Approximate Analytical Models for Networked Servers Subject to MMPP Arrival Processes

Bruno Ciciani, Andrea Santoro, Paolo Romano
Dipartimento di Informatica e Sistemistica
Università di Roma "La Sapienza"
{ciciani, santoroa, romanop}@dis.uniroma1.it

Abstract

Input characterization to describe the flow of incoming traffic in network systems, such as the GRID and the WWW, is often performed by using Markov Modulated Poisson Processes (MMPP). Therefore, to enact capacity planning and Quality-of-Service (QoS) oriented design, the model of the hosts that receive the incoming traffic is often described as a $MMPP/M/1$ queue. The drawback of this model is that no closed form for its solution has been derived. This means that evaluating even the simplest output statistics of the model, such as the average response times of the queue, is a computationally intensive task and its usage in the above contexts is often unadvisable.

In this paper we discuss the possibility to approximate the behavior of a $MMPP/M/1$ queue with a computational effective analytical approximation, thus saving the large amount of calculations required to evaluate the same data by other means. The employed method consists in approximating the $MMPP/M/1$ queue as a weighted superposition of different $M/M/1$ queues. The analysis is validated by comparing the results of a discrete event simulator with those obtained from the proposed approximations, in the context of a real case study involving a GRID networked server.

1 Introduction

In the context of queuing theory, one well known model for system evaluation is the $M/M/1$ queue, which is often appreciated for its fast computability. However, workload characterization studies of networked systems, such as the GRID and the WWW, show that the incoming traffic behavior in such systems must rely on more complex models than a simple Poisson process [2, 3, 6, 9, 12, 16].

In order to capture the typical characteristics of the incoming traffic (such as self-similar behavior, burstiness behavior, and long range dependency) one of the most used models is the Markov Modulated Poisson Process ($MMPP$) [14, 15, 11, 17, 13], which is simply a Poisson process whose mean value changes according to the evolution of a Markov Chain [5]. However evaluating even the average response times of the $MMPP/M/1$ queue is a computationally intensive task, thus making it unfit for, e.g., capacity planning of a large scale system.

In this paper we discuss the possibility to approximate the behavior of a $MMPP/M/1$ queue analytically, thus saving the large amount of calculations required to evaluate the same data by other means. The method employed consists in approximating that queue as a weighted superposition of different $M/M/1$ queues. Specifically, we derive an approximation that overestimates the $MMPP/M/1$ response time, which could be employed in the context of capacity planning for Quality-of-Service (QoS) oriented system design. Tightness of the overestimation vs the real response time is supported in our study via a comparison between the output provided by our analytical solutions and simulation results based on real traces describing the traffic incoming to a networked GRID server [10].

The rest of the paper is structured as follows. Section 2 shows how to derive a reasonable approximation that consistently overestimates the response time/queue length of a $MMPP/M/1$ queue. In Section 3, simulation results for a validation of the analysis are presented. Finally, conclusions and future work are discussed in Section 4.

2 The Analysis

2.1 Rationale

The object of this paper is to derive a stochastic process which approximates the behavior of a $MMPP/M/1$ queue by exploiting results in the context of the evaluation of $M/M/1$ queues. We will show that such an approximation consistently overestimates the $MMPP/M/1$ queue length and response time, and that such an overestimation is tight for realistic settings for the $MMPP/M/1$ parameters (i.e., when considering traces from a real world GRID server). Note that when the value of a variable is consistently overestimated, its cumulative distribution function is consistently underestimated. Hence, the following relation holds on the response time/queue length cumulative distribution functions of the proposed approximation and on those of the original $MMPP/M/1$: $F_{MMPP/M/1}(\mathbf{r}) > F_{approximation}(\mathbf{r})$. In other words, the approximating process provides a lower bound on the cumulative distribution function of the response time and queue length of a $MMPP/M/1$ queue, thus from now on we will refer to this approximation as to a "lower bound".

The relevance of identifying a (tight) lower bound approximation for a $MMPP/M/1$ queue is that any computation that makes sure that $F_{approximation}(\mathbf{r}) > A$ also implies that $F_{MMPP/M/1}(\mathbf{r}) > A$. Hence, this approximation can be used for capacity planning purposes in the context of, e.g., QoS oriented design of networked servers, without incurring the risk of underestimating the computational demand of the system with consequent violations of any establised Service Level Agreement [4].

