Analysis and enhancement of network solutions using geometrically batched traffic

David Thornley, Harf Zatschler

Department of Computing
Imperial College
Huxley Building, 180 Queen's Gate
London SW7 2RH
UK
{djt,hz3}@doc.ic.ac.uk

Abstract. Networks of GE/GE queues provide bursty traffic, and are motivated in work by Kouvatsos *et al.* We provide an improved approximation method using processes defined on *bands* of queue lengths that can be employed within systems that exhibit significant correlation. These bands allow construction of a piece-wise queue length dependent arrival process, which improves the match between the approximation and the true traffic and above the use of a single geometrically batched process. This adds to the range of methods for approximation in geometrically batched networks, thus broadening their applicability.

1 Introduction

Burstiness can be modelled by various means, the conceptually simplest being with the use of either modulated arrivals with contrasting rates, or explicitly with batch transitions. In the latter method, we can either use deterministically selected batch sizes of finite or infinite range. One particular batch size distribution - the geometric distribution - is motivated by a maximum entropy formulation for the behaviour of a G/G/1 queue as a form for constraining a moment matching procedure. This form of traffic is conceptually validated by its ability to produce a range of batch sizes in which larger batches are less likely. The particular distribution has characteristics which enable efficient solutions [7] for exact performance measures.

In previous work [4] we have approached the use of geometric batches from a different angle to maximum entropy [1] and [2], explicitly solving queues allowing multiple independently batched arrival processes.

Our approach is enabled by the particular form of the geometric distribution which allows manipulation of the Kolmogorov balance equations to provide an equivalent ensemble of constraints which are soluble using spectral expansion [3] or matrix geometric techniques [5, 6]. An early form of the solution methods is given in [7], which solves exactly using spectral expansion, but does not provide multiple arrival streams. A later method, which include multiple arrival streams, enabled initial experiments [4] on open networks seeking to enhance the accuracy

of approximated solutions. Using a join in contrast to the aggregation techniques underlying other work shows some improvements in accuracy.

Work undertaken by Kouvatsos provides examples of the behaviour of performance measures resulting from the use of techniques motivated by maximum entropy analyses. We explain ours from the particular viewpoint of explicit solution of system specific balance equations for the queues. It is noted in *e.g.* [2] that many performance measures resulting from the use of geometric batches are pessimistic. Performance measures are strongly dependent on the detail of the batch size distribution, leading to significant errors in predictions when using approximation techniques.

In this paper, we look at an example of queue length correlated arrival processes. This motivates an examination of the stark contrast between the arrival process specification of a GE/GE/1/L queue and the traffic it has to represent.

2 Closed networks

A particular example where batching is problematic is the closed network, or an open network with finite capacity. Kouvatsos commonly includes blocking in his approximations as a well-justified means for simplifying some of the issues. Using our specific form of solution, we examine the detail of the batch size distributions to find a more accurate approximation method.

Since in a closed network there is a constant number of jobs, queue behaviour is strongly correlated through the network when distributed batch sizes are used. We choose to approximate the queue length dependent arrival process by setting the rate and batch size parameters of the arrival process to values specific to bands of queue length.

The traffic in a closed network is subject to a constraint that the number of jobs in the system remains constant, and finite, whereas the geometric distribution used to provide batch sizes for our queues is unbounded. When using a network solution method based on the steady state of queues, as we and Kouvatsos and others do, this can lead to the apparent departure of a batch of jobs larger than the prescribed content of the whole network, and this oversized batch can make its way around along a complete path in an open network, and resonate around a closed network. We therefore follow the convention that batch sizes are truncated to the largest feasible batch size, should this occur.

