MAP-AMVA: Approximate Mean Value Analysis of Bursty Systems

Giuliano Casale
SAP Research
CEC Belfast, UK
giuliano.casale@sap.com

Evgenia Smirni
College of William & Mary
Williamsburg, VA
esmirni@cs.wm.edu

Abstract

MAP queueing networks are recently proposed models for performance assessment of enterprise systems, such as multi-tier applications, where workloads are significantly affected by burstiness. Although MAP networks do not admit a simple product-form solution, performance metrics can be estimated accurately by linear programming bounds, yet these are expensive to compute under large populations.

In this paper, we introduce an approximate mean value analysis (AMVA) approach to MAP network solution that significantly reduces the computational cost of model evaluation. We define a number of balance equations that relate mean performance indices such as utilizations and response times. We show that the quality of a MAP-AMVA solution is competitive with much more complex bounds which evaluate the state space of the underlying Markov chain. Numerical results on stress cases indicate that the MVA approach is much more scalable than existing evaluation methods for MAP networks.[1]

1. Introduction

The management of multi-tier web architectures and enterprise systems requires the continuous application of analytical methods to predict scalability and responsiveness under changing load intensities. Analytical models are fundamental for performance assessment, for capacity planning, for early identification of systems that need hardware upgrades, and to drive software reconfiguration decisions. Mean Value Analysis (MVA) [1] and its variations (e.g., approximations for workloads with high service time variability [8]) provide a comprehensive set of analytical tools that has been extensively used in the literature for the performance assessment of systems because of their ease of use and intuitive appeal. Typically, the only challenge in MVA models is input parameterization, i.e., given a specific system in operation how to best measure the workload demands and to be used as inputs to MVA [14]. However, as it often turns out, taking into account the multiple dependencies between servers or the complex characteristics of modern workloads (e.g., high-variability coupled with temporal locality and burstiness across different time scales) can be very challenging as classic capacity planning models cannot support such workload features [12]. Indeed, using MVA or other classic models for the performance predictions of systems with workload burstiness results in dramatic errors [12].

The recently proposed class of MAP queueing networks [3] provides a solution to this limitation of existing models by introducing a new framework to describe the effects of workload burstiness on the performance of distributed systems that may be modeled as a network of queues. A MAP queueing network is a generalization of a product-form queueing network [1] in which service times are no longer limited to be independent of each other (e.g., exponential or Coxian); instead, they can be more general point-processes known as Markovian Arrival Processes (MAPs) [10]. Workloads with periodicities, time dependence, or burstiness can be modeled as a MAP [6] and used within a MAP queueing network to describe accurately the service characteristics at the different resources; see [6] for a tool for automatic workload fitting into MAPs. Remarkably, the performance effects of workload burstiness that are unpredictable with classic models, such as the dynamic bottleneck switch phenomenon between resources that is frequently observed in real multi-tier systems [11], can be captured very accurately with MAP networks [2, 11]. Nonetheless, the only known approximation method for MAP queueing networks is the class of linear reduction (LR) bounds proposed in [3], which uses a probabilistic description of the models within a linear optimization program to compute bounds on metrics such as throughput, queue-lengths, and state probabilities. Although accurate, LR bounds can require very large time and space requirements if the model has many jobs or queues. In addition,

[1] This work is partially supported by NSF grants CNS-0720699 and CCF-0811417 and by the InvestNI/SAP MORE project.

the computation of LR bounds on large models can be subject to numerical difficulties because of the small values taken by the equilibrium state probabilities in very large state spaces [3]. As a result, there is a need to have approximation techniques for MAP queueing networks which are more economic and numerically stable than LR bounds, while preserving high approximation accuracy.

In this paper we introduce MAP-AMVA, a new approximation for MAP queueing networks targeted to models with burstiness or temporal dependence in the service process of the resources, but that can also be applied to networks with general independent renewal service. This new analytical evaluation is based upon formulas that are very similar to the ones of the original MVA. Following the path of classic queueing network theory, MAP-AMVA departs from the probabilistic approach used in [3] to a new approximation based on mean value analysis (MVA), in which only mean performance indexes are computed instead of individual state probabilities as in the LR bounds of [3]. Further, the proposed MAP-AMVA formulas have immediate performance interpretations, therefore they are intuitively appealing and simple to match with values that are measured directly in real systems. For instance, MAP-AMVA uses the concept of the mean *queue-length seen upon arrival* at the various stations, which we show captures analytically for the first time the phenomenon of bottleneck switch described in [11].

The contribution of this paper are as follows: (1) we describe a framework of exact equations which characterize the behavior of MAP queueing networks in terms of mean performance indexes only, such as utilization, throughput, and queue-length seen upon arrival; (2) we obtain the MAP-AMVA approximation scheme by numerical evaluation of these relations within a linear program, which is yet several orders of magnitude cheaper to solve than the LR bounds and with computational requirements that are independent of the total population size; (3) we illustrate the approximation accuracy of the technique on models with dynamic bottleneck switch that cannot be approximated by MVA and general-independent models.

The remainder of the paper is organized as follows. In Section 2, we illustrate the practical importance of MAP queueing networks using a case study of a real multi-tier architecture in which we show the prediction inaccuracies of classic models that ignore burstiness in service workloads. Section 3 gives model notation. In Section 4 we introduce the MAP-AMVA approach using a simple queueing network with two stations; results are extended to general networks in Section 5. The MAP-AMVA approximation method is introduced in Section 6 and validated in Section 7. Due to limited space, theorem proofs are all reported in the accompanying technical report [4].

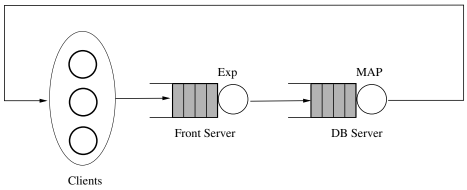


Figure 1. MAP queueing network model of the TPC-W system considered in Section 2.

2. Motivation

We point to [3] and references therein for background on research results related to this paper. In this section, we focus on explaining why the class of MAP queueing networks is relevant for practical performance assessment.

