, this rate is set to:

$$\mu_1 = \frac{\sum_{i=2} t_1 \pi_{A,i}}{\sum_{i=1} t_1 \pi_{A,i}} \tag{1}$$

The formulation thus far corresponds to the basic approximation when u_A is identical to $a_{A,0}$, o_A is identical to a_{A1}, and $t_1 = \infty$. We introduce additional states $a_{A,2}$ through $o_{A,n}$ to split the occupied queue approximation into n regions. The transition rates between these states are to be calculated to best approximate the behaviour of the queue. As a starting point, we have measurements of the marginal queue length probabilities of queue A calculated in a previous iteration of queue A, and these can be used to set $P(a_{A,i})$ for all $i > 0$ thus:

$$\begin{aligned} P(a_A, 0) &= \pi_{A,0} \\ P(a_{A,i}) &= \sum_{j=t_{i-1}}^{t_i} \pi_{A,j} \end{aligned}$$

This then allows us to calculate appropriate values for μ_i and λ_i based on a set of Kolmogorov balance equations as follows:

$$\begin{aligned} P(a_{A,0})\lambda_0 &= P(a_{A,1})\mu_1 \\ P(a_{A,i})(\lambda_i + \mu_i) &= P(a_{A,i-1})\lambda_{i-1} + P(a_{A,i+1})\lambda_{i+1} \\ P(a_{A,n-1})\lambda_{n-1} &= P(a_{A,n})\mu_n \end{aligned}$$

There are therefore $n + 1$ balance equations, and $2n - 1$ unknowns ($\lambda_1 \ldots \lambda_{n-1}$, $\mu_1 \ldots \mu_n$). When $n = 2$, this system is exactly constrained. If $n > 2$, a constraint on the values of λ_i and μ_i will be required to complete the solution. Alternately, it might be appropriate to model the individual queue levels up to some maximum, and aggregate the levels above into a single state. This would allow the upward transition rates between the states representing single queue lengths to be set as in expression 1, and the downward rates are provided by the balance equations.

8 Conclusions

The simplest form of joint activity approximation does not perform significantly better than other simple approximations, but the addition of states to represent regions of queue lengths to improve the degree of match may improve this performance. We expect particular benefit when we move to modelling geometrically batched queues using this joint activity technique, such as those examined in the companion paper [3]. Here, information about the departure batch size distribution is of particular value, and this is strongly dictated by the queue length from which the batch is to be drawn. We expect the banded information about queue length probabilities calculated using the suggested techniques here will allow us to improve the approximation of network traffic in batched networks.

References

1. Nigel Thomas, David Thornley and Harf Zatschler: Approximate solution of a class of queueing networks with breakdowns. In proceedings European Simulation Multiconference (ESM 2003), Nottingham, 9th-11th June 2003, Society for Computer Simulation International.
2. 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.
3. David Thornley and Harf Zatschler: Analysis and enhancement of network solutions using geometrically batched traffic, in Proceedings of the Nineteenth Annual UK Performance Engineering Workshop, University of Warwick, July 9th-10th 2003.