3 A simple example network

We have simulated a simple closed network of three infinite queues in a feedback loop, each with the same processing characteristics: a Poisson process of rate 1 generates departures of batches of geometrically distributed random size S chosen on the geometric distribution, such that $P(S = s) = (1 - \phi)\phi^{s-1}$ with the parameter ϕ initially set to 0.9 to give large batch sizes to emphasize the effect. Each queue receives traffic from one other queue with precisely the same characteristics, and sends its output to one other queue. There is a total of 60

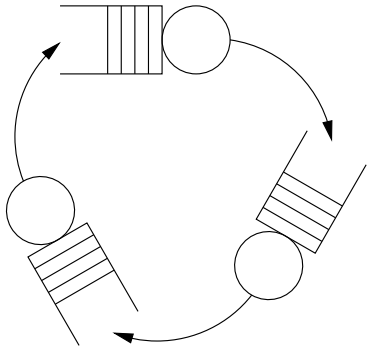

Fig. 1. example feedback network

customers in the network. It is clear that the network under consideration is entirely symmetric so that we limit the investigation to one particular queue of the three as the other's behaviour at steady state is identical.

This example is the simplest network to require approximate solution methods, and serves as a basis for discussion of issues in the use of geometric batches.

3.1 Queue length dependent arrival processes

In order to provide an initial overview of the involved traffic characteristics, a simulation of the above network was performed to 2×10^9 departure events. In addition to recording queue length distributions, the simulation was instrumented to record the inter-arrival distributions and the sizes of batches arriving at a queue for each given queue length of the target queue.

From the nature of the feedback network, we can expect a large amount of correlation to exist between the queue lengths of the three queues. The more customers one particular queue contains, the fewer are left over for the others in this closed system. As a consequence, the arrival process at the queue is effectively throttled at large queue lengths, and reaches zero when the queue contains all 60 customers. The correlated behaviour therefore gives rise to a queue length dependent arrival process, illustrated in Figure 2.

At low queue lengths, the inter-arrival delay and arrival batch distributions behave as would be expected from an exponential rate and geometric batch size. With increasing queue length, the inter-arrival distribution flattens and develops a local maximum, whereas the batch distributions become visibly truncated at the batch size that would cause all customers within the system to be within this queue.

