Analytic Performance Models for Bounded Queueing Systems

Praveen Krishnamurthy
Roger D. Chamberlain

Praveen Krishnamurthy and Roger D. Chamberlain, "Analytic Performance Models for Bounded Queueing Systems," in *Proc. of Workshop on Advances of Parallel and Distributed Computing Models*, April 2008 (associated with IPDPS).

Dept. of Computer Science and Engineering
Washington University
Campus Box 1045
One Brookings Dr.
St. Louis, MO 63130-4899

Analytic Performance Models for Bounded Queueing Systems[1]

Praveen Krishnamurthy and Roger D. Chamberlain
Dept. of Computer Science and Engineering, Washington University in St. Louis
{praveen,roger}@wustl.edu

Abstract

Pipelined computing applications often have their performance modeled using queueing techniques. While networks with infinite capacity queues have well understood properties, networks with finite capacity queues and blocking between servers have resisted closed-form solutions and are typically analyzed with approximate solutions. It is this latter case that more closely represents the circumstances present for pipelined computation. In this paper, we extend an existing approximate solution technique and, more importantly, provide guidance as to when the approximate solutions work well and when they fail.

1. Introduction

Many applications can be effectively parallelized using pipelining techniques. Examples include sensor-based signal processing, biosequence analysis, text search, graphics processing, etc. When these applications are deployed on a pipeline of computational resources, queueing theory is a powerful tool for analyzing the performance of these systems. With effective performance evaluation possible prior to system construction, design choices can be made cognizant of the performance implications of those choices.

Traditionally, physically pipelined systems are modeled as a tandem queueing network. Nodes in the network represent tasks executing on a computational resource and queues buffer tasks between the nodes. The challenge in using queueing theory to model systems of this type is the explict need to incorporate the effects of bounded queues that exist between the stages of a computation pipeline. If an interstage queue is full, the upstream computations must block, and the majority of analytic results for queueing systems assume a lossy model, where incoming jobs are discarded. Here, we extend existing models that provide approximate solutions for blocking queueing networks and assess the conditions under which the approximations used in the models are reasonable.

[1]This work is supported by NSF grants CNS-0313203 and CCF-0427794. P. Krishnamurthy is currently with NVIDIA Corp.

Figure 1 illustrates the type of queueing network being modeled. An arrival process provides jobs at some mean rate λ. Each server provides service to jobs with mean rate μ_i. Upon completion of service, an individual job is either delivered to the downstream node (with probability d_i) or discarded (with probability $1 - d_i$). This feature can be used to model a filter computation, in which the results of processing at a node determine whether or not the job in question is to be passed further down the pipeline. Biosequence search is an example of this type of application.

Blocking is modeled as follows. Each node has an associated maximum capacity, K_i, representing the physical queue present in the computational pipeline, and if a node i is at capacity the upstream server at node $i - 1$ experiences blocking. Here, we assume a blocking after service model, in which the upstream node completes service for its job and then waits for the downstream node to have sufficient space to accommodate its output. Given the above information as input, the analytic model provides an estimate of the maximum sustainable throughput for the queueing system.

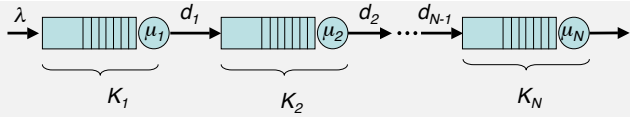


Figure 1. A tandem queueing network.

In this paper, we briefly review the analytic modeling techniques of Perros and Altiok [8], which provide an approximate solution for the case when node service time distributions are exponentially distributed and when the mean service rates at nodes are approximately equal. We have modified the original model to accomodate the deliver probabilities. We then extend this technique to include non-exponential service time distributions. The primary contribution of the paper is an assessment of when the analytic model works well and when the fundamental approximations underlying the technique are not met and the model therefore yields potentially erroneous performance predictions. We finish the discussion with a summary of related work and conclusions.

978-1-4244-1694-3/08/$25.00 ©2008 IEEE

2. Analytic Models

The model of [8] attempts to determine the maximum throughput supportable by queueing networks of the form illustrated in Figure 1. Table 1 summarizes notation.

Table 1. Notation and description of terms.

