

Versatile Models of Systems Using MAP Queueing Networks

Giuliano Casale, Ningfang Mi, and Evgenia Smirni

College of William and Mary
Department of Computer Science
Williamsburg, VA
{ casale, ningfang, esmirni }@cs.wm.edu

Abstract

Analyzing the performance impact of temporal dependent workloads on hardware and software systems is a challenging task that yet must be addressed to enhance performance of real applications. For instance, existing matrix-analytic queueing models can capture temporal dependence only in systems that can be described by one or two queues, but the capacity planning of real multi-tier architectures requires larger models with arbitrary topology.

To address the lack of a proper modeling technique for systems subject to temporal dependent workloads, we introduce a class of closed queueing networks where service times can have non-exponential distribution and accurately approximate temporal dependent features such as short or long range dependence. We describe these service processes using Markovian Arrival Processes (MAPs), which include the popular Markov-Modulated Poisson Processes (MMPPs) as special cases. Using a linear programming approach, we obtain for MAP closed networks tight upper and lower bounds for arbitrary performance indexes (e.g., throughput, response time, utilization). Numerical experiments indicate that our bounds achieve a mean accuracy error of 2% and promote our modeling approach for the accurate performance analysis of real multi-tier architectures.

1 Introduction

Capacity planning of modern computer systems requires to account for temporal dependent features in workloads, such as short-range or long-range temporal dependence that

create burstiness among consecutive requests. Recent measurements in real systems show that temporal dependent processes can be prevalent in a variety of different settings, including multi-tier architectures and disk drives [4, 6]. However, there is currently a lack of understanding on how to quantify performance degradation due to temporal dependence and to counteract its negative performance effects.

Consider, for example, multi-tiered systems, a prevalent architecture of today's Web sites. We observed that burstiness in the service process of *any* of the tiers may result in very high user response times even if the bottleneck resource in the system is not highly utilized, while measured throughput and utilizations of *all* other resources are also modest [4]. When burstiness is not considered, this underutilization may falsely indicate that the system can sustain higher capacities and mislead the capacity management.

In collaboration with researchers at Seagate Research we built an e-commerce server according to the TPC-W e-commerce benchmark to identify the presence of temporal dependence in different tiers of the system. A high-level overview of the experimental set-up is illustrated in Figure 1, which also shows the flow of requests. TPC-W defines think times of clients to be exponentially distributed, implying that *there is no temporal dependence due to the generation of client requests in the system*. Burstiness, as characterized by the autocorrelation function, is nonetheless observed in various flows in the system as shown in the right graph of Figure 1. According to our analysis, the origin of these bursty flows is the service process in the front server and is an effect of caching/memory pressure (details can be found in [4]). Furthermore, because of the closed-loop nature of the system, burstiness propagates in the flows of the *entire* system, severely affecting end-to-end client response times. In [4] we built closed queueing network models to understand the observed behavior of the TPC-W experiments. Figure 2 presents the queueing network that captures the TPC-W flow of requests in the multi-tiered system. Figure 3 presents average performance measures obtained

This research was funded by the National Science Foundation under grants ITR-0428330 and CNS-0720699.

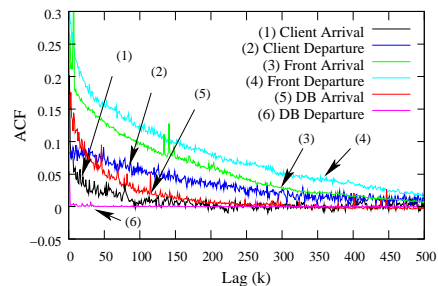
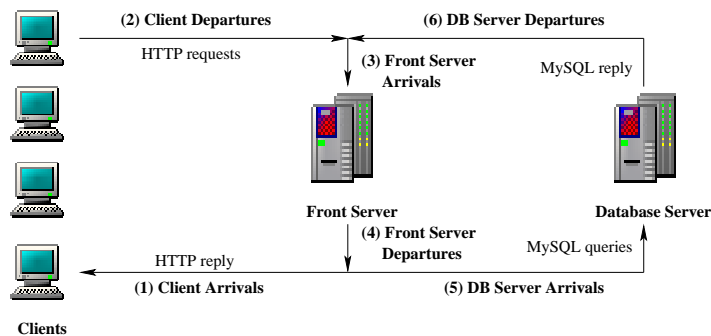


Figure 1. TPC-W experimental environment (left) and autocorrelation flows in various marked points of the system under the the default browsing mix with 384 emulated browsers (right).