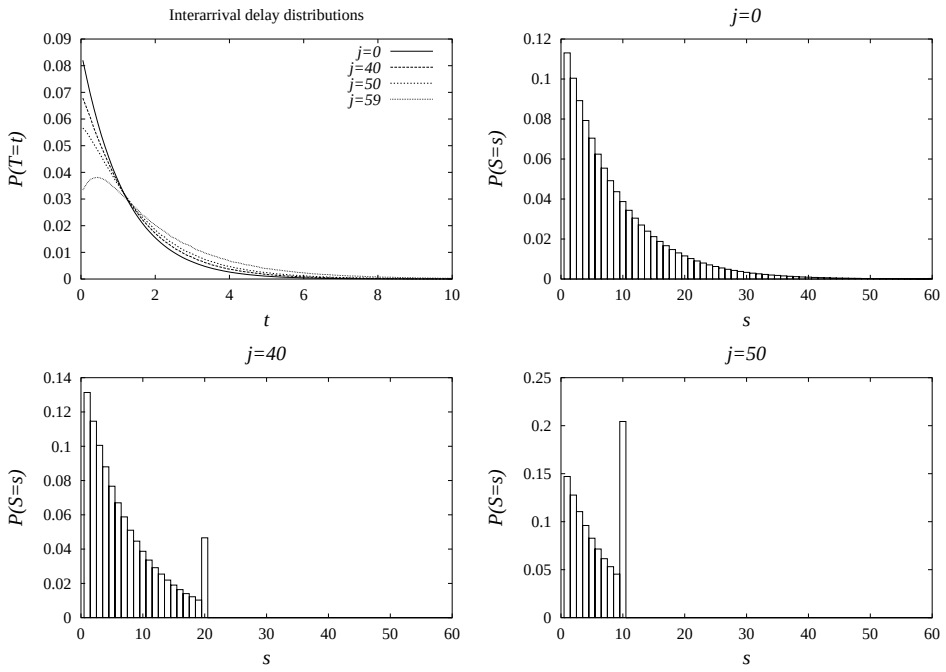

Fig. 2. Inter-arrival and Batch Distributions at selected queue lengths

3.2 Matching Methodology

We aim to match the actual arrival distributions as closely as possible using exponential and geometric distributions, so that the queue length distributions can be derived without the need for prior simulation. We chose these particular distribution types due to familiarity and easy availability of exact steady-state solvers derived from [4] as well as due to maximum entropy considerations. For the matching process, we explicitly recast each individual queue to be finite of length 60. This change causes otherwise unbounded geometric batches to be truncated at the same length as those of the actual arrival batch process shown in Figure 2 and at the same time forcing an effective arrival rate of zero at $j = 60$. We do not lose any state space detail due to this transformation as any states for which $j \geq 61$ are unreachable.

The marginal distributions for different j are too dissimilar to be matched individually, so that our current strategy centers on achieving approximations to the first moments – shown in Figure 3 – of the inter-arrival and batch distributions for a range of j.

When matching the arrival process by means of a unique, global exponential rate together with a geometric batch parameter, we obtain a result which essen-

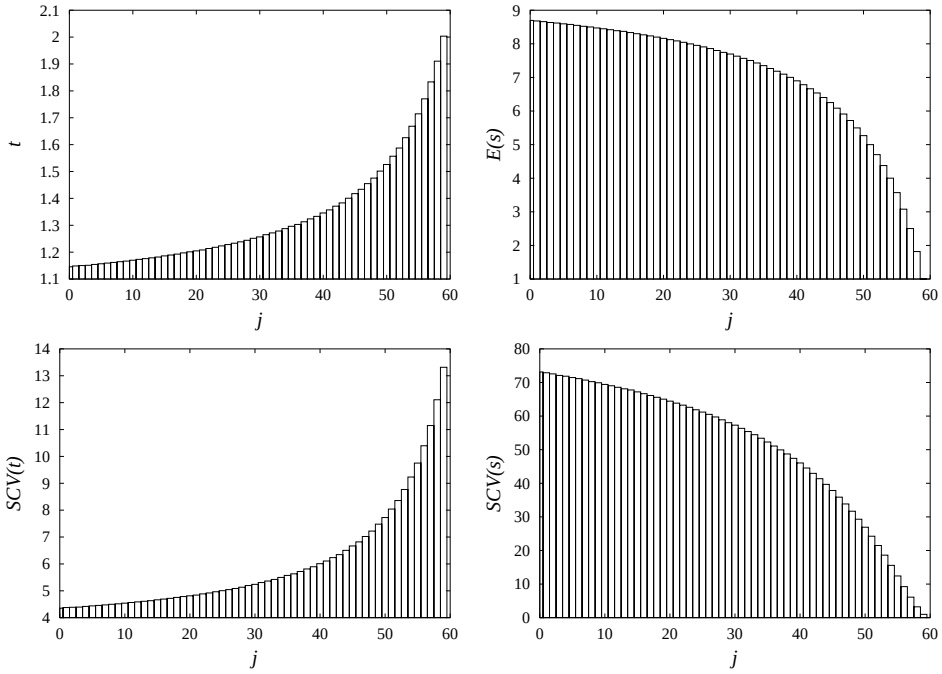

Fig. 3. Mean and SCV of Inter-arrival and Batch Distributions

tially follows that set by maximum entropy methods [1]. The use of a global rate throughout the queue yields a rather crude approximation of the arrival process, which we aim to improve upon by allowing for certain ranges of queue lengths to have differing arrival process characteristics. To this end, we are developing a methodology to split a queue into *bands* of queue lengths, within which arrival streams[1] have different rate and batch parameters. Below, we present the use of a global rate and that of two bands, $B_1 = \{0, \ldots, 30\}$ and $B_2 = \{31, \ldots, 60\}$ by giving the approximated arrival processes and contrast the derived solutions.