2.2 Approximation Construction

Consider a $MMPP/M/1$ queue, and let the MMPP that models the incoming traffic be composed by H states (S_1 ... S_H). We use the notation $M_i/M/1$ to refer to a $M/M/1$ queue whose average arrival rate is the λ_i observed in the generic S_i and the service rate μ is a constant among all the S_i. The analytical approximations studied in this paper are based on the following observation: if the MMPP stays in state S_i long enough without transitioning to another state, the mean response time and mean queue length at time t reach the same steady state observed for the corresponding $M_i/M/1$ queue. Those values are pinned on the same steady state value of $M_i/M/1$ as long as the MMPP does not change its state from S_i.

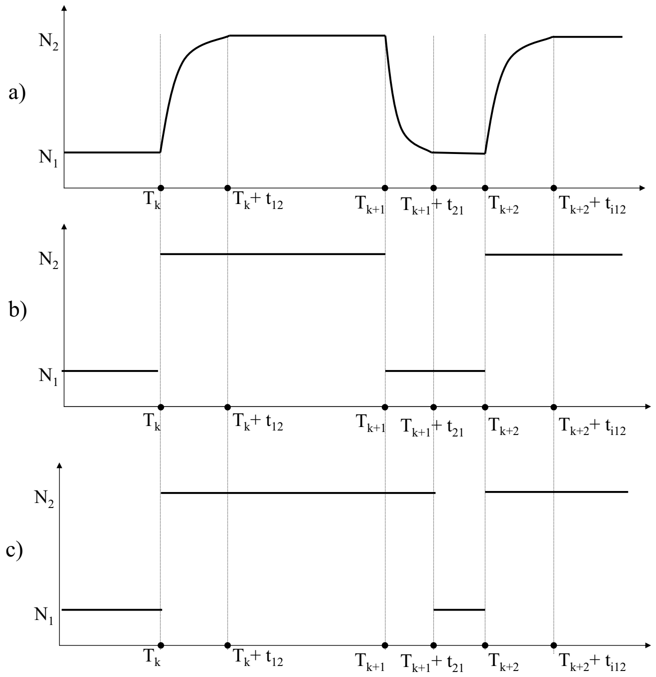


Figure 1. a) $MMPP/M/1$ behavior; b) Unbiased Approximation Behavior; c) Lower Bound Process Behavior.

As an example, let us consider a MMPP composed by two states. The mean number of resident requests at time t in a $MMPP/M/1$ queue is shown in Figure 1.a. Each instant T_k, T_{k+1} and T_{k+2} represents a transition between the two states of the MMPP. Therefore the evolution of the mean queue length value of the $MMPP/M/1$ can be described as follows: each time a state transition occurs there is a transient phase (t_{12} for a transition from S_1 to S_2 and t_{21} for a transition from S_2 to S_1) after which the mean queue length of the MMPP reaches the steady state of the correspondent $M_i/M/1$ (the behavior of the mean queue length during the transition is evaluated according the methodology shown in Appendix A). As soon as another state transition occurs, a new transient phase starts, a new steady state is reached and so on ([1]).

In this paper we consider two approximations of the behavior of a $MMPP/M/1$ based on weighted superpositions of the H steady state $M_i/M/1$ queues. For each ap-

[1] It is interesting to note that analytically a $M/M/1$ queue never reaches steady state, but merely approaches it asymptotically. Since we are studying an approximation, from now on we consider that a $M/M/1$ "reached steady state" when the difference between the mean queue length at time t and its theoretical value at steady state differ no more than an arbitrary value ϵ.

proximation we study the behavior of the average of queue length and response time, as well as their cumulative distribution functions.

2.3 Unbiased Approximation

The most obvious approximation can be derived by adopting the asymptotic probabilities for the MMPP to stay in each state S_i as the weights of the approximation. Specifically, by denoting (i) with Q_i the steady state queue length of $M_i/M/1$ and (ii) with p_i the asymptotic probability for the MMPP to stay in state S_i, the mean queue length of the $MMPP/M/1$ queue can be approximated as $Q = \sum_{i=1}^{H} p_i Q_i$, which would generate a queue process as the one shown in Figure 1.b. An analogous technique could be applied to derive the mean response time of the $MMPP/M/1$ queue, but there is an important difference. Specifically, the mean queue length is an integral average evaluated over time. Contrariwise, the mean response time is an average evaluated over the number of incoming requests, which are not distributed equally over time (during the state S_i the rate of requests is λ_i, while during the state S_j the rate of requests is a different amount λ_j, as in Figure 2). Therefore the mean response time of the $MMPP/M/1$ can still be a weighted sum of the mean response times of the $M_i/M/1$ queues, expressed as $R = \sum_{i=1}^{H} w_i R_i$, but the weights ($w_i$) are not simply composed by the asymptotic probabilities of the MMPP (as in the case of the average queue length) but must be scaled to keep into account the different arrival rate per each state. Hence $w_i = \frac{p_i \lambda_i}{\sum_{j=1}^{H} p_j \lambda_j}$. Finally, as far as the cumulative distribution and probability density functions of those parameters are concerned, they can also be derived as a weighted superposition of the correspondent functions of the separate $M_i/M/1$, the weights being those described above, respectively.