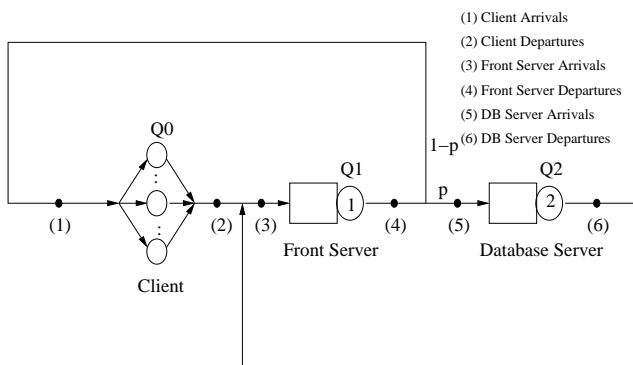


Figure 2. A simple queueing model of TPC-W.

from two different parameterizations of the models of Figure 2 and, to facilitate comparison, also the corresponding measurements from our TPC-W testbed. The first model parameterization that explicitly captures temporal dependence (in terms of autocorrelation) in the front server is shown on the first row of bargraphs. Model and measurement results are in excellent agreement. The agreement between model and measurements diminishes quickly if the same model parameterization uses uncorrelated processes throughout the closed system, see second row of bargraphs. That is, ignoring temporal dependence results in severely underestimated response times and queue lengths as well as overestimated server utilizations at all tiers.

Starting from these results, we are now investigating new models and characterization techniques to account for burstiness in the performance evaluation of multi-tier architectures. Capacity planning based on product-form queueing networks has been extensively used in the past, since these models enjoy simple solution formulas and low computational cost of approximation algorithms [3]. However, modern Web, parallel, and storage systems often exhibit high variability in their service processes and are therefore best modeled by networks of queues with general independent (GI) service [3]. Nevertheless, although much more

accurate than product-form networks, solution techniques developed for models with GI service are insufficient for robust performance predictions if the service process is autocorrelated. An illustrative example is shown in Figure 4, which illustrates the inaccuracy of basic Markov chain decomposition techniques [2], commonly used for the evaluation of non-product-form networks, when applied to autocorrelated models. The figure plots the utilization at queue 1 for a basic network with two queues in tandem as the number of jobs in the network grows. Decomposition shows unacceptable inaccuracies as soon as the number of processed requests N increases beyond a few tens. Similarly, the general ABA bounds [3], also shown in the figure, cannot approximate performance well, except at very low or very high utilization.

We overcome these limitations of existing approximation techniques by providing a bound analysis methodology for queueing networks with autocorrelated workloads. Because of the complexity of their analysis, only small autocorrelated models based on one or two queues have been considered in the literature, mostly in matrix analytic methods research; therefore, models such as the one in Figure 2 cannot be studied analytically with existing techniques. We instead define and study a class of closed queueing networks where service times are modeled by Markovian Arrival Processes (MAPs), a family of point processes which can easily model general distributions and temporal dependent features such as burstiness in service times [5], and that admits an accurate bound analysis. Our bounds derive from the analysis of the Markov process underlying the MAP queueing network. Specifically, we use a new linear programming approach that remains computationally efficient also on models with large populations and large number of servers.

This document is organized as follows. We overview our bounding approach in Section 2 introducing the new concept of marginal balances. In Section 3 we give numerical evidence that our bounds can effectively characterize the performance of autocorrelated queueing networks. Finally, Section 4 concludes the paper and outlines future work.

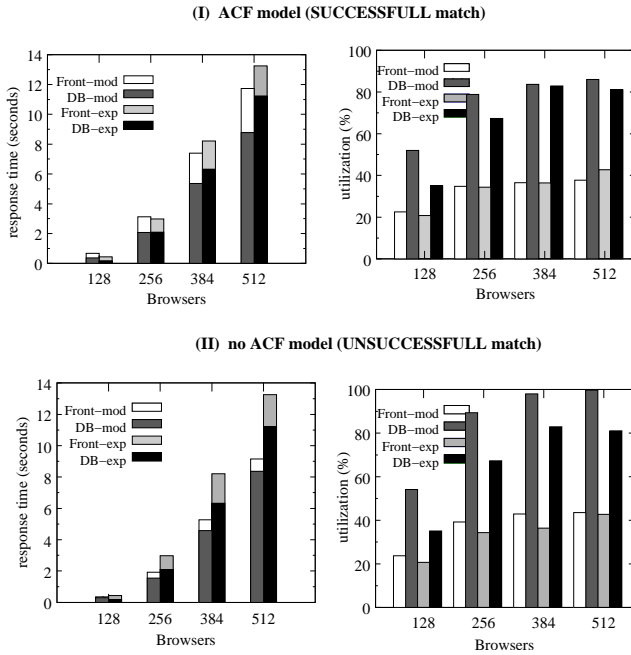


Figure 3. Performance measures for the model in Figure 2: request response time and server utilization. Results are presented for two queueing models: one that captures autocorrelation in the service processes for the front server (first row, successful match) and one with no autocorrelation in the service process of the front server (second row, unsuccessful match). Classic models that are used for capacity planning do not consider autocorrelation and would result in an unsuccessful match. To facilitate comparison, experimental results are also presented (labeled “exp”).