3.3 Approximated Arrival Process

The calculated global arrival rates shown in Figure 4 are necessarily a crude approximation to the actual values taken, with a noticably better fit in the two-band case. Batch distributions are significantly easier to fit for this problem, but there is still systematic underestimation of batch sizes for $j < 20$ as well as overestimation for $j > 20$ using global rates. The use of two bands improves the fit as before.

[1] this can also apply to service and negative customer streams, if so desired

The closer fit of the arrival process that can be achieved with bands also gives a closer match for the queue length distribution. This can bee seen in Figure 5, where the probability at an individual queue length is almost always more accurate for the two band case.

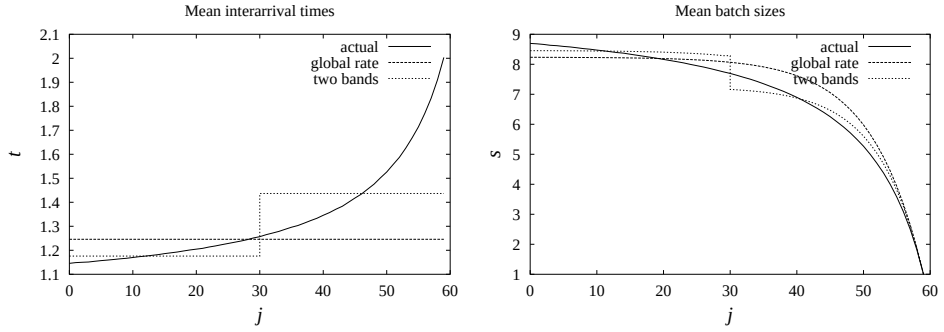

Fig. 4. Actual and matched arrival processes for $\phi = 0.9$

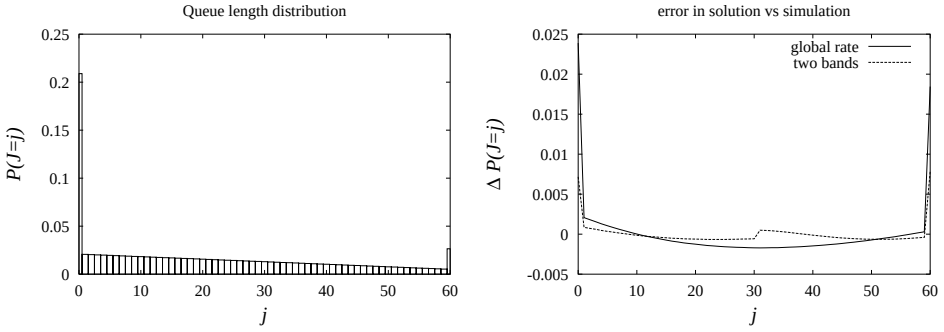

Fig. 5. Queue length distribution and errors for $\phi = 0.9$

4 Fixed point and perturbation

The preceeding analysis tells us that approximate solutions for networks of GE/GE/1 queues are necessarily rather coarse, and we would like to be able

to modify the parameterization in a systematic manner, not relying on prior knowledge of the solution, to improve the fit. Since we cannot do this in the general case without simulation of the network, we propose an approach which we hope will be applicable to network design and optimization.

For a given network geometry and paramaterization (processing rates, routing probabilities) we can simulate the network (expensively) to find its steady state, and from this, measure throughput. We can then model this network using GE/GE queues, with rates and batching carefully selected to accurately match the steady state. This in practice means having a Poisson rate and batch distribution parameter which give the correct mean arrival rate, but with the two parameters selected to create a good match in, for example, the mean queue length of the target queue, in order to measure mean sojourn time correctly.

To perform the search for the optimum, we need to vary the the parameterization of the real network. To assess the performance of this network, the model parameters also need to be changed to track this. The GE arrival process is a strict approximation - it very rarely accurately reflects features of the traffic in relation to each other (see discussion of apparent batching distributions earlier). Thus, we need to pick a locus for the queue parameters to follow for a given change in behaviour. In the above example where we wish to match mean sojourn time, we could chose, for example, to scale the modelled Poisson rate while keeping the batch parameter constant to track the mean traffic rate. Further research may reveal more sophisticated loci to achieve a better result.