However this simple approximation is not immediately usable for the purposes described in Section 2.1. In fact, as shown in Figure 3, the error committed vs the real behavior of the $MMPP/M/1$ queue is given by the grayed out areas in the figure, which are the areas comprised between the transition from state S_i to S_j (for the real $MMPP/M/1$ queue) and the immediate transition to the steady state of S_j (as assumed in the analytical approximation) and vice versa. However, the two areas would tend to cancel each other since the error introduced when passing to a state with a higher utilization factor is positive (i.e. the approximation is already overestimating the average queue length on its own), while the error introduced by transitions to states with lower utilization factors are negative (the approximation is underestimating the average queue length). This means that it is not possible to obtain guarantees of overestimation from this approximation. In the following we will describe a variant of this approximation allowing the achievement of the above mentioned guarantee.

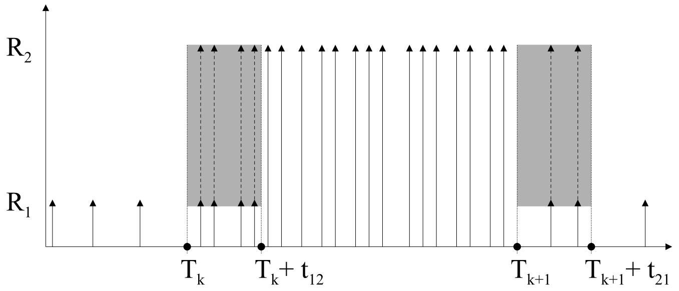


Figure 2. Behavior of the Average Response Time in a MMPP.

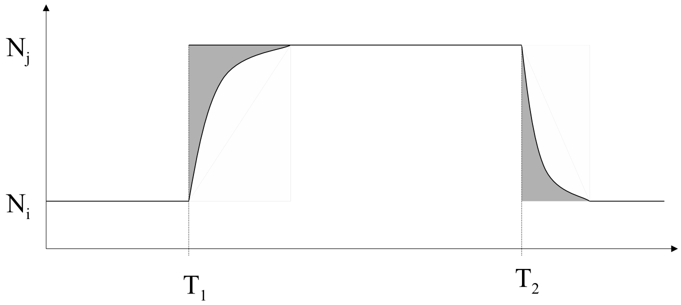


Figure 3. Difference Between Unbiased Approximation and Exact $MMPP/M/1$ Behavior.

2.4 Lower Bound Approximation

As shown in Figure 3, the error committed by the analytical approximation is given by the grayed out areas in the figure, which is the area comprised between the transition from state S_i to S_j (for the real $MMPP/M/1$ queue) and the immediate transition to the steady state of S_j (as assumed in the analytical approximation). Even with this intuitive understanding, exact evaluation of the error is not straightforward. This is due to the fact that the mean queue length during the transition from steady state of S_i to steady state of S_j can be analytically derived from the works in

[1, 8] (see Appendix A for the derivation), but the expression is not easily integrable.

However, a lower bound process on the queue length can be constructed by matching the $MMPP/M/1$ behavior during the steady state period, while systematically overestimating the queue length during transient periods, as shown in Figure 1.c. The generation of such a process employs the same technique described in Section 2.3 except that it requires to modify the probabilities p_i to reflect the different proportion among the average times spent in each of the MMPP states.

By the same considerations, the lower bound process on the response time can be also derived using the analogous approximation in Section 2.3, except that the weights w_i are changed according to the modified p_i.

Now we show a procedure to be used in generating the lower bound process. We assume that the following parameters are known:

- α_{jk}. The transition rates between every S_j and S_k of the MMPP.
- λ_i. The interarrival time for the incoming requests to the queue, when the MMPP is in state S_i.
- μ. The average service time required by the queue to serve each request (does not include the time spent waiting in the queue).