2 MAP Queueing Networks

We illustrate our bounding approach using the example model shown in Figure 5, however we remark that the discussion in this section readily generalizes to models with larger number of queues or different topology. The model in Figure 5 represents two application servers processing requests incoming from a shared communication link modeled by queue 1. Queue 1 and queue 2 have exponential service times; queue 3 has instead MAP service, thus we can use here non-exponential service time distributions, e.g., hyperexponential, and temporal dependent features, e.g., short-range dependence. Figure 6 shows the underlying Markov process of the MAP network in Figure 5 with routing probabilities $p_{1,1}, p_{1,2}, p_{1,3} = 1 - p_{1,1} - p_{1,2}$ at the first queue and $p_{2,1} = 1, p_{3,1} = 1$, at the remaining queues. For simplicity of illustration, the MAP service is specified by a Markov process composed by two states (phases). This means that while the Markov process is in state 1 we assume that exponentially-distributed service is offered with mean rate λ_1 ; instead, in state 2 the service times are exponentially-distributed with a different mean

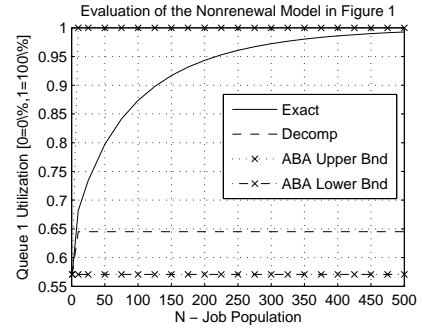


Figure 4. Exact global balance solution of a two-queue closed network with the ABA bounds [3] and the decomposition-aggregation approximation [2].

rate $\lambda_2 \neq \lambda_1$. This state-space based description of service times allows to define non-exponential distributions, e.g., hyperexponential, and by properly selecting the frequency of jump between the two states it is also possible to modulate the temporal dependent properties of the service times to approximate short or long range dependence. In Figure 5, grey is used for states where the MAP is in phase 1; in white states the MAP is in phase 2; additional notation is described in the caption.

The basic idea of our bounding approach is as follows. We consider a queue k , $1 \leq k \leq 3$, and the group of states where its queue-length is equal to a given n , $1 \leq n \leq N$. We then determine separating cuts, henceforth called *marginal cuts*, that isolate this group of states from the rest of the graph. Drawing all cuts for all possible choices of the queue $k = 1, 2, 3$ and of the queue-length size $n = 0, 1, 2$ we obtain the grid of dashed lines shown in Figure 7.

We make the crucial observation that the underlying global balance equations of the Markov process can be aggregated to describe *only* the probability fluxes across the marginal cuts of the grid. Although conceptually simple, the observation is striking in that it opens the way to efficiently estimate performance indexes by focusing only on simpler aggregate quantities and without the need of evaluating individually each state of the Markov process. To the best of our knowledge, this is the first time that an exact aggregation technique is proposed for arbitrary-size non-product-form networks. For the example in Figure 6, we may formulate a marginal cut balance for queue 2 as

$$p_{1,2}\mu_1\pi(n_1 \geq 1, n_2) = \mu_2\pi(n_2 \geq 1, n_2 + 1), \quad (1)$$

for $1 \leq n_2 \leq N - 1$, where the marginal probability $\pi(n_j \geq 1, n_k)$ is the probability that queue j is busy while queue k has n_k enqueued jobs. More general formulas accounting for possible phase changes in the MAP queues

can be similarly derived, leading to the derivation of thirteen distinct types of marginal balance equations. The main properties of this result are twofold:

Computational tractability. Overall, the number of probability terms appearing in marginal cut balances similar to (1) for a model with N jobs and M queues is $M^2(N + 1)$, which scales efficiently with the model size. This is a fundamental result, because the number of terms in the global balance equations grows as

$$\binom{M + N - 1}{N},$$

which explodes combinatorially as the number of queues or jobs in the model grows. In comparison, the marginal cut balance description is always computationally feasible.