Symbol	Description
N	The number of the nodes (server units) in the tandem network
T_j	The assumed mean throughput of the queueing network (measured at the input) for the j^{th} iteration of the algorithm
t_i	The throughput at node i of the network
λ_i	Hypothetical lossy mean arrival rate into node i of the network, the jobs are assumed to have an exponentially distributed interarrival time
μ_i	Mean service rate of node i
d_i	Prob. that a job completing service at node i needs service at (is delivered to) node $i+1$
K_i	The maximum number of jobs that can be present at any node i, including the one (if any) in service
$p_i(x)$	Prob. of having x jobs at node i in the network
π_i	Prob. of a job leaving node i being blocked from entering node $i+1$ of the network

The analytic model is an iterative approach in which:

(1) an initial throughput assumption is made for the network, T_0;

(2) the implications of that throughput are evaluated at each node of the network in turn, moving from back to front;

(3) at node 1, the resulting throughput, t_1, is compared to the initial assumption, T_0; and

(4) if convergence has not yet been achieved, a new throughput assumption, T_{j+1}, is formulated from T_j and t_1.

In step (1), the initial throughput assumption is computed as the minimum service rate of each of the nodes.

$$T_0 = \min\left\{\mu_1, \min_{2\le i\le N}\left\{\mu_i / \prod_{s=1}^{i-1} d_s\right\}\right\} \qquad (1)$$

In step (2) for iteration j, starting with $j = 0$, each node of the network is considered in isolation, using standard single-server queueing models to predict the queue occupancy distribution $p_i(x)$ at node i, which can be used to predict the upstream blocking probability, π_{i-1}. For example, when the service time distribution is exponential with mean rate μ_i, the queue occupancy is given by

$$p_i(x) = \frac{(1-\lambda_i/\mu_i)(\lambda_i/\mu_i)^x}{1-(\lambda_i/\mu_i)^{K_i+1}}, 0 \le x \le K_i + 1, \qquad (2)$$

where λ_i is the mean rate of a (lossy) arrival process chosen to support the throughput at node i, $t_i = \prod_{1\le s\le i} d_s T_j$. When the service time distribution is phase-type, the queue occupancy is determined using the techniques of Neuts [6]. For both cases,

$$\pi_{i-1} = \mu_i p_i(K_i + 1)/\lambda_i. \qquad (3)$$

Moving from node i to node $i-1$ (i.e., from the back to the front of the network), node $i-1$'s service distribution is altered to account for blocking downstream. The altered distribution is modeled as phase-type in which a job, upon completion of original service, enters a blocking phase with probability π_{i-1}.

In step (3), the resulting throughput at node 1, t_1, is compared with the current estimate T_j. In step (4), if convergence has not yet been achieved, a new throughput estimate, T_{j+1}, is formulated and the technique iterates from step (2). Additional details of both the original model of [8] and its extensions can be found in [5].

Both the original model with exponential service distributions and the extended model with phase-type service distributions were each tested against a queueing network simulator using 200 synthetically generated queueing networks. The parameters for each of these experiments are given in Tables 2 and 3. For the phase-type distributions, a squared coefficient of variation less than one, $c^2 < 1$, implies a distribution with a tighter tail than an exponential distribution. This would model, for example, an algorithmic stage in a pipeline that has closer to a deterministic compute time. A squared coefficient of variation greater than one, $c^2 > 1$, models an algorithmic stage with a highly variable compute time (i.e., the tail of the distribution is heavier than that of an exponential distribution).

Figure 2 shows the results of these experiments, comparing the throughput of the queueing network as predicted by the analytic model with that predicted by the queueing network simulator. Simulation results are plotted with 95% confidence intervals, which are tight enough as to appear as a single point at this scale on the graphs.

We conclude from these plots that, generally, the analytic model works quite well for the cases explored in the random experiment. Of the 400 synthetically generated networks, only 35 have a discrepancy between the analytic and simulation models of more than 10% (5 from the exponentially distributed service time experiments and 30 from the phase-type service time experiments). In the next section, we specifically explore the circumstances under which the analytic model works well and also performs poorly.

Table 2. Range of parameters tested for exponentially distributed service times.

Symbol	Range of values
N	{2,3, . . . , 10}
K_i	{5,10,20,30,40,50,60,70,80,90,100,110,120}
μ_i	{10,20, . . . , 1000}

Table 3. Range of parameters tested for phase-type service time distributions.