From the above parameters, to generate the straightforward approximation described in Section 2.3 we would just need to evaluate the weights w_i and apply them accordingly. However, since we want to generate the lower bound process described in Section 2.1, we need to modify the p_i employed by those weights so that they are augmented (or reduced) to reflect that each time there is a transition between two different $M_i/M/1$ queues, always the highest response time among the two transitions must be considered. Thus, considering that p_i also represents the amount of time spent by the MMPP in state S_i during one time unit, the modified probability is derived by: (i) adding to p_i a factor proportional to the time spent during each transition to lower utilization factors, and (ii) subtracting from p_i a factor proportional to the time spent during each transition from higher utilization factors (because that time is added to the probability p_j of the state with a higher utilization factor).

Hence, the following step-by-step procedure describes how to modify the p_i to obtain the lower bound process:

1. Evaluate the steady state probabilities for each state S_i of the MMPP. This can be done by using standard results in queuing theory ([7]). Denote such probabilities with p_i.

2. Evaluate the transition period for each transition listed above. According to standard queuing theory, the mean queue length $N_i(t)$ of a $M_i/M/1$ queue during a transition period can be evaluated according to the formula presented in Appendix A. We know that $\lim_{t\to\infty} N_i(t) = N_i$ and although $N_i(t)$ cannot be easily integrated we can still evaluate its value for specific values of t. Thus to evaluate the duration of a transition period we compute $N_i(t)$ for different values of t until we obtain a value for which the difference between $N_i(t)$ and N_i differ less than an arbitrary ϵ. Denote the transition period from S_i to S_j with t_{ij} ([2]).

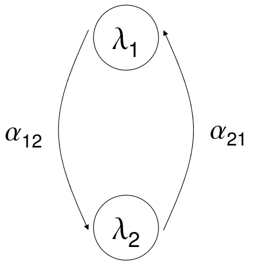
3. Evaluate the modified probabilities p'_i by using the formula: $p'_i = p_i + \sum_j^{\lambda_i>\lambda_j} p_i\alpha_{ij}t_{ij} - \sum_j^{\lambda_i<\lambda_j} p_j\alpha_{ji}t_{ji}$. In other words, the modified probability p'_i is generated by adding to each probability p_i the probability to be in a transition period from state S_i to a state S_j having lower request arrival rate, and subtracting from it the probability to be transitioning to state S_i from a state S_j having higher request arrival rate. It can be easily verified that p'_i are still probabilities since $\sum_{i=1}^{H} p'_i = 1$.

4. Generate the lower bound processes by performing a weighted superposition of the output processes of the different S_i with the newly derived p'_i. All the relevant statistics (mean value, density function, cumulative function) may be derived likewise.

3 Validation

In this section we aim at evaluating the tightness of our approximation techniques in realistic settings for what concerns the parameters space of the $MMPP/M/1$ queue. At this end, we compare the results provided by our approximations with those obtained via explicit simulation of a $MMPP/M/1$ queue. We set the $MMPP$ characterizing parameters on the basis of the results reported in [10], which has shown, via real traces analysis, the feasibility to model incoming traffic to a GRID server by means of a $MMPP/M/1$ model. According to the data reported in this work, the incoming traffic of the analyzed GRID server

[2] Note that in general t_{12} is not equal to t_{21}. Transitioning from an higher utilization factor to a lower one is typically a faster operation than the opposite one.



MMPP Parameters	
λ_1	22.1
λ_2	7.16
α_{12}	0.17
α_{21}	0.08

Figure 4. MMPP Model Employed for the Validation Study (Parameters Values from [10]).

Load Level	$\frac{Transients\ Duration}{Steady\ States\ Duration}$
Low	0.12%
Medium	1.48%
High	21.66%

Table 1. Normalized Duration of Transient Periods While Varying System Load.

can be modeled by a 2-state MMPP model, whose parameters are reported in Figure 4. Namely, transition rate α_{12}, from state S_1 to state S_2, is 0.17, while the reciprocal transition rate α_{21}, from state S_2 to state S_1, is 0.08. The request arrival rates λ_1 and λ_2, associated to state S_1 and S_2, are equal, respectively, to 22.1 and 7.16.