Exactness. The bounding method derives from the exactness of marginal cut balances. Let $\mathbf{A}\vec{\pi} = \mathbf{b}$ be the set of all possible marginal cut balance equations for the queueing network model under study, and define $f(\vec{\pi})$ as a linear combination of probabilities which represents a performance metric of interest. Useful quantities that can be computed in terms of a linear function $f(\vec{\pi})$ are, e.g., utilizations, throughputs, or mean, variance and higher moments of queue-lengths; we also show later how to compute response times. We determine bounds on these performance indexes by computing a bound on $f(\vec{\pi})$ using a simple linear program in the form

$$f_{min} = \min f(\vec{\pi}) \text{ subject to } \mathbf{A}\vec{\pi} = \mathbf{b}, \vec{\pi} \geq 0,$$

for lower bounds or

$$f_{max} = \max f(\vec{\pi}) \text{ subject to } \mathbf{A}\vec{\pi} = \mathbf{b}, \vec{\pi} \geq 0,$$

for upper bounds. The computational costs of linear programs of marginal cut balances are very good for practical applications, e.g., we have solved the linear program for a model with 10 MAP(2) queues and $N = 50$ jobs using an interior point solver in approximately four minutes; for $N = 100$ the solution of the same model is found in approximately ten minutes suggesting very good scalability in the population size. Examples of the linear programming bounds described before are given in the next section.

3 Accuracy Evaluation and Examples

We assess the accuracy of the proposed bounds using the following methodology. We use both randomly-generated models and representative case studies. In the random models, we evaluate bound maximal relative error with respect to the exact solution of the MAP network computed by global balance. Due to the state space explosion, experiments using exact global balance solutions are prohibitive

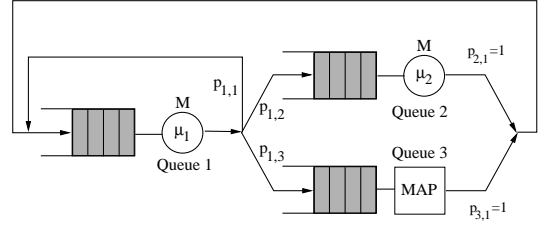


Figure 5. Example network composed by two exponential queues and a MAP queue.

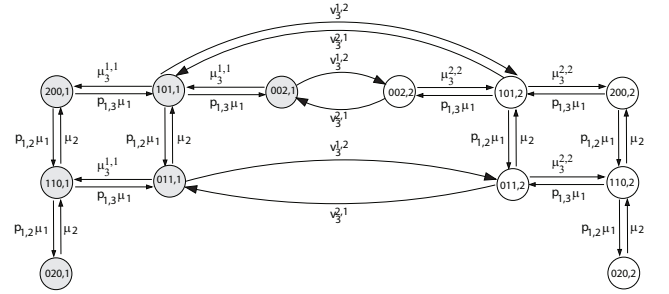


Figure 6. Underlying Markov process of the network in Figure 5 in the simple case when the MAP is a MMPP(2) process; the job population is $N = 2$. The first two queues are exponential with rates μ_1 and μ_2 , respectively; the third queue is a MAP with two phases where, e.g., $v_3^{1,2}$ and $\mu_3^{1,2}$ are the probabilities of moving from phase 1 to phase 2 without job completions, respectively. (002, 1) indicates that the exponential queues are idle and the MAP queue has two jobs and is in phase 1; in (110, 2), the phase 2 is the phase left active by the last served job.

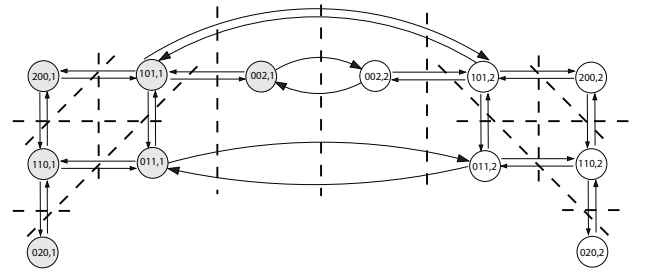


Figure 7. State space grid defined by marginal cuts.

for MAP networks with more than three queues and population $N \geq 100$. Therefore to fully explore the accuracy of the proposed bounds compared to the exact solution, we focus on models with three queues. Mean, coefficient of variation, skewness, and autocorrelation geometric decay rate at MAP(2) servers are also drawn randomly. For each model, we compute upper and lower limits X_{max} and X_{min} on the mean throughput $f(\vec{\pi}) = X$. Then, using Little's Law we get the response time bounds $R_{min} = N/X_{max}$

	M	Maximal Relative Error			
		mean	std dev	median	max
R_{max}	3	0.013	0.021	0.004	0.141
R_{min}	3	0.022	0.020	0.019	0.126

Table 1. Results of Random Experiments

and $R_{max} = N/X_{min}$ which are used to compute absolute relative errors from the exact response time R . We do not report errors on other measures due to lack of space, but we remark that they are in the same range of response time errors.