2.1 Burstiness and System Scalability

To illustrate the effectiveness of MAP queueing networks in the performance evaluation of real systems, we consider a capacity planning model of a real multi-tier system and we follow the modeling approach in [11]. The application is evaluated using the TPC-W benchmark, which simulates the operations of an online bookstore. The architecture is composed by a web server (Apache), an application server (Tomcat), and a back-end database (MySQL 5.0); all machines run Linux Redhat 9.0. The web server and the application server are installed on the same front server, a 1-way 3.2 GHz Pentium-D; the database runs on a 2-way 3.2 GHz Pentium-D, and the TPC-W requests depart from two 2-way 3.2 GHz Pentium-D client machines.

In the TPC-W benchmark, the HTTP requests are generated by a set of N clients that submit a new request only after completing the download of the previously requested page. Think times are exponentially distributed. Because of the closed-loop structure of the TPC-W workload, the number of simultaneous active user sessions is upper bounded by N and thus the system can be modeled as a closed queueing network with N jobs, where each job models a page download. Figure 1 models the multi-tier system as a MAP queueing network composed by two resources representing the front and database servers, respectively, preceded by a delay server that models the think times between submission of consecutive requests.

The service processes of the two queues in Figure 1 are parameterized from measurements of the standard "browsing mix" workload of TPC-W [11]. That is, the service process at the front server is modeled as an exponential process with mean service time $S_1 = 5.58$ ms; the think times have mean equal to $Z = 500$ ms. The service process at the database, instead, is found to be significantly affected by burstiness and therefore it is fitted using a two-phase MAP with mean $S_2 = 3.26$ ms, squared coefficient of variation $SCV = 16$, skewness $SKEW = 8.58$, and lag-1 autocorrelation coefficient[2] $\rho_1 = 0.40$. A MAP server works similarly to an hyper-exponential or Erlang server: the current rate of service depends on the active phase of the underlying Markov process. The main difference is that this can depend on past history, thus making MAPs capable of representing also time-varying properties, i.e., a MAP server describes a time series of service times and not only their distribution as in hyper-exponential or Erlang models.

Table 1 compares MAP queueing network predictions on the TPC-W system for the front server utilization against the measured values and the prediction of product-form networks analyzed by the MVA algorithm and of queueing networks with general independent (GI) service solved exactly by global balance [8]. The last column reports the measured utilization values at the front server on a two hours experiments with the browsing mix of TPC-W, where utilization samples are collected every five seconds. We first make the obvious observation that modeling methods do not provide results that are identical to the measured values because of the inherent approximation involved in a modeling process and possibly also due to slight measurement inaccuracies. Table 1 indicates that for small populations all methods from the literature are accurate. However, as the load grows and the effects of burstiness become more evident, the much increased accuracy of MAP queueing networks solutions is immediately visible compared to product-form and GI models, which for the largest population suffer severe errors up to 35.48% (0.9997) and 26.98% (0.9370) of the measured utilization (0.7379), respectively. MAP queueing networks, instead, have a small approximation error also on these problematic cases making the case of being much more robust that MVA and GI models in performance prediction of real systems under burstiness conditions. LR bounds [3] have been invented to evaluate systems where an *exact* MAP queueing network solution by global balance is computationally infeasible due to state space explosion, yet LR bounds can still suffer computational complexity limitation as we show in Section 7. This computational limitations provide motivation for the development of the MAP-AMVA approximation scheme.

	Front Server Utilization			
N	MAP QN [3]	MVA	GI [8]	*Real System*
25	0.2631	0.2727	0.2659	*0.2733*
50	0.4550	0.4875	0.4578	*0.4602*
75	0.6405	0.7194	0.6466	*0.6495*
100	0.7800	0.9479	0.8410	*0.7445*
150	0.7687	0.9997	0.9370	*0.7379*

Table 1. Performance modeling of the TPC-W e-commerce system for an increasing number of users N.

3. MAP Queueing Network Models

We consider a closed single-class MAP queueing network with M queues connected with arbitrary topology. Jobs at all queues are scheduled according to a First-Come-First-Served (FCFS) policy and both service and routing are load-independent. The mean service time at resource i is denoted by S_i. The term $p_{i,j}$ is the routing probability from queue i to queue j; the probability matrix $P = [p_{i,j}]$ implies a mean number of visits V_i of jobs at resource i which is equal to the probability of state i if P is regarded to as the probability matrix of a discrete-time Markov chain. The job population (also called multiprogramming level, number of users, or number of customers) is denoted by N. For simplicity of exposition, throughout the paper we consider a model where the first $M - 1$ queues have exponentially distributed service times, while the service times of queue M are modeled as a Markovian Arrival Process (MAP) with K phases[3]; the extension to the general case of the presented results is straightforward and follows the same ideas. Essentially, having additional MAP servers involves only additional summations in the equations developed throughout the paper to account for the different states and rates of service of the MAP server.

We start with a few definitions and notations for MAP queueing networks [3].The continuous-time Markov chain (CTMC) underlying the queueing model has state space $\mathcal{S}(N)$ that considers the distribution of jobs across the network and the current state of the MAP service process at queue M, i.e.,

$$\mathcal{S}(N) = \{(\vec{n}, k) \,|\, \textstyle\sum_{i=1}^{M} n_i = N \wedge n_i \geq 0 \wedge 1 \leq k \leq K\},$$

where n_i is the number of jobs in queue i and k is the current state of CTMC underlying the MAP. We indicate with $q_{i,j}^{k,h}$, $1 \leq k, h \leq K, 1 \leq i, j \leq M$, the transition rate from state

[2] The autocorrelation coefficients are a way to summarize fundamental properties of workload burstiness. In particular, large positive lag-1 autocorrelation, such as in this example, indicates that consecutive service times have usually similar magnitude, which leads to aggregation in bursts.

[3] A Markovian Arrival Process (MAP) with K phases can be described by two square matrices (D_0, D_1) both of order K, see [10] for a comprehensive overview. Essentially, a MAP is obtained from a continuous-time Markov chain (CTMC) by conventionally associating some user-selected transitions in the infinitesimal generator Q to the occurrence of the events that we want to model, e.g., service completion events. By placing these tagged transitions in D_1 and the remaining in D_0, such that $Q = D_0 + D_1$, a MAP is defined and its moment and correlations represent exactly the moment and correlations of the time between these user-specified events; see [10] for moment and correlation formulas in MAPs. In this way, if one is able to define a proper D_0 and D_1 couple, then a trace of events (e.g., trace of service times) can be fitted accurately by a MAP.