In [10] the service rate μ of the analyzed GRID server is not reported. Hence, we decide to treat μ as the independent parameter of a sensibility analysis aimed at evaluating the accuracy of both the unbiased and the lower bound approximations, respectively defined in Sections 2.3 and 2.4. At this purpose, we consider three different scenarios, representative of low, medium and high load situations. Specifically, we consider three different μ values corresponding to 10, 2, 1.25 times the maximum arrival rate λ_1: this determines server utilization factors for the $M_1/M/1$ queue respectively equal to 10%, 50% and 80%.

In Table 1 we report, for the three considered load scenarios, the sum of the duration of the transient periods (from state S_1 to S_2 and viceversa) normalized to the sum of the average permanence period in the two $MMPP$ states. Transient periods are evaluated using the methodology reported in Appendix A, and considering the transient period from S_i to S_j concluded when the mean queue length deviates no more than 5% from the mean queue length of the steady state S_j. The data in Table 1 shows that, in light and medium load scenarios, the duration of transient periods is negligible with respect to the permanence time in the steady states. Conversely, in the case of transitions from/to states with higher utilization factors, the relative weights of transient periods grow up to around 20% of the average permanence in states S_1 and S_2. This experimental data confirms the validity of the basic intuition underlying our approximation approach, namely that, when considering realistic parametric settings for a $MMPP/M/1$ queue, the permanence time in each state S_i of the $MMPP$ is long enough to allow reaching the steady state of the corresponding $M_i/M/1$ queue. Additionaly, the data in Table 1 highlights that, in high load scenarios, the relative increase of the transient periods duration may actually expose our analytical approximation techniques to higher errors.

The above inferences are confirmed by the data in Table 2. These data show that the deviation introduced by both the unbiased and the lower bound approximations is almost null in low and medium load scenarios. It is also interesting to note that while, as expected, the approximation errors show an increasing trend as the server load increases, the deviation still remains very limited. In fact, according to the data in Table 2, the percentual error for the average response time and queue length computed through the unbiased approximation never grows larger than 3%, whereas the deviation of the lower bound approximation peaks at 5.4% for the average response time and at 10.7% for the average queue length. Note that the unbiased approximation determines a minor deviation with respect to the lower bound one. This is due to that, as already hinted in Section 2.3, in the unbiased approximation the positive errors committed during transitions to state S_1, having higher utilization factor, tend to cancel with the negative errors due to transitions to state S_2, having lower utilization factor (see Figure 3). We recall, however, that the unbiased approximation does not provide any guarantee to overstimate the $MMPP/M/1$ output variables.

In this validation not only we aim at comparing the mean values of the aforementioned random variables, but rather at evaluating the whole statistical behavior of the proposed analytical approximations. Therefore, we next proceed through the plots in Figures 5, 6 and 7 to compare the cumulative distribution functions (C.D.F.) obtained from the approximations described in Section 2.4 with those produced via simulation of the $MMPP/M/1$. We omit plotting the C.D.F.s related to the low load scenario since, just like in the medium load case (see Figure 6), the curves obtained via simulation and via analytical approximations are indis-

Load	Mean Resp. Time Error		Mean Queue Length Error		Max Resp. Time Error (CDF)		Max Queue Length Error (CDF)	
Level	Lower Bound	Unbiased	Lower Bound	Unbiased	Lower Bound	Unbiased	Lower Bound	Unbiased
Low	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%	<0.1%
Medium	1.7%	0.1%	0.5%	0.2%	0.2%	0.1%	0.4%	<0.1%
High	5.4%	2.53%	10.7%	2.98%	1.98%	1.3%	2.57%	0.7%

Table 2. Deviation of Lower Bound and Unbiased Approximations from MMPP Simulation Output.