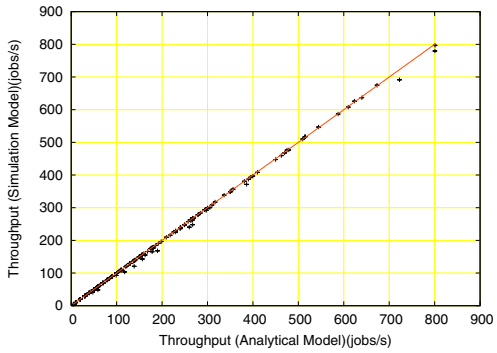
Symbol	Range of values
N	{2,3, . . . , 10}
K_i	{5,10,20,30,40,50,60,70,80,90,100,110,120}
μ_i	{10,20, . . . , 1000}
c^2	{0.5,0.8,1.1,1.3,1.6,2.0,3.0,4.0,5.0,7.5,10,15,20}

3. Assessing the Analytic Models

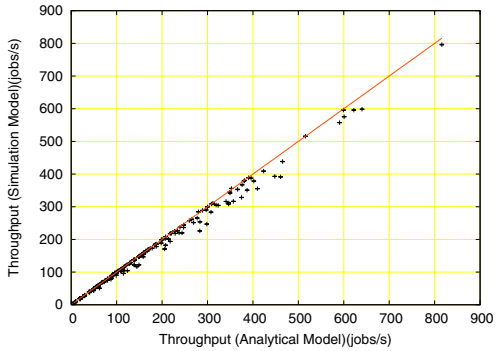
Detailed examination of the randomly generated test cases leads us to pay particular attention to the following aspects of the analytic model.

1. Queueing networks for which there is clearly one bottleneck node do not require this level of analysis. The throughput of the system is dominated completely by the throughput of the slowest node, and all upstream nodes effectively serve as extensions of the queue associated with the bottleneck node.
2. A queueing system becomes non-work-conserving when the queue associated with a node alternately is empty (starving the node) and full (blocking the upstream node). This circumstance is more likely as: a) the size of the queue between nodes gets smaller, b) the service rates for two adjacent nodes are similar to one another, and c) the variability in the service distribution of a node increases.
3. The quality of the analyic model throughput results are closely tied to the blocking probability experienced by upstream nodes.

To explore the above observations, a set of test cases were developed to examine the associated parameter space explicitly. Figures 3 and 4 represent results from 2 node experiments for which each node has an exponentially distributed service time (i.e., the squared coefficient of variation, $c_i^2 = 1, 0 \leq i \leq 1$). In the first experiment (Figure 3), the service rates for both nodes are equal ($\mu_i = 300$ jobs/s, $0 \leq i \leq 1$), the capacity of the upstream node is $K_0 = 100$ (chosen to be large enough as to not impact the throughput), and the capacity of the downstream node, K_1, ranges from 5 to 100. This first experiment is designed to explore the impact of queue size on the analytic model.



(a) Exponential service distribution

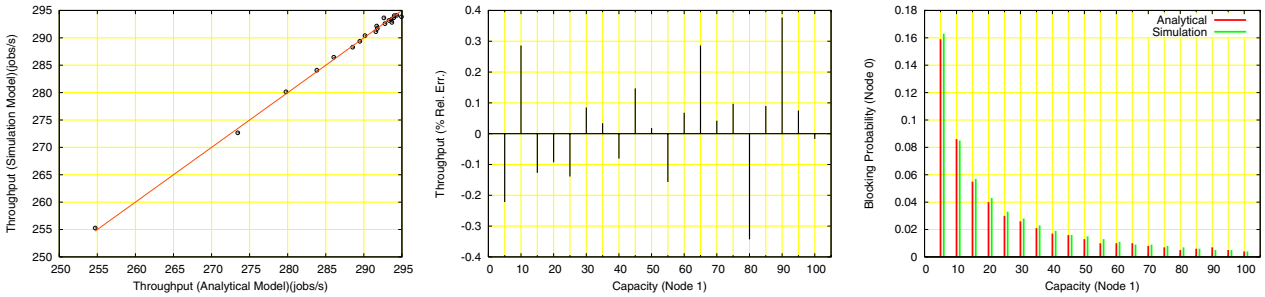


(b) Phase-type service distribution

Figure 2. Throughput of analytical model versus simulation.