$(\vec{n}, k)$ to state $(\vec{n}-e_i+e_j, h)$, where the notation e_i indicates a vector of all zeros except for a one in the ith position. Intuitively, these transitions can be the result of either a job service completion, of a state jump performed by the MAP, or both; we point to [3] for analytical expression of $q_{i,j}^{k,h}$ as a function of the model parameters (service rates, $p_{i,j}$, and the jump rates that define the MAP process).

In the MAP-AMVA approach, we focus on the computation of mean performance indices only, such as utilization, throughput, queue-length, and response times. Let $\mathbb{P}[\vec{n}, k]$ be the equilibrium probability of state $(\vec{n}, k) \in \mathcal{S}(N)$. The mean queue length at resource i is given by $Q_i(N) = \sum_{k=1}^{K} Q_i^k(N)$, where

$$\textstyle\sum_{i=1}^{M} Q_i(N) = N \tag{1}$$

and $Q_i^k(N) = \sum_{\vec{n}} n_i \mathbb{P}[\vec{n}, k]$ is the (unconditional) mean queue-length at i when the MAP of server M is in state k. Similarly, the utilization is $U_i(N) = \sum_{k=1}^{K} U_i^k$, where

$$U_i(N) \leq 1 \tag{2}$$

and $U_i^k = \sum_{\vec{n}: n_i \geq 1} \mathbb{P}[\vec{n}, k]$ is the utilization of station i when the MAP process of queue M is in state k. The mean throughput $X_i(N)$ at server i is then given by

$$X_i(N) = \textstyle\sum_{k=1}^{K} \sum_{h=1}^{K} \sum_{j=1}^{M} q_{i,j}^{k,h} U_i^k(N),$$

and by Little's law [1] the average response time at i is $R_i(N) = Q_i(N)/X_i(N)$. Due to the forced flow law [1], the system throughput is always equal to $X(N) = X_i(N)/V_i$ for any choice of the station i. From [3], the utilizations at the MAP server are known to be related by

$$\begin{aligned} &\left(\textstyle\sum_{\substack{h=1 \\ h \neq k}}^{K} \sum_{j=1}^{M} q_{M,j}^{k,h} \right) U_M^k(N) \\ &\qquad = \textstyle\sum_{\substack{h=1 \\ h \neq k}}^{K} \left(\sum_{j=1}^{M} q_{M,j}^{h,k} U_M^h(N) \right) \end{aligned} \tag{3}$$

which is a consequence of the global balance equations that govern the MAP process at queue M. Further,

$$\begin{aligned} &\textstyle\sum_{k=1}^{K} \sum_{h=1'}^{K} \sum_{j=1}^{M} q_{i,j}^{k,h} U_i^k(N) \\ &\qquad = \textstyle\sum_{m=1}^{K} \sum_{k=1}^{K} \sum_{j=1}^{M} q_{j,i}^{k,m} U_j^k(N), \end{aligned} \tag{4}$$

for any choice of i and j, imposes equilibrium between the mean rates of the input and output flows of the different resources according to the flow balance rule [1]. Equations (1)-(4) are fundamental for the development of the MAP-AMVA approach introduced in the next sections.

4. Does Mean Value Analysis Apply?

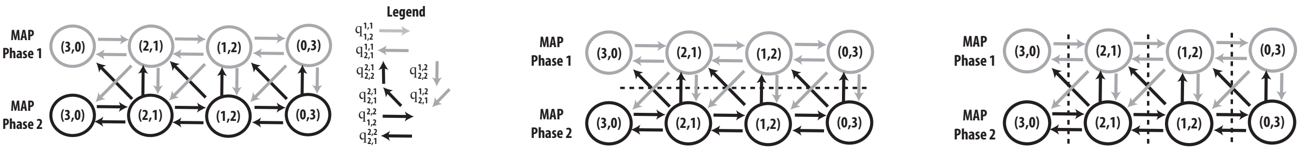
In this section, we investigate the applicability of classic MVA theory in the context of MAP queueing networks. Established MVA approximation of models with non-product-form features such as blocking, priorities, or high-variability have been obtained by an approximation methodology where one first describes the mean queue-length seen on arrival by a job joining a resource's queue [1], then proper corrections are introduced to account for the specific features of the model, and finally numerical solutions are obtained by fixed-point iteration similarly to the Bard-Schweitzer approximate MVA algorithm [1]. However, similar results do not exist for MAP networks. Here, we try to fill this gap by obtaining exact analytical formulas to describe mean MAP network performance and we draw insights from these observations leading to the MAP-AMVA approximation scheme presented in this paper.

Case Study. To evaluate the applicability of the MVA approach, we focus on a small MAP network with $N = 4$ jobs and $M = 2$ queues, one with exponential service times and one with two-phase MAP service times (i.e., $K = 2$). We assume that the two queues are in tandem. Thanks to the simplicity of the model, the underlying CTMC is the finite quasi-birth death (QBD) process shown in Figure 2. Markov models of such small dimension can be solved easily by global balance [1], but this technique has limited applicability on models with more than two queues because it scales poorly with the model size.

In order to obtain an MVA-like solution of the QBD using an approach different from global balance and that could remain efficient on general MAP queueing networks, MAP-AMVA evaluates the QBD in Figure 2 as follows. We start by considering a set of *partial balance* conditions on the QBD state space. These conditions are equations that impose equilibrium between the probability fluxes exchanged through horizontal or vertical cuts that separate the QBD states into disjoint partitions. For example, Figure 2(b) shows an horizontal cut that separates the state-space into two partitions, one describing the evolution of the network when the MAP server of resource M is in phase 1 (gray states), the other when the MAP is in phase 2 (black states). The partial balance arising from this cut imposes that the probability flux departing from the states where the second resource is busy and the MAP is in phase 1 is equal to the flow departing from the corresponding states where the MAP is in phase 2; it can be verified easily that this partial balance is exactly (3) in the case $K = 2$ and $M = 2$, thus it is immediately expressed in terms of mean values only, i.e., utilizations. Our goal is to obtain also from vertical cuts similar relations for mean queue-lengths and utilization that may help in understanding the mean performance of the system.