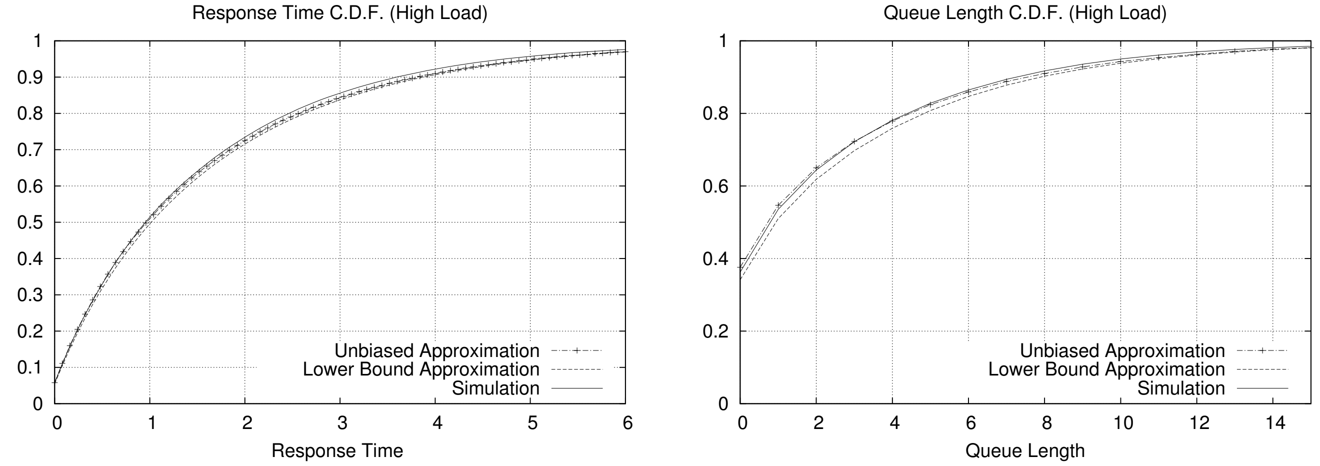


Figure 5. Cumulative Distribution Functions for Response Time and Queue Length (Heavy Load).

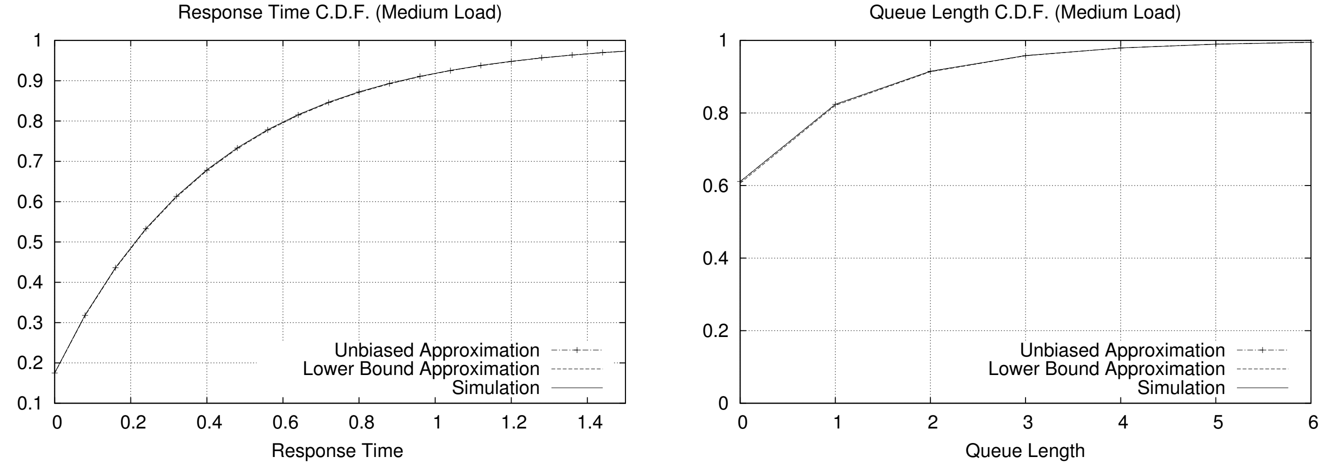


Figure 6. Cumulative Distribution Functions for Response Time and Queue Length (Medium Load).