The further the search strays from the simulated parameterization (which we take to be an accurate reflection of the real network's behaviour), the more approximate will be our result. We might suggest this has a metaphor in the Taylor expansion with a limited number of terms. The larger the deviation from the base case, the larger the errors. To counter this, if we believe the solution we seek is too far from the simulated case, we seed the search space with further simulations from which the perturbation search can be carried out. In this manner, we can balance efficiency against accuracy of the solution.

As a first step, we will perturb our network by varying the service batch parameter ϕ common to all three queues in the system to investigate what the effect of batched services has on the optimal matched solution using maximum entropy (global rate) and two bands. The two graphs in Figure 6 show a surprisingly strong relationship between ϕ and the matched parameters.

Within the considered range, $\phi = 0 \ldots 0.9$, the matched batch parameters $\hat{b}$ are proportional to and strictly smaller than ϕ. Among the rates for the same range of ϕ, there exists a linear relationship between ϕ and $\frac{1}{\mu - \hat{r}_\phi}$, where $\hat{r}_\phi$ is a matched rate and $\mu = 1$ is the departure rate of the queues. Taking advantage of these simple linear dependencies, it is now possible to achieve the optimal matched parameters for either global rate or two band approximated arrival processes by way of simple linear interpolation between two known data points. Allowing for numerical inaccuracies, the parameters resulting from interpolation are identical to those obtained through simulation and subsequent matching.

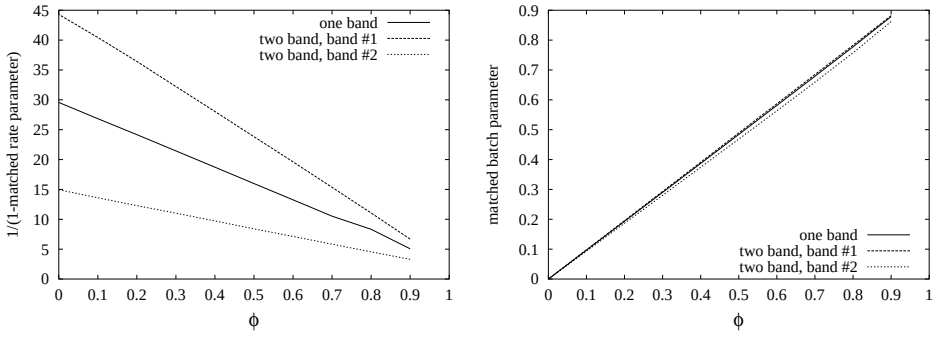

Fig. 6. Relationship between ϕ and matched parameters

Figure 7 shows the arrival process and solution for $\phi = 0.5$. Matched parameters were derived by linear interpolation between data points at $\phi = 0.25$ and $\phi = 0.6$

5 Further Work

In the above approximation, we have chosen the bands $B_1 = \{0, \ldots, 30\}$ and $B_2 = \{31, \ldots, 60\}$ somewhat arbitrarily. In general, it is not necessarily sufficient to use equally-sized bands, since the means (as in Figure 3) may vary significantly more within one region than another. An improved fit may be achieved by extending B_2, over a smaller region at the top of the queue, where the arrival rates vary most. A method to derive optimal band sizes will be developed in future work.

The use three of more bands allows for better approximations still, but is counterweighted by the additional computational expense incurred by each additional band. Every band necessitates the use of a separate Spectral Expansion (or Matrix Geometric) solution pass, so that it can become more efficient to explicitly solve the entire Markov chain when too many bands are present.

In [4] we allowed for arrivals from two distinct sources to be explicitly superimposed – a technique that was not applied here as every queue had exactly one input. We will develop a methodology to facilitate both superimposed and banded streams within the same framework.