4.1 Queue Seen on Arrival

The goal of this subsection is to demonstrate that mean performance indices such as utilizations and mean queue-



(a) QBD (b) Horizontal Cut (c) Vertical Cuts

Figure 2. Finite quasi-birth death (QBD) process of the two queue MAP network considered in Section 4. A state label (n_1, n_2) indicates the queue-length of the exponential and the MAP queue, respectively.

lengths are inter-related by a number of equations that involve mean quantities only and that, surprisingly, do not require detailed low-level probabilistic information about the model, such as residual service times seen by arrival jobs or probability of finding the MAP server in a certain state[4], that are instead ubiquitous in non-product-form queueing analysis [1]. This is fundamental to understand the critical quantities that should be estimated by MAP-AMVA.

Let us consider the vertical cuts in Figure 2(c) which impose statistical equilibrium between the departure processes of the resources. Each cut provides a partial balance

$$\begin{aligned}\textstyle\sum_{k=1}^{K}(q^{k,1}_{2,1} + q^{k,2}_{2,1})\mathbb{P}[n_2, k]& \\ = q^{1,1}_{1,2}\mathbb{P}[n_2 - 1, 1] + q^{2,2}_{1,2}\mathbb{P}[n_2 - 1, 2],& \quad (5)\end{aligned}$$

for all possible job populations $1 \leq n_2 \leq N$ at the second resource, where we omit the population n_1 from notation since this is immediately computed as $n_1 = N - n_2$. The left-hand side of (5) describes departures from the MAP queue, the right-hand side characterizes departures from the exponential queue. Within the MAP-AMVA approach, we transform expressions involving probabilities into relations involving only mean values by weighted summation of the probabilistic expressions and by MAP filtration [9] applied to study the state of a resource upon arrival of a new job. For example, summing all possible balances (5) and weighting the contribution of each term by n_2, the left-hand side of (5) gives

$$\begin{aligned}\textstyle\sum_{k=1}^{K}(q^{k,1}_{2,1} + q^{k,2}_{2,1})\sum_{n_2=1}^{N} n_2\mathbb{P}[n_2, k]& \\ = \textstyle\sum_{k=1}^{K}(q^{k,1}_{2,1} + q^{k,2}_{2,1})Q^k_2(N),& \quad (6)\end{aligned}$$

while the right-hand side of (5) sums to

$$\textstyle\sum_{k=1}^{K} q^{k,k}_{1,2}\sum_{n_2=1}^{N} n_2\mathbb{P}[n_2 - 1, k] = X(N)V_2(1 + A_2(N)),$$

where V_2 is the mean number of visits of jobs to queue 2 and $A_2(N)$ is the mean queue-length seen by a job upon arrival at the MAP resource (not counting itself). The last observation follows by MAP filtration, since we have that the probability that a job arriving to the second resource finds there $n_2 - 1$ enqueued jobs while the MAP is in phase 1 is $\mathbb{P}^{arr}[n_2 - 1, 1] = q^{1,1}_{1,2}\mathbb{P}[n_2 - 1, 1]/(X(N)V_2)$, see [4] for a proof. According to the above observations, we have that the per-phase queue-lengths at the second resource are related by

$$\textstyle\sum_{k=1}^{K}(q^{k,1}_{2,1} + q^{k,2}_{2,1})Q^k_2(N) = X(N)V_2(1 + A_2(N)), \quad (7)$$

and multiplying both sides by $Q_2(N)$ we get

$$Q_2(N) = X(N)D_2(N)(1 + A_2(N)), \quad (8)$$

which is consistent with classic MVA theory and where the "mean service demand" $D_2(N)$ is

$$D_2(N) = \left(\frac{V_2}{\sum_{k=1}^{K} f^k_2(N)(q^{k,1}_{2,1} + q^{k,2}_{2,1})}\right), \quad (9)$$

with $f^k_2(N) = Q^k_2(N)/Q_2(N)$ being the fraction of the customers which waits in the MAP queue while the server is in phase k and $(q^{k,1}_{2,1} + q^{k,2}_{2,1})$ is the total rate of service in that phase. Here $D_2(N)$ summarizes in a single number several delay terms such as mean service time and residual time of the job found in service by a new arrival; $D_2(N)$ can be thought as a simple way to compact this complexity in a single number as if the service process would be exponential instead of MAP. Using similar passages in the case of the exponential queue we obtain

$$Q_1(N) = X(N)D_1(1 + A_1(N)), \quad (10)$$

which is again consistent with classic MVA theory and where $D_1 = S_1V_1$ is the mean service demand at the exponential queue which does not have service phases.

Discussion. Equations (8)-(10) state surprising properties of MAP networks. A first counter-intuitive property is that the residual time for completing the job in service when a new job arrives does not need to be explicitly involved in the computation of the mean queue-length and

[4] Note that, for MAP queue with service process (D_0, D_1), the residual time is computed as $\vec{a}^T(-D_0)^{-1}\vec{1}$, where $\vec{1} = (1, 1, \ldots, 1)$ and $\vec{a} = (a_1, a_2, \ldots, a_K)$ is defined by the probability a_k that an arriving job finds the job in service with the MAP in phase k. Therefore, residual times in MAPs are defined probabilistically and do not immediately simplify to mean quantities like in $M/G/1$ systems [1] since these are obtained under renewal assumption that do not apply in general to MAP networks.

consequently also of the mean response time $R_2(N) = Q_2(N)/X(N)$. This is surprising, since each job suffers a different residual time according to the arrival order and the statistical correlations between service times; thus, because of the correlations, one would *not* expect to factor out mean performance metrics into simple products such as (8). Equation (8) suggests instead that we do not need to explicitly account for residual times as long as we provide in the analysis detail on the mean queue-lengths in the different MAP server phases (i.e., the $Q_2^k(N)$ terms).