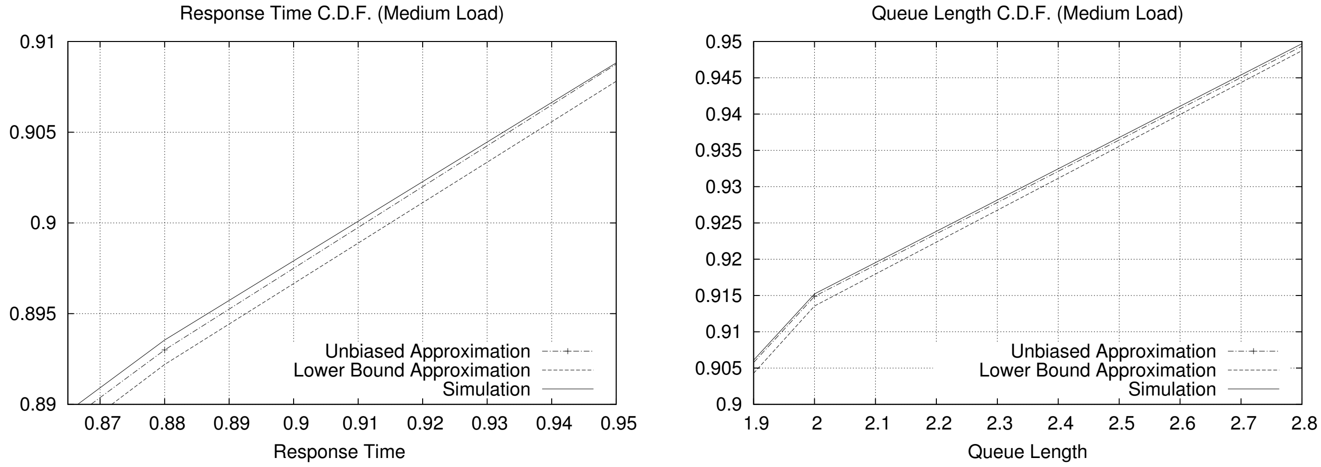


Figure 7. Zoom of the Cumulative Distribution Functions for Response Time and Queue Length (Medium Load).

tinguishable. However, in Figure 7 we report a (very strong) zoom of the C.D.F.s plots obtained in the medium load scenario, so to be able to visually quantify the relative distance among the C.D.F.s. The plot in Figure 5 shows indeed that even in the less favourable high load scenario, the C.D.F. curves produced by the lower bound and unbiased approximations lie very close to the one obtained via simulation. This is confirmed by the data reported in the four rightmost columns of Table 2, which show that the maximum puntual error of the queue length and response time lower bound C.D.F.s equal to 2.57% and 1.98%. Let us further analyze the worst case scenario of high load, and consider the 90-th or the 95-th percentiles of the response time, i.e., classical probability thresholds for Service Level Agreements established for QoS purposes. By the plot in Figure 5, we observe that the distance between the simulated $MMPP/M/1$ and the lower bound approximation does not exceed 5%, confirming the tightness of our approximation techniques even in this less favorable scenario.

Finally, yet importantly, we observe that the C.D.F. plots of the lower bound analytical approximation consistently underestimate the distribution of the output variables of the real $MMPP/M/1$ over all its length, thus validating the objective described in Section 2.1.

4 Future Work

In this paper we examined the feasibility of approximating a $MMPP/M/1$ queue with a weighted superposition of $M/M/1$ queues. By choosing appropriate weights we derived an approximation that guarantees an overestimation of the $MMPP/M/1$ output, hence providing lowerbounds on the cumulative distribution functions of the $MMPP/M/1$ output.

However during all the conducted experiments we have noticed that also the approximation employing the “unbiased” weights (as described in Section 2.3) slightly overestimates the real $MMPP/M/1$ output, while actually exhibiting better precision. The reason why the distributions of the “unbiased” approximation is likely to underestimate the distributions of the real $MMPP/M/1$ (instead of overestimating them) is that the transient time after T_2, shown in Figure 3, is *usually* much shorter than the transient time after T_1, thus the overestimation error normally oversteps the underestimation error. However, we cannot exclude the possibility of pathological cases in which this may not happen, thus we cannot propose such an approximation as a consistent, more precise lower bound. Theoretically establishing whether the “unbiased” approximation is a consistent, more accurate lower bound on the $MMPP/M/1$ behavior is part of our future work.

References

[1] J. Abate and W. Whitt, “Transient Behavior of the M/M/1 Queue Via Laplace Transforms”, Advances in Applied Probability, vol. 20, No. 1, 1988, pp. 145-178.

[2] L. Breslau, P. Cao, L. Fan, G. Phillipps, and S. Shenker, “Web Caching and Zipf-like Distributions: Evidence and Implications”, Proc. of IEEE INFOCOM, 1999.

[3] M. Crovella and A. Bestavros, “Self-similarity in World-Wide-Web traffic: Evidence and possible causes.”, IEEE/ACM Transactions on Networking, Vol.3, No.3, Jun. 1994, pp. 226-244.

[4] Y. Diao, B. Ciciani, C. H. Crawford, “Enforcing Quality of Service Using Decentralized Runtime Feedback Control”, Proc. of the 29th Int. Computer Measurement Group Conference, 2003, pp. 627-638

[5] W. Fischer and K. Meier-Hellstern, “The Markov-modulated Poisson process (MMPP) cookbook”, Performance Evaluation, Vol.18, No.2, Sep. 1993, pp. 149-171.