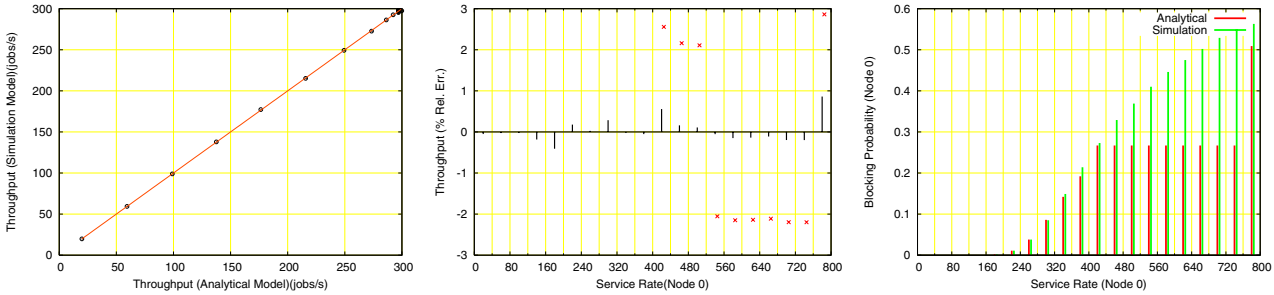
For each of the experiments performed, we plot the throughput as predicted by the simulation model vs. the throughput as predicted by the analytic model. The straight line reference that is added to the plot represents perfect alignment between the two models. The second plot for each experiment shows the relative error in the analytic model as a function of the independent variable being varied for the experiment (e.g., downstream node capacity for Figure 3(b)). The third plot for each experiment compares the probability that the upstream node is blocked for each of the analytic and simulation models. Although not shown explicitly in the first plot, the low-throughput points correspond to the cases of low downstream node capacity (and corresponding high upstream blocking probability).

The graphs of Figure 3 correspond to a parameterization explicitly covered by the original models in [8]. Clearly, there is close correspondence between the analytic model predictions and the simulation model predictions, both for overall throughput and blocking probability for the upstream node. For small downstream queue sizes, the throughput drops off and the blocking probability increases, as one would expect.



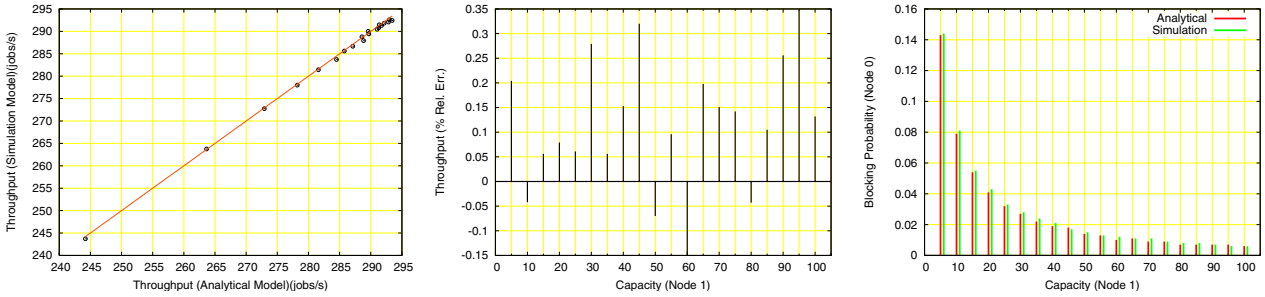
(a) Throughput (b) Throughput relative error (c) Upstream blocking probability

Figure 3. Throughput and upstream blocking probability predicted by analytical and simulation models ($c_0^2 = 1$, $c_1^2 = 1$, $\mu_0 = 300$ jobs/s, $\mu_1 = 300$ jobs/s, $K_0 = 100$, $K_1 \in [5, 100]$).



(a) Throughput (b) Throughput relative error (c) Upstream blocking probability

Figure 4. Throughput and upstream blocking probability predicted by analytical and simulation models ($c_0^2 = 1$, $c_1^2 = 1$, $\mu_0 \in [20, 780]$ jobs/s, $\mu_1 = 300$ jobs/s, $K_0 = 100$, $K_1 = 10$).



(a) Throughput (b) Throughput relative error (c) Upstream blocking probability

Figure 5. Throughput and upstream blocking probability predicted by analytical and simulation models ($c_0^2 = 1$, $c_1^2 = 2$, $\mu_0 = 300$ jobs/s, $\mu_1 = 300$ jobs/s, $K_0 = 100$, $K_1 \in [5, 100]$).

Further examination of the individual cases (out of the randomly generated tests) with poor correlation between the analytic and simulation throughputs indicates a pair of circumstances under which the analytic model inadequately corresponds to the physical system being modeled. The first circumstance is the case in which a downstream node is sufficiently slower than the immediate upstream node that, in effect, the upstream node (and its associated queue) act as an extention to the queue associated with the downstream node. The second circumstance is the case in which the service process for the upstream node differs sufficiently from an exponential distribution that a Poisson model for the arrival process for the downstream node is no longer effective. We will examine each of these circumstances individually.