A second observation is that the structure of (8)-(10) is much harder to recursively approximate than product-form models, because response times depend not only on the queue-lengths seen on arrival, but also on the particular workload fractions $f_2^k(N)$ observed for population N. This tells us that basic interpolations of the queue-length seen on arrival using a model with population $N - 1$ are not enough to approximate the model, since one should also gain knowledge on how $Q_2(N)$ is distributed over the $Q_2^k(N)$ terms. Furthermore, (8)-(10) provide means to understand if the arrival theorem rule $A_j(N) = Q_j(N - 1)$ used in classic MVA analysis of product-form models [1] applies, at least approximately, to MAP queueing networks. We first observe that, by inserting the condition $n_2 = N - n_1$ into the derivation of (7)-(10), one could easily show that

$$A_1(N) + A_2(N) = N - 1, \qquad (11)$$

which is consistent with the product-form case where

$$A_1(N) + A_2(N) = Q_1(N - 1) + Q_2(N - 1) = N - 1.$$

Note that the validity of $A_1(N) + A_2(N) = N - 1$ for general closed networks has been recently proved by Varki *et al.* [13], thus this result is expected. Despite property (11), in MAP queueing networks it is simple to find cases where $A_j(N) >> Q_j(N - 1)$ or $A_j(N) << Q_j(N - 1)$, thus invalidating the conjecture that simple product form-like approximations apply easily to MAP queueing networks as we show in the next example.

Numerical Example. To exemplify the typical difficulty in approximating MAP networks, consider again the two-queue model and assume that the exponential server has mean $S_1 = 0.5$, while the MAP service times have mean $S_2 = 1$, squared coefficient-of-variation $SCV = 20$, skewness $SKEW = 7.0$, and lag-1 autocorrelation coefficient $\rho_1 = 0.40$. Consider the problem of approximating the queue-length seen on arrival: for a population $N = 20$, the mean queue-length seen by an arrival at the first resource computed by global balance is $A_1(N) = 15.338$. Yet, when one evaluates the mean queue-length of the first resource on the population $N - 1$ it is found that $Q_1(N - 1) = 5.8774$ which indicates that the classic MVA approximation $A_1(N) \approx Q_1(N - 1)$ is dramatically inaccurate and therefore traditional approaches cannot be used reliably with MAP queueing network models.

The results of this example also illustrate a peculiar characteristic of MAP queueing networks which cannot be captured by product-form models: the dynamic bottleneck switch phenomenon [11]. In fact, if we look at the mean queue-lengths at equilibrium for population $N = 20$ these are $Q_1(N) = 5.0644$ and $Q_2(N) = 14.936$, therefore it is legitimate to wonder: why the queue seen on arrival at the first exponential station is $A_1(N) = 15.338$ when the mean queue-length is only $Q_1(N) = 5.0644$? why station one is the bottleneck resource exactly at the instant of arrival of a new job thus creating the worst conditions for the response times? Similar questions hold also for the MAP queue where it is $A_2(N) = N - 1 - A_1(N) = 3.662$ and $Q_2(N) = 14.936$. Let us first observe that $A_1(N)$ is the queue-length seen by jobs arriving from the MAP queue 2 to the exponential queue 1. In the example, the MAP service times are such that 92% of the jobs receive fast service rate at queue 2, while only 8% are served there with a slow rate. Thus, out of 100 arrivals to queue 1, the queue seen on arrival is 92 times the one seen by jobs that are served quickly at the MAP queue 2 and only 8 times the one seen by slow jobs. However, for jobs served quickly at the MAP queue, the bottleneck in the queueing network is the exponential queue 1, thus 92% of the times the queue seen on arrival will be long and this explains the large value $A_1(N) = 15.338$ in this example. This is also true because in presence of positive autocorrelations the MAP serves for an extended period of time jobs that are all fast or all slow: therefore, when the MAP rate changes from slow to fast (or vice-versa) the queueing network has the time to reach equilibrium and move the bottleneck to the temporarily slowest server in the network. This makes the arrival queue-length seen by jobs very different according to the way they are served at the MAP. This property does not hold for product-form models and also for GI queueing networks where independent service times do not give to the network the time to reach equilibrium and shift the bottleneck position. This is because in GI models the service sequence of large or small jobs is random.

5 MAP-AMVA on General Models

The findings of the previous sections indicate that mean performance indices of a MAP networks such as queue-lengths and throughputs (hence by Little's Law also response time and utilizations) can be related to each other if: 1) we also consider in the analysis the queue-lengths seen on arrival by a new job; 2) we describe mean performance as observed during each possible active phase of the MAP server. In this section, we generalize the exact relations we have found in the previous section to MAP networks with arbitrary routing, arbitrary number of queues and MAP phases. Algorithms for approximating the mean performance indexes are given in Section 6.

5.1 Queue-Level Equations

When evaluating partial balances over general state spaces, one should first extend the definition of vertical and horizontal cuts used in the QBD example in Section 4. Consider a MAP with K phases and isolate a specific phase k; then a *generalized horizontal cut* for phase k is immediately obtained by the plane that separates the set of states where the MAP is in phase k from all other states of the network. Similarly, a *generalized vertical cut* is obtained by choosing a station i and population $n_i > 1$, and considering the plane that separates states where i has n_i jobs from states where it has $n_i - 1$ jobs, regardless of the active MAP phase. These generalized cuts specialize to the vertical and horizontal cuts illustrated in Figure 2(b)-(c) in the case $M = 2$ and $K = 2$. Further, by manipulations along the same lines of Section 4, we can obtain also from these cuts a set of generalized MVA relations. Let us begin by considered the MVA relations arising from horizontal cuts.

Theorem 1. *The generalized horizontal cuts over the state space of a MAP network give the following MVA balance*

$$\sum_{\substack{h=1\\h\neq k}}^{K}\sum_{w=1}^{M} q_{M,w}^{k,h} Q_M^k + \sum_{m=1}^{K}\sum_{j=1}^{M-1} q_{M,j}^{m,k} U_M^m = \sum_{j=1}^{M-1} q_{j,M}^{k,k} U_j^k + \sum_{\substack{h=1\\h\neq k}}^{K}\sum_{w=1}^{M} q_{M,w}^{h,k} Q_M^h, \qquad (12)$$

for all MAP phases $k = 1, \ldots, K$.