3.1 Random Models

We evaluate on 10000 random models the maximal relative error with respect to the exact response time over all populations $1 \leq N \leq 100$. Table 1 indicates that the proposed bounds perform extremely well for all models. The mean error is 1 – 2% for both bounds with a standard deviation of 0.02; the median is less than the mean, indicating that the asymmetry of the error distribution is more concentrated on small errors. The maximum error is found to be 14.2% for the upper response time bound and 12.6% for the lower bound. We have inspected carefully these cases and found that models with more than 10% error in at least one of the two bounds account for only the 1% of the total number of experiments.

3.2 Case Studies

We also illustrate the accuracy of the proposed bounds on the model shown in Figure 5 with routing probabilities $p_{1,1} = 0.2$, $p_{1,2} = 0.7$, $p_{1,1} = 0.1$. The MAP queue 3 has $CV = 4$ and geometric autocorrelation decay-rate $\gamma_2 = 0.5$. Figure 8 shows the utilization and response time bounds of queue 3 as a function of the number of requests in the system. The bounds of both utilization and response times are very close to the exact value on most populations. Both bounds converge to the asymptotic exact, a feature that is not always found in bounds for queueing networks. For different values of the routing probabilities the results are even tighter than in Figure 8.

4 Conclusions

Autocorrelated service processes are often found in workloads of storage systems and Web servers [4, 6], but existing queueing network models are unable to correctly model the performance degradation due to temporal dependence. We have found a solution to this problem by studying a new class of MAP closed networks that supports autocorrelated service. Experimental results indicate that our

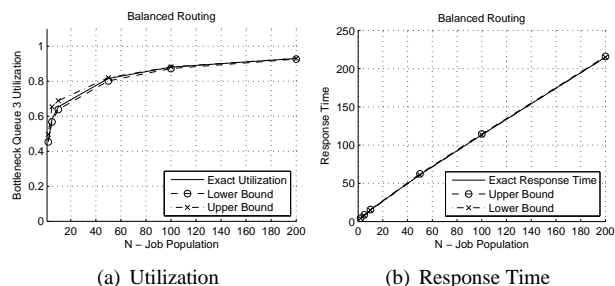


Figure 8. Case Study Results.

bounds are extremely accurate, showing on average a 2% relative accuracy error on the response time and therefore can provide robust estimate of the performance of real systems.

Starting from these results future work will focus on defining dynamic resource allocation policies that strive to minimize request round-trip times under temporal dependent workloads. This can be done both at the system-level by exploring in real time (e.g., with the proposed bounds) alternative network configurations that lead to improved performance or at component-level by smart scheduling disciplines that take advantage of the temporal dependence properties of the workload locally. Finally, a fundamental research to be carried out is the parameterization of MAP service processes from measurements. There is currently a lack of methods for fitting higher-order properties of the service times, such as skewness or higher-order spectra, which can be relevant in the accurate characterization of system performance. Our preliminary results indicate that queueing models with MAPs parameterized up to third-order statistical properties can be several orders of magnitude more accurate in prediction accuracy than standard second-order parameterizations [1].

References

- [1] G. Casale, E.Z. Zhang and E. Smirni Characterization of Moments and Autocorrelation in MAPs *SIGMETRICS Performance Evaluation Review*, 35(1):27-29, 2007.
- [2] P. Courtois. Decomposability, instabilities, and saturation in multiprogramming systems. *CACM*, 18(7):371–377, 1975.
- [3] E. D. Lazowska, J. Zahorjan, G. S. Graham, and K. C. Sevcik. *Quantitative System Performance*. Prentice-Hall, 1984.
- [4] N. Mi, Q. Zhang, A. Riska, E. Smirni, and E. Riedel. Performance impacts of autocorrelated flows in multi-tiered systems. *Perf. Eval.*, 64(9-12):1082–1101, 2007.
- [5] M. F. Neuts. *Structured Stochastic Matrices of M/G/1 Type and Their Applications*. Marcel Dekker, New York, 1989.
- [6] B. Schroeder and G. A. Gibson. Understanding disk failure rates: What does an MTTF of 1,000,000 hours mean to you? *ACM Trans. Storage*, 3(3):8, 2007.