3.1. Test 1

Figure 4 shows the results of an experiment designed to explore the first circumstance described above. Here, the capacity of the downstream node is fixed at $K_1 = 10$, the mean service rate of the downstream node remains at $\mu_1 = 300$ jobs/s, and the mean service rate of the upstream node, μ_0, is varied between 20 and 780 jobs/s. The system throughput closely tracks the upstream throughput while $\mu_0 < \mu_1$ (the middle and left of the first plot, Figure 4(a)), stabilizing near μ_1 as μ_0 exceeds μ_1. While the throughput results still represent a good match between the analytic and simulation models, there is a clear discrepancy between the two for the upstream blocking probability. In effect, the upstream node's queue is essentially acting as an extention of the downstream node's queue, and the model is not effectively characterizing this fact.

An explicit check for this condition can be formulated, in which the throughput is determined for a single-node system comprised of the downstream node's server with a capacity equal to the sum of the two nodes' capacities. We call this check "test 1." For an exponential downstream server with with arrival rate λ, service rate μ, and queue capacities K_1 and K_2 the maximum sustainable throughput is $t_{max} = \big(1 - \frac{1-\lambda/\mu}{1-(\lambda/\mu)^{K_1+K_2+1}}\big)\mu$. If the analytic model's predicted throughput for some node i exceeds the limit imposed by test 1 (i.e., $t_i > t_{max}$), we conclude that the analytic model is giving erroneous results. Cases where this occurs are marked by an "x" in Figure 4(b), and this mark will be inserted on the middle graph for all of the remaining figures whenever test 1 fails.

While [8] dealt with exponentially distributed service times, we have extended the model to address more general service distributions. Figures 5 through 7 show the results of experiments in which there is increased variability in the service time of the downstream node. In the experiments of Figures 5 and 6, the service rates for both nodes are again equal ($\mu_i = 300$ jobs/s, $0 \leq i \leq 1$), the capacity of the upstream node is $K_0 = 100$, and the capacity of the downstream node, K_1, ranges from 5 to 100. What differs from the first experiment is the squared coefficient of variation, c_1^2, for the downstream node, which is set to 2 in Figure 5 and 5 in Figure 6. The results show a close match between analytical and simulation models for the entire range of queue sizes explored, both for throughput and for upstream blocking probability.

The sensitivity of the analytic models to dissimilar service rates is again illustrated in Figure 7. As in Figure 4, the service rate for the upstream node is varied over the range $\mu_1 \in [20, 780]$ jobs/s and the capacity for both nodes is fixed at $K_0 = 100$ and $K_1 = 10$. The results here are similar to the results of Figure 4. When the upstream service rate is low, the analytic model accurately reflects this fact. When the two service rates are comparable, the analytic model again performs well. As the upstream service rate exceeds the downstream service rate, the model eventually fails test 1 as indicated in the relative error plots.

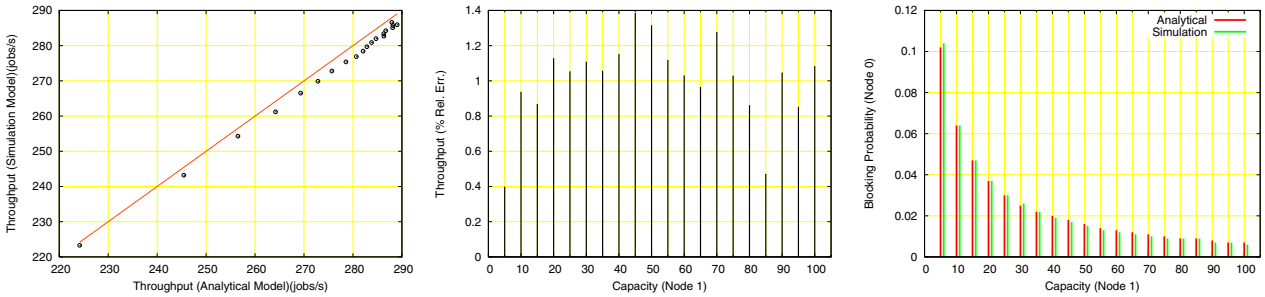
The next set of experiments investigates the case where the upstream node's service distribution varies from exponential. Figures 8 through 10 show the results of experiments where the service rates for the two nodes are returned to be equal ($\mu_0 = \mu_1 = 300$ jobs/s), the service distribution of the downstream node is returned to exponential ($c_1^2 = 1$), the capacity of the downstream node is varied ($K_1 \in [5, 100]$), and the squared coefficient of variation of the upstream node is different for each individual experiment. ($c_0^2 \in [0.8, 1.3]$).

Across the board, these experiments show good results for the analytic model, with a close match between the analytic model's predictions and that of the simulation model.