Proof. Theorem proof is given in [4]. □

The interest for (12) is that it relates the per-phase mean queue-lengths $Q_M^k(N)$ and thus provides immediate information that can be exploited to approximate the workload fractions $f_M^k(N) = Q_M^k(N)/Q_M(N)$. To understand precisely how the workload fractions interact with mean queue-lengths and the mean queue-length seen on arrival in the general case, we now study the balance arising from the generalized vertical cuts.

Theorem 2. *The generalized vertical cuts over the state space of a MAP network give the following MVA balance*

$$Q_i(N) = D_i(N)X(N)(1 + A_i(N)) \qquad (13)$$

for $i = 1, \ldots, M$, where $A_i(N)$ and

$$D_i(N) = \left(\frac{V_i}{\sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{j=1}^{M} q_{i,j}^{k,h} f_i^k(N)}\right) \qquad (14)$$

are respectively the mean queue-length and the mean service demand seen on arrival at queue i, and $f_i^k(N) = Q_i^k(N)/Q_i(N)$ is the fraction of the mean load at queue i that waits when the server is in phase k. In particular, if i is an exponential server then $D_i(N)$ simplifies to D_i, which is the mean service demand of the queue.

Proof. Theorem proof is given in [4]. □

Finally, we have the following corollary.

Corollary 1. *The mean queue-lengths seen on arrival satisfy the relation $\sum_{i=1}^{M} A_i(N) = N - 1$.*

Proof. As observed earlier in the paper, the theorem is a specialization of the recent results obtained in [13]. □

The above results provide an immediate generalization of the concepts discussed in the special case of the QBD in Section 4 to MAP networks with arbitrary size. Therefore, the discussion of the above results is identical to that in Section 4 and confirms the validity of our observations on the applicability of the MVA approach to MAP networks.

5.2 Network-Level Equations

The balances we have explored so far describe the state of a queue upon arrival of a new job. Consider the following questions: when a new job arrives to queue i, what is the state of the other queues in the network? How this affects the state of queue i? These are non-trivial questions that try to capture the essence of non-product-form models, where resources cannot be assumed nearly-independent of each other as in product-form networks. We now obtain a characterization of the state of the network seen at arrival instants at queue i.

Consider the instant of arrival of a new job at queue i; let us recall that $A_i(N)$ is the mean queue-length of resource i seen by the new arrival. Define

$E_j^i(N) =$ mean queue-length of resource j when a new job arrives to queue i (also from self-loops).

Clearly, $A_i(N) = E_i^i(N)$ and because of the closed nature of the system

$$\sum_{j=1}^{M} E_j^i(N) = N,$$

for any choice of the arrival destination queue i. We can now determine the relation between mean system performance and $E_j^i(N)$ values.

Theorem 3. *The state of the network seen upon arrival of a job to queue i satisfies the following balance*

$$\sum_{k=1}^{K}\sum_{h=1}^{K}\Big(\sum_{\substack{j=1\\j\neq i}}^{M} q_{i,j}^{k,h} Q_i^k(N) + \sum_{j=1}^{M} q_{j,i}^{k,h} Q_j^k(N)\Big) = \Big(N + M - 1 + \sum_{\substack{l=1\\l\neq i\neq j}}^{M} (1 + E_l^i)\Big)V_i X(N). \qquad (15)$$

Proof. Theorem proof is given in [4]. □

In the next section, (12) and Theorem 3 are used in the definition of the MAP-AMVA approximation.

6 MAP-AMVA Approximation Algorithm

We propose to use the equations derived in the previous section to characterize approximately the mean performance of the MAP network. The MAP-AMVA approximation first considers an optimization program based on the constraints (4), (12), (13) for $i = 1 \ldots M$, (3) for $k = 1 \ldots K$, (1), and (2). The unknowns of this optimization program are the queue-lengths $Q_i^k(N)$, the utilizations $U_i^k(N)$, the arrival queues $A_i(N)$, the network state variables $E_j^i(N)$ for $j \neq i$, and the throughput $X(N)$. To further restrict the range of feasible solution and obtain more accurate results, the optimization program is augmented with the following constraints

$$Q_i^k(N) \leq NU_i^k(N), \quad 1 \leq k \leq K, 1 \leq i \leq M,$$

which derive from the probabilistic-level definition of queue-length;

$$\textstyle\sum_{j=1}^M Q_j^k(N) \geq NU_i^k(N), \quad 1 \leq k \leq K, 1 \leq i \leq M,$$

which follows by noting that the mean queue-lengths $Q_j^k(N)$ are computed also on states where queue i is idle.

Based on the above constraints, the optimization program can use any objective function involving utilization, queue-length, throughput, or combinations thereof. These allows to compute minimum or maximum values for the mean performance of the system. *Note that these values are immediately bounds on the exact solution since all formulas used in the optimization program are exact.* Pointwise MVA estimates can then be obtained using well-known bound-based approximation schemes, such as the harmonic mean approximation proposed in [7].

For example, suppose that the mean throughput $X(N)$ is investigated. Then using the MAP-AMVA optimization program bounds $X^{min}(N)$ and $X^{max}(N)$ are immediately obtained by respectively considering, e.g., minimization and maximization programs on the utilization $U_1(N)$ and then scaling the result by the mean service demand of the first resource. Afterward, an harmonic approximation $X^{apx}(N)$ is obtained as

$$X^{apx}(N) = \frac{2X^{min}(N)X^{max}(N)}{X^{min}(N) + X^{max}(N)}. \tag{16}$$

An advantage of this approximation scheme is that the maximal relative error $\varepsilon_{rel}(N)$ is given by [7]

$$\begin{aligned} \varepsilon_{rel}(N) &= \max_{X(N)} \frac{|X^{apx}(N) - X(N)|}{X(N)} \\ &= \frac{X^{min}(N) - X^{max}(N)}{X^{min}(N) + X^{max}(N)}, \end{aligned} \tag{17}$$

where the maximization is over $X(N) \in [X^{min}(N), X^{max}(N)]$. The index $\varepsilon_{rel}(N)$ provides a simple way to assess the quality of performance estimates obtained by the bound-based harmonic approximation.