6 Conclusions

Geometric batches give us a range of batch sizes, and behaviours which are susceptible to a range of solution techniques. The accuracy of models using this feature is largely to be judged by simulation, as any approximate analytic solutions lose information about correlation, blocking and instantaneous population

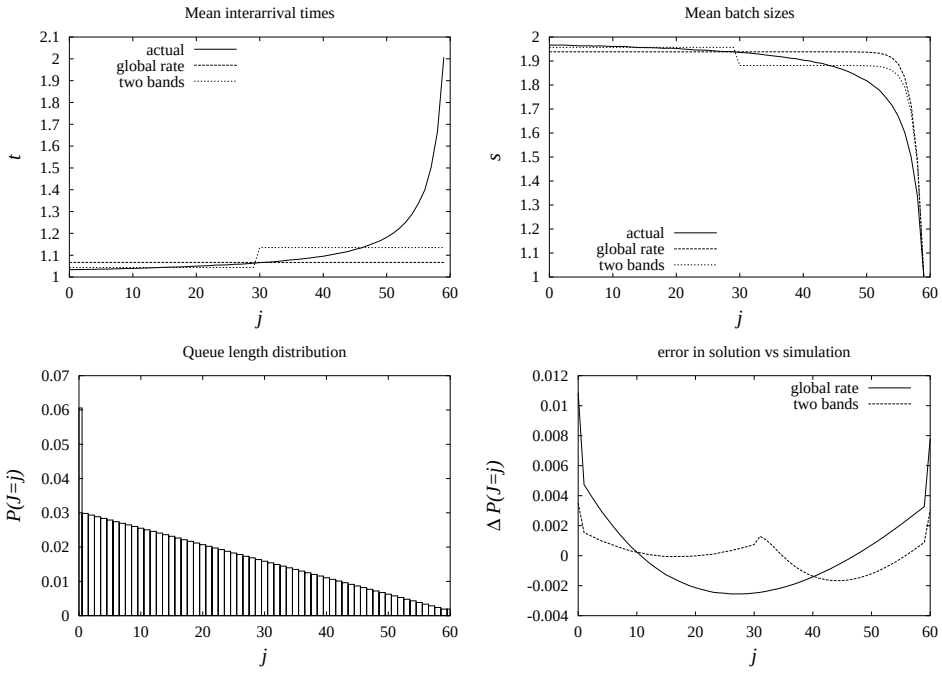

Fig. 7. Arrival processes, queue length distribution and error for the network with $\phi = 0.5$

levels. As with any other methods, the best we can do is identify classes of problem to which the particular approximations are most applicable, or produce methods which actively acknowledge errors, or are always pessimistic. We have proposed a method which has built in checkpoints for accuracy. Inaccuracies in using geometric batches in network solutions generally take the form of over-estimation of queue lengths, and hence exaggerated sojourn times. But the mis-estimation is quite well behaved when dealing with queues in isolation, so we might sensibly hope for systematic means for modifying the methods to account for this, and hence improve network solutions in further research.

References

1. Kouvatsos D, Tabetaouel N: A maximum-entropy priority approximation for a stable G/G/1 queue. ACTA Informatica 27 (3): 247-286 1989
2. Kouvatsos D, Awan I: Entropy maximisation and open queueing networks with priorities and blocking. Performance Evaluation 51 (2003) 191-227.

3. 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.
4. P.G. Harrison, D.J. Thornley, H. Zatschler: Geometrically batched networks. In proceedings (ISCIS 17) Seventeenth International Symposium On Computer and Information Sciences October 28-30, 2002 University of Central Florida Orlando, Florida
5. M.F. Neuts: Matrix Geometric Solutions in Stochastic Models. Johns Hopkins University Press, Baltimore, 1981.
6. G. Latouche, V. Ramaswami: A logarithmic reduction algorithm for quasi-birth-death processes. Journal of Applied Probability (1993)
7. Ram Chakka, Peter G. Harrison: A Markov modulated multi-server queue with negative customers - The MM CPP/GE/c/L G-queue. Acta Informatica 37(11-12): 881-919 (2001)