3.2. Test 2

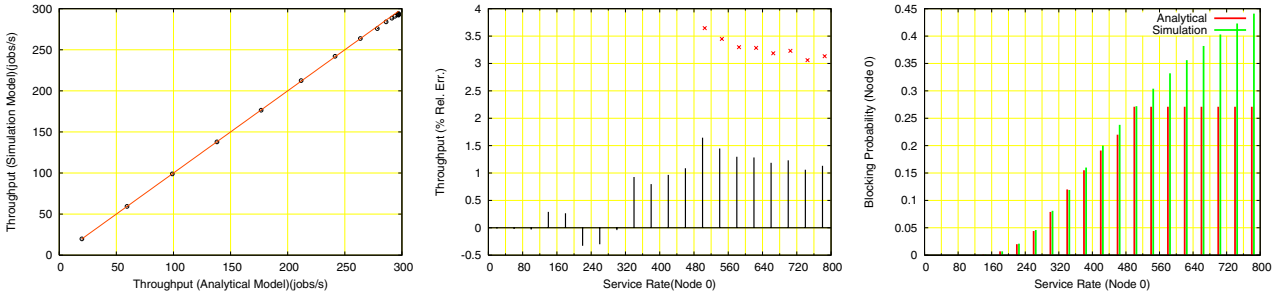
The second circumstance described above in which the model fails to adequately represent the system being studied is when the tail of the service distribution of the upstream node is sufficiently heavy so as to invalidate the implicit assumption of a Poisson arrival process at the intermediate node(s). This is illustrated in Figure 11, which shows the case where the squared coefficient of variation of the upstream node is increased to 2. Here, the tail of the upstream service time distribution is significantly greater than that of an exponential distribution.

The results in this last figure clearly show the analytic model failing to accurately predict the throughput and blocking probability. This leads us to a second check on the analytic model, called "test 2." Unfortunately, unlike test 1 which has sound theoretical underpinnings, test 2 is empirically based. Essentially, test 2 is a restriction on the range of input values supported by the model. If an upstream node with a $c^2 \geq 2$ experiences any blocking from the corresponding downstream node (i.e., $\pi > 0$), test 2 says the analytic model will fail to accurately represent the throughput of the network. Note that this is an intentionally conservative test. Figure 11 indicates acceptable errors for



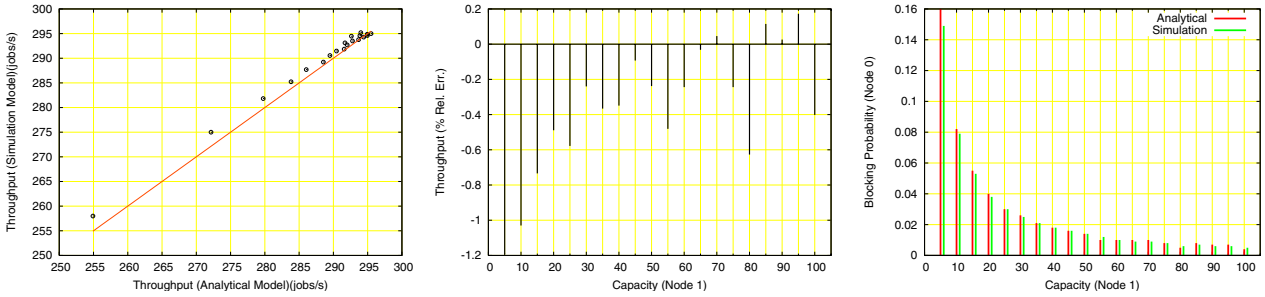
(a) Throughput (b) Throughput relative error (c) Upstream blocking probability

Figure 6. Throughput and upstream blocking probability predicted by analytical and simulation models ($c_0^2 = 1$, $c_1^2 = 5$, $\mu_0 = 300$ jobs/s, $\mu_1 = 300$ jobs/s, $K_0 = 100$, $K_1 \in [5, 100]$).



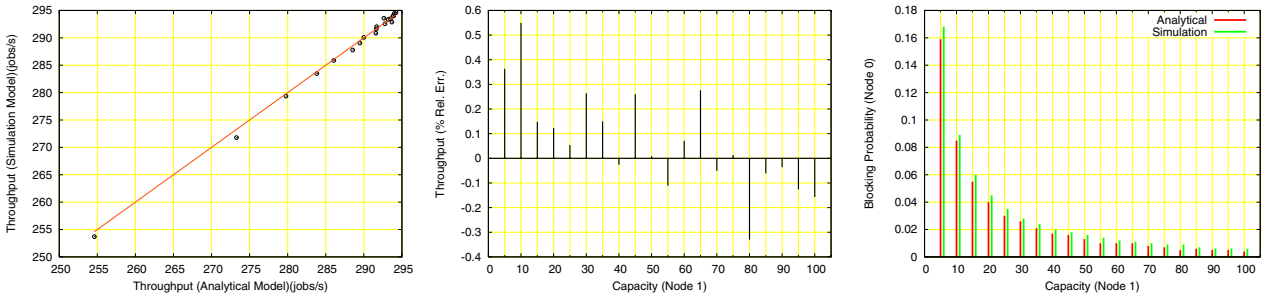
(a) Throughput (b) Throughput relative error (c) Upstream blocking probability