6.1 Numerical Solution

A direct numerical solution of the MAP-AMVA optimization program described above is challenging because of the non-linearities in equations such as (13). We therefore propose a simple linearization scheme which preserves the exactness of the representation. This approach also suggests an additional constraint which improves the quality of the approximation.

Define $\widetilde{E}_i^{j,k} = \sum_{\vec{n}:n_j \geq 1} n_i \mathbb{P}[\vec{n}, k]$ as the mean queue-length of station i while j is busy and the MAP is in phase k. Then we have the following result concerning the linearization of the MAP-AMVA optimization program.

Theorem 4. *The following linearization scheme*

$$V_i X(N) A_i(N) \rightarrow \textstyle\sum_{k=1}^K \sum_{h=1}^K \sum_{j=1}^M q_{j,i,k,h} \widetilde{E}_i^{j,k} \tag{18}$$

$$V_i X(N) E_i^l(N) \rightarrow \textstyle\sum_{k=1}^K \sum_{h=1}^K \sum_{j=1}^M q_{j,i,k,h} \widetilde{E}_l^{j,k} \tag{19}$$

preserves the exactness of the MAP-AMVA equations.

Proof. Theorem proof is given in [4]. □

An interesting consequence of this result is that we can add to the optimization program resulting from the linearization also the constraint

$$Q_i^k \geq \widetilde{E}_i^{j,k}, \quad 1 \leq i \leq M, \ 1 \leq j \leq M, \ 1 \leq k \leq K,$$

which holds true since Q_i^k is computed on a larger state space than $\widetilde{E}_i^{j,k}$ with the same analytical expression.

Finally, we remark the fundamental fact that *none* of the equations defined in this paper grows in cardinality with the population N. Therefore, the MAP-AMVA approximation is $O(1)$ with respect to this parameter and therefore it is expected to overcome the performance issues of LR bounds in the approximation of models with hundreds or thousands of jobs. A performance comparison of MAP-AMVA and LR bounds confirming this property is given in Section 7.

7 Numerical Validation

We illustrate the accuracy and computational efficiency of the MAP-AMVA method using case studies. Due to limited space, we focus on stress cases where MAP queueing network solutions cannot be approximated accurately by classic MVA models.

	MAP-AMVA			
N	CPU [sec]	MEM [MB]	iter	nonzeros
1000	< 1	0.2	23	153
2000	< 1	0.2	23	153
3000	< 1	0.2	23	153

	LR bounds			
N	CPU [sec]	MEM [MB]	iter	nonzeros
1000	3	54	1038	320363
2000	10	108	2089	640363
3000	28	161	3916	960363

Table 2. Comparison of computational costs of MAP-AMVA and LR bounds approximations.

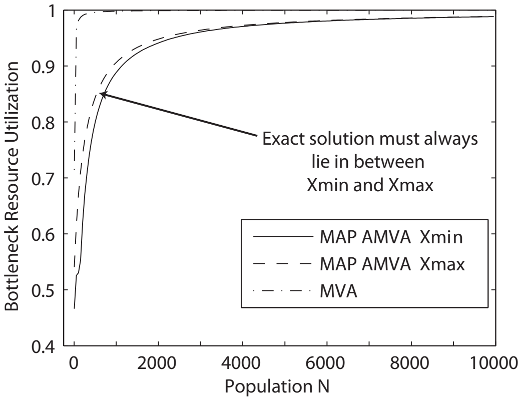


Figure 3. MAP-AMVA accuracy in case study 1. Exact solution is always included between the MAP-AMVA X_{min} and X_{max} values, but it is computationally prohibitive to obtain by global balance for $N > 100 - 200$. LR bounds (not shown in the figure) have slightly better accuracy, but their computational costs are several orders of magnitude larger than for the MAP-AMVA method, see Table 2.

7.1 Case Study 1

We first study a stress case considered in the literature for the evaluation of queueing networks with high service variability, i.e., Balbo's network discussed in [3]. This network has been used also in [3] for the validation of the LR bounds. The model is composed by three resources which are traversed by jobs according to the routing matrix

$$P = [p_{i,j}] = \begin{pmatrix} 0.2 & 0.7 & 0.1 \\ 1.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 \end{pmatrix}.$$

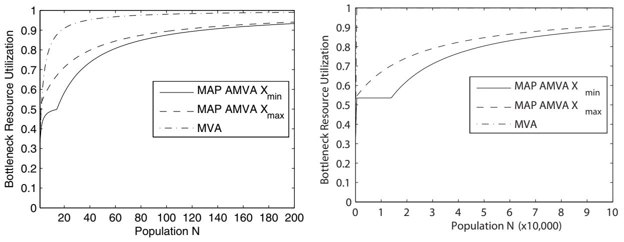
This routing matrix induces the mean number of visits $V_1 = 0.5556$, $V_2 = 0.3889$, and $V_3 = 0.0556$. The first two stations have exponential service times with mean service time $S_1 = 0.0280$ and $S_2 = 0.0400$, respectively. The third resource has a MAP server with mean service time $S_3 = 0.2800$, squared coefficient-of-variation $SCV = 24$, skewness $SKEW = 7.8843$, and lag-1 autocorrelation coefficient $\rho_1 = 0.4313$.[5] Note in particular that these values imply bottleneck switch since the bottleneck station changes with the active MAP phase: when the fast state of the MAP is active, the jobs queue at the bottleneck stations 1 and 2; when the slow phase of the MAP is active, queue 3 becomes temporarily the bottleneck.

The MAP-AMVA results obtained using the linearization approach described in Section 6.1 are shown in Table 2 and Figure 3. The figure shows the trend of the minimum and maximum bottleneck utilization bounds obtained from the MAP-AMVA linear program and compares it with the MVA solution of the corresponding product-form model that ignores workload burstiness. Indeed, the example under consideration is representative of models that cannot be approximated effectively as product-form, since the MVA values are very far from the feasible region for the exact MAP queueing network solution which is delimited by the upper and lower MAP-AMVA bounds. The MAP-AMVA results are in general extremely accurate: the gap between upper and lower bound is negligible for the great majority of the populations. The largest gap is found at very small populations ($N < 10$), where the maximum relative error is $\max_N \varepsilon_{rel}(N) = 12.5\%$; however, these populations do not represent a computational challenge and can be solved for maximal accuracy by a global balance solution which gives exact results. However, as the population increases, exact solutions cannot be computed by global balance if the model contains more than $100 - 200$ jobs, hence in this example one would not be able to explore exactly the performance in the large range of bottleneck utilizations between 60% and 90% and approximate techniques such as LR bounds and MAP-AMVA are needed. We remark that LR bounds follow very closely the MAP-AMVA curves, yet at much higher computational costs as discussed below.