[6] A. Horvath and M. Telek, “A Markovian Point Process Exhibiting. Multifractal Behavior and Its Application To Traffic Modeling”, Proc. of MAM4, Adelaide, Australia, 2002.

[7] L. Kleinrock, “Queuing Systems”, Volume I: Theory, John Wiley & Sons, 1975.

[8] W. Leguesdron, J. Pellaumail, G. Rubino and B. Sericola, “Transient analysis of the M/M/1 queue”, Advances in Applied Probability, No.25, 1993.

[9] W. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, “On the self-similar nature of the Ethernet traffic (extended version)”, IEEE/ACM Transactions on Networking, Vol.2, No.1, Feb. 1994, pp 1-15.

[10] H. Li, M. Muskulus, and L. Wolters, “Modeling Job Arrivals in a data-intensive Grid”, Proc. 12th Workshop on Job Scheduling Strategies for Parallel Processing, 2006.

[11] A. Nogueira, P. Salvador, R. Valadas, and A. Pacheco, “Fitting elf-similar traffic by a superposition of MMPPs modeling the distribution at multiple time scales”, IEICE Transactions, Vol.E87-B, No.3, 2004, pp. 678-688.

[12] V. Paxson and S. Floyd, “Wide Area Traffic: the Failure of Poisson Modeling”, IEEE/ACM Transactions on Networking, Vol 3, No. 3, pp. 226-244, 1995.

[13] A. Riska, M. Squillante, S. Yu, Z. Liu, and L. Zhen, “Matrix-Analytic Analysis of a MAP/PH/1 Queue Fitted to Web Server Data”, Proc. of Int. Conference on Matrix Analytic Methods in Stochastic Models, July 2002.

[14] P. Rodriguez, C. Spanner, and E. W. Biersack, “Analysis of Web Caching Architectures: Hierarchical and Distributed Caching”, IEEE/ACM Transactions on Networking, Vol.9, No.4, Aug. 2001, pp. 404-418.

[15] P. Salvador, R. Valadas, and A. PAcheco, “Multiscale Fitting Procedure using Markov Modulated Poisson Processes”, Telecommunication Systems, Springer, Vol.23, No.1-2, 2003, pp. 123-148.

[16] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, “Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level”, IEEE/ACM Transactions on Networking, Vol.5, No.1, Feb. 1997, pp 71-86.

[17] T. Yoshihara, S. Kasahara, and Y. Takahashi, “Practical Time-Scale Fitting of Self-similar traffic with Markov Modulated Poisson process”, Telecommunication Systems, Vol.17, No.1-2, 2001, pp 185-211.

Appendix A

This appendix describes the calculations employed to derive the the transient behavior of the $M/M/1$ queue length. According to classical queuing theory[7], in the generic case the mean queue length $N(t)$ is given by $N(t) = \sum_{k=0}^{\infty} P_k(t)k$ where $P_k(t)$ is the probability for the queue to have k requests active at time t. Therefore to estimate $N(t)$ during the transition we need to evaluate the correspondent $P_k(t)$.

When the $MMPP/M/1$ settles on the steady state of S_i, the $P_k(t)$ is time independent with a well-known analytical solution[7]. However when a transition between the states S_i and S_j occurs, we have a situation in which $P_k(t)$ is again time-dependent until the steady state for S_j is reached. The value of $P_k(t)$ is evaluated as follows. Immediately before the transition occurs, the queue contains h requests with probability P_h (time-independent). Moreover, the works in [1, 8] derive analytical expressions for the probability $P_{h,k}(t)$, i.e. the probability for the queue to contain k request at time t conditioned to the fact that the queue contained h requests at time 0. Hence $P_k(t) = \sum_{h=0}^{\infty} P_{h,k}(t)P_h$ and $N(t) = \sum_{k=0}^{\infty} \sum_{h=0}^{\infty} P_{h,k}(t)P_h k$.

Note that these calculations for $N(t)$ involve infinite sums. Our decision on when the sums must be stopped is a logical AND among the two following conditions:

1. The sum of the P_h (or $P_k(t)$ (depending on whether we are summing over h or k) is larger than 0.9999. This means that our sum has already covered almost the whole probability space.

2. The summed value is smaller than 10^{-10} (the involved amounts of $N(t)$ are at least on the order of 10^{-1}).

By using this technique we have observed that the steady state values observed by the numerical calculation of the expression derived above, differ from the theoretical ones only at the fifth significative digit.