Figure 7. Throughput and upstream blocking probability predicted by analytical and simulation models ($c_0^2 = 1$, $c_1^2 = 2$, $\mu_0 \in [20, 780]$ jobs/s, $\mu_1 = 300$ jobs/s, $K_0 = 100$, $K_1 = 10$).



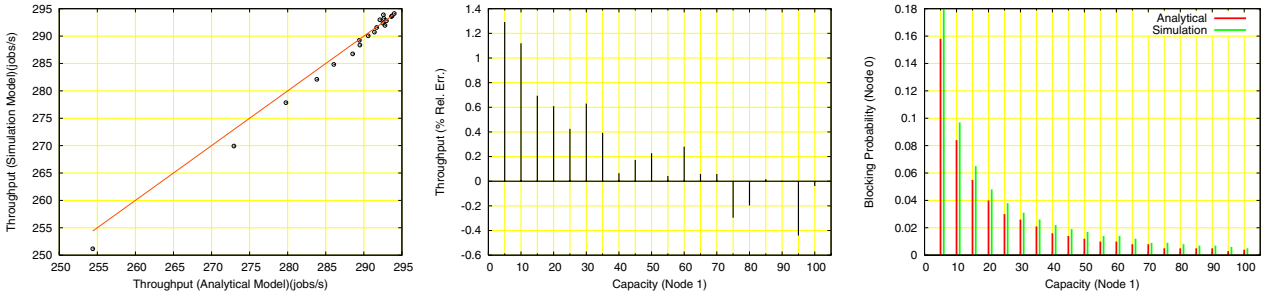
(a) Throughput (b) Throughput relative error (c) Upstream blocking probability

Figure 8. Throughput and upstream blocking probability predicted by analytical and simulation models ($c_0^2 = 0.8$, $c_1^2 = 1$, $\mu_0 = 300$ jobs/s, $\mu_1 = 300$ jobs/s, $K_0 = 100$, $K_1 \in [5, 100]$).



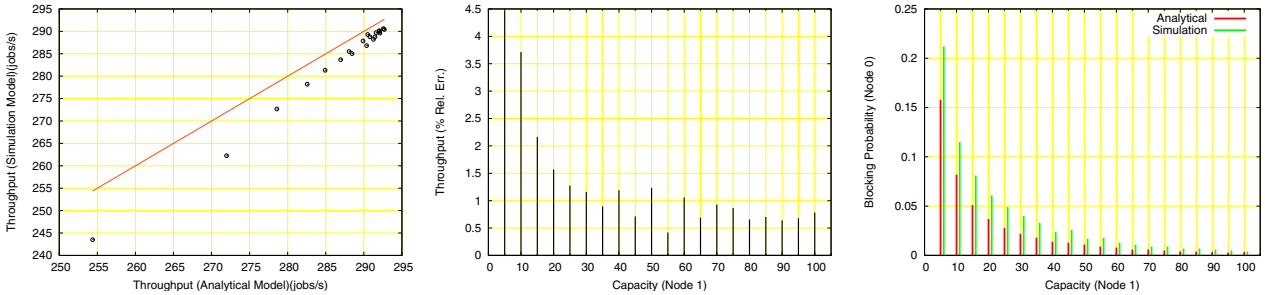
(a) Throughput (b) Throughput relative error (c) Upstream blocking probability

Figure 9. Throughput and upstream blocking probability predicted by analytical and simulation models ($c_0^2 = 1.1$, $c_1^2 = 1$, $\mu_0 = 300$ jobs/s, $\mu_1 = 300$ jobs/s, $K_0 = 100$, $K_1 \in [5, 100]$).



(a) Throughput (b) Throughput relative error (c) Upstream blocking probability

Figure 10. Throughput and upstream blocking probability predicted by analytical and simulation models ($c_0^2 = 1.3$, $c_1^2 = 1$, $\mu_0 = 300$ jobs/s, $\mu_1 = 300$ jobs/s, $K_0 = 100$, $K_1 \in [5, 100]$).