Table 2 compares the computational costs of LR bounds and the MAP-AMVA approximation in the sub-range of populations $N = 1000 - 3000$. The table indicates the marked superiority of MAP-AMVA in terms of computational requirements that are negligible in all cases and $O(1)$ with respect to the population size. In addition, the dimensionality of the linear program itself is much less scalable than for the MAP-AMVA, since the number of non-zeros is about $10^5 - 10^6$ for LR bounds, while it dramatically re-

[5] These values parameterize a MAP(2) service process with

$$D_0 = \begin{pmatrix} -0.26777 & 0 \\ 0 & -52.589 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0.24287 & 0.024896 \\ 0.36939 & 52.22 \end{pmatrix}.$$



(a) no-burst

(b) high-burst

Figure 4. MAP-AMVA results on case study 2.

duces to $1.5 \cdot 10^2$ for MAP-AMVA, thus saving about $2-3$ orders of magnitude with respect to the LR bounds. Therefore, this example strongly argues for the effectiveness of MAP-AMVA in reducing the computational costs of the analysis while preserving high approximation accuracy as illustrated by the very tight bounds in Figure 3.

7.2 Case Study 2

In the second case study we examine the sensitivity of the results presented above to the burstiness intensity. This is interesting to illustrate the effectiveness of model approximation as the characteristics that are specific to MAP queueing networks become stronger. We consider the same network discussed in the first case study and we evaluate the bottleneck resource utilization when burstiness at the MAP server is removed (case *no-burst*) or increased by one hundreds times according to the index of dispersion metric (case *high-burst*) [10]. The results of this experiment shown in Figure 4 are qualitatively similar to the one presented for Case Study 1, with the gap between the upper and lower bounds being typically small, and the product-form approximations given by MVA lying far outside the feasible region for the exact solution. For the case of high-burst the MVA curve is not even clearly visible on Figure 4. In particular, the high-burst case is found slightly harder to approximate than the no-burst case: note in particular that evaluation up to 90% utilization of the bottleneck resource requires to reach a population of $100,000$ users, which is done in few seconds and with negligible memory consumption by MAP-AMVA.

8. Conclusion

We have presented the MAP-AMVA approximation for MAP queueing networks. We have illustrated in Section 2 that this class of capacity planning models is the only capable of predicting correctly the performance impact of burstiness over a distributed system that may be modeled as a network of queues. The main contribution of the MAP-AMVA technique is the identification of *exact* equations that relate mean performance indexes in a MAP queueing network and the definition of a linear programming approach to approximate these indexes. Numerical results indicate that our linear programming-based approach achieves good accuracy at negligible computational costs.

In future extension of this work we will provide more details on the application of the approach to models with several MAP servers. Further, we are currently evaluating techniques to incorporate into the MVA approach delay servers that are useful to represent user think times between submission of consecutive requests. This would be useful since the LR bounds cannot be applied to this case which requires more expensive bounds known as QR bounds [5].

References

[1] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi. *Queueing Networks and Markov Chains*. John Wiley and Sons, 2006.

[2] G. Casale, N. Mi, L. Cherkasova, and E. Smirni. How to parametrize models with bursty workloads. *ACM Perf. Eval. Rev., HOTMETRICS Special Issue*, 36(2):38–44, 2008.

[3] G. Casale, N. Mi, and E. Smirni. Bound analysis of closed queueing networks with workload burstiness. In *Proc. of ACM SIGMETRICS 2008*, pages 13–24. ACM Press, 2008.

[4] G. Casale and E. Smirni. An approximate mean value analysis algorithm for MAP queueing networks. TR WM-CS-2009-02, College of William and Mary, 2009.

[5] G. Casale, N. Mi, and E. Smirni. Model-Driven System Capacity Planning Under Workload Burstiness. In *IEEE T. Computers*, to appear.

[6] G. Casale, E. Zhang, and E. Smirni. KPC-toolbox: Simple yet effective trace fitting using markovian arrival processes. In *Proc. of QEST Conference*, pages 83–92, 2008.

[7] D. L. Eager, K. C. Sevcik. Bound hierarchies for multiple-class queueing networks. *JACM*, 33(1):179–206, 1986.

[8] D. L. Eager, D. Sorin, and M. K. Vernon. AMVA techniques for high service time variability. In *Proc. of ACM SIGMETRICS*, pages 217–228. ACM Press, 2000.

[9] D. A. Green. *Departure processes from MAP/PH/1 queues*. PhD thesis, The University of Adelaide, 1999.

[10] A. Heindl. *Traffic-Based Decomposition of General Queueing Networks with Correlated Input Processes*. Ph.D. Thesis, Shaker Verlag, Aachen, 2001.

[11] N. Mi, G. Casale, L. Cherkasova, and E. Smirni. Burstiness in multi-tier applications: Symptoms, causes, and new models. In V. Issarny and R. E. Schantz, editors, *Middleware*, LNCS 5346, pages 265–286. Springer, 2008.

[12] N. Mi, Q. Zhang, A. Riska, E. Smirni, and E. Riedel. Performance impacts of autocorrelated flows in multi-tiered systems. *Performance Evaluation*, 64(9-12):1082–1101, 2007.

[13] E. Varki, L. Dowdy, and C. Zhang. Quick performance bounding techniques for computer and storage systems with parallel resources. *Under review, available for download at http://www.cs.unh.edu/ varki/publication/bound.ps*.

[14] Q. Zhang, L. Cherkasova, and E. Smirni. A regression-based analytic model for dynamic resource provisioning of multi-tier applications. In *Proc. of ICAC*, 27–27, 2007.