(a) Throughput (b) Throughput relative error (c) Upstream blocking probability

Figure 11. Throughput and upstream blocking probability predicted by analytical and simulation models ($c_0^2 = 2$, $c_1^2 = 1$, $\mu_0 = 300$ jobs/s, $\mu_1 = 300$ jobs/s, $K_0 = 100$, $K_1 \in [5, 100]$).

all but the lowest downstream queue capacities; however, test 2 leads us to disregard the analytic model's predictions.

3.3. Random Experiments

Returning to the random experiments of Figure 2, we now consider the implications of the two tests developed above. Table 4 shows the results of applying test 1 to the 5 cases in Figure 2(a) for which the relative error in the analytic model's throughput prediction is greater than 10%. Test 1 catches each of these cases.

Table 4. Test 1 check for experiments with $>$ 10% error in Figure 2(a).

Exp. #	Analytic model throughput (jobs/sec)	t_{max} (jobs/sec)	Result
24	117.6	105.3	Fail
47	47.6	40.9	Fail
84	138.6	122.9	Fail
116	190.1	177.6	Fail
120	58.8	49.1	Fail

Turning our attention to the remaining 30 cases of greater than 10% error (in Figure 2(b)), 25 of these 30 networks failed test 2 and the remaining 5 failed test 1. The tests successfully detected every case in which the analytic model's performance predictions were greater than 10% error.

We note here that there could still be cases where one could have false positives from the analytical model. The analytical model could potentially pass the above tests and still produce incorrect estimates. Also, test 2 is conservative in nature and one could potentially be discarding results of the analytical model for networks where the model works well. In essence tests 1 and 2 are necessary but not sufficient checks on the correctness of the analytical model.

4. Related Work

The models presented in this paper rely heavily on previous work in queueing theory, especially on solutions to single-server queues with Poisson arrival processes and either exponentially distributed service times or phase-type service times. The case for exponentially distributed service times is a classic one whose solution is given in [4]. The case for phase-type service times requires a matrix geometric solution developed by Neuts [6]. In each of the above models, the arrival process is assumed to be lossy (i.e., if a job arrives to find a full queue, that job is discarded). The two assumptions that make the models in this work approximate rather than exact are: 1) Poisson arrivals at the input to each node, and 2) the use of a lossy arrival mode to solve for the queue occupancy distributions.

The seminal work on networks of queues was by Jackson [2, 3], in which he showed that, for infinite capacity queues, the steady state probabilities are found via a product form solution in which each queue is treated individually. This was subsequently expanded by Baskett et al. [1] to cover a range of service disciplines and queueing protocols. Two approaches to modeling queueing networks with finite capacity queues and associated upstream blocking are state-space models and approximate models. State-space models have the difficulty that the model does not scale well with increasing nodes in the network [7]. The approximate models of Perros and Altiok [8] are the ones we have extended and assessed here.

5. Conclusions

In this paper we have extended the approximate analysis of Perros and Altiok [8] for queueing networks with finite queues and upstream blocking to handle more general service distributions. Phase-type distributions with squared coefficient of variation less than 1 are often better models of the computational requirements of pipelined applications, since many algorithm stages are unlikely to have as heavy a tail as an exponential distribution. The model consistently works well under these circumstances.

We have also assessed both the original model and our extensions to determine conditions under which the approximate modeling assumptions cause the results to be questionable. Two tests are developed that enable the user to rule out specific cases where the models are known to fail.

References

[1] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *J. ACM*, 22(2):248–260, 1975.

[2] J. Jackson. Network of waiting lines. *Management Science*, 5(4):518–521, 1957.

[3] J. Jackson. Jobshop-like queueing systems. *Management Science*, 10(1):131–142, 1963.

[4] L. Kleinrock. *Queueing Systems, Volume 1: Theory*. John Wiley & Sons, 1975.

[5] P. Krishnamurthy. *Performance Evaluation for Hybrid Architectures*. PhD thesis, Dept. of Computer Science and Engineering, Washington Univ. in St. Louis, Dec. 2006.

[6] M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. John Hopkins Univ. Press, 1981.

[7] R. Onvural, H. Perros, and T. Altiok. On the complexity of the matrix-geometric solution of exponential open queueing networks with blocking. In *Proc. of Int'l Workshop on Modelling Techniques and Performance Evaluation*, pages 3–12, 1987.

[8] H. Perros and T. Altiok. Approximate analysis of open networks of queues with blocking: Tandem configurations. *IEEE Trans. Soft. Eng.*, 12:450–461, 1986.