Model-Driven System Capacity Planning Under Workload Burstiness

Giuliano Casale, Ningfang Mi, Evgenia Smirni
College of William and Mary
Computer Science Department
23187-8795 Williamsburg, Virginia
{casale, ningfang, esmirni}@cs.wm.edu

***Abstract*—In this paper, we define and study a new class of capacity planning models called *MAP queueing networks*. MAP queueing networks provide the first analytical methodology to describe and predict accurately the performance of complex systems operating under bursty workloads, such as multi-tier architectures or storage arrays. Burstiness is a feature that significantly degrades system performance and that *cannot* be captured explicitly by existing capacity planning models. MAP queueing networks address this limitation by describing computer systems as closed networks of servers whose service times are Markovian Arrival Processes (MAPs), a class of Markov-modulated point processes that can model general distributions and burstiness. In this paper, we show that MAP queueing networks provide reliable performance predictions even if the service processes are bursty.**

We propose a methodology to solve MAP queueing networks by two state space transformations, which we call Linear Reduction (LR) and Quadratic Reduction (QR). These transformations dramatically decrease the number of states in the underlying Markov chain of the queueing network model. From these reduced state spaces, we obtain two classes of bounds on arbitrary performance indexes, e.g., throughput, response time, utilizations. Numerical experiments show that LR an QR bounds achieve a mean accuracy error of 2%. We also illustrate the high effectiveness of the LR and QR bounds in the performance analysis of a real multi-tier architecture subject to TPC-W workloads that are characterized as bursty. These results promote MAP queueing networks as a new robust class of capacity planning models.

I. Introduction

Capacity planning of modern computer systems requires to account for the presence of nonrenewal features in workloads, such as short-range or long-range temporal dependence which significantly affect performance [26]–[28], [36], [40]. A typical example of temporal dependence is workload burstiness, where the sizes of consecutive jobs processed by the system are correlated, e.g., the arrival of a long job is likely to be followed by the arrival of another long job (and vice-versa for short jobs). Time-varying workloads of this type are naturally modeled as nonrenewal workloads with temporal dependence among consecutive requests.

Because of the complexity of their analysis, only small nonrenewal models based on one or two queues have been considered in the literature, mostly in matrix analytic methods research [31]. We address the current lack of more general provisioning models by introducing and analyzing a new class of closed queueing networks which can account for temporal dependence in the service processes. Our analysis enables for the first time the analytical performance evaluation of complex environments with nonrenewal workloads and immediately finds application in the capacity planning of multi-tier architectures and storage systems.

Capacity planning based on product-form queueing networks [4] has been extensively used in the past, since these models enjoy simple solution formulas and low computational cost of exact and approximate algorithms [11], [25]. However, modern Web, parallel, and storage systems often exhibit high variability in their service processes [2] and are therefore best modeled by networks of queues with first-come first-served (FCFS) queues and general independent (GI) service [7], [17], [34], [38]. Queueing networks with general independent (GI) service [7], [17], [34], [38] have been proposed as a solution, but although these models are much more accurate than product-form networks, they *cannot* be used for robust performance predictions if the service process is nonrenewal [9].

In this paper, we overcome the limitations of existing modeling techniques by providing a bound analysis methodology for queueing networks with nonrenewal workloads. We define and study a class of closed queueing networks, that we call *MAP queueing networks*, where service times are modeled by Markovian Arrival Processes (MAPs). MAPs belong to a family of point processes which can easily model general distributions as well as the main features of nonrenewal workloads, such as autocorrelation in service times [31]. Algorithms for fitting measurements into MAPs are available [1], [12], [23], [39].

Because of the well-known difficulty of extending exact solution formulas outside the product-form case, we study bound analysis techniques for MAP networks. With the exception of the general ABA bounds [30], which provide good estimates only for very low or very high population values, no bounding techniques for nonrenewal networks exist and this is due to the lack of exact results which are needed to prove the bounding property. In this paper, we overcome this classic limitation and obtain provable bounds on performance indexes also in non-product-form networks.

The proposed nonrenewal bounds derive from the analysis of the Markov process underlying the MAP queueing network. Because of the state space explosion, the queueing network's equilibrium behavior cannot be determined exactly, but we argue that it can still be bounded accurately by describing the

system with "reduced" state spaces, which we call *marginal state spaces*. Marginal state spaces capture the behavior of the network conditioned on a given queue being busy or idle and can be obtained from two transformations called Linear Reduction (LR) and Quadratic Reduction (QR).

The fundamental property of marginal state spaces obtained by the LR transformation is that the number of states grows *linearly* with the number of jobs in the network; thus, the analysis remains computationally tractable also on models with large populations. For example, a MAP queueing network with three queues and one hundred jobs has underlying Markov chain composed by 10^{12} states, but LR marginal state spaces reduce it to about $3,600$ states only. We then derive *exact* balance equations for the equilibrium behavior of the LR marginal state spaces. The number of these exact equations grows combinatorially with the model size, but it is yet insufficient for determining exactly the equilibrium probabilities of the marginal state spaces. Here, we illustrate how these formulas can be combined with linear programming [5], [29] for the computation of bounds on mean value indexes.

We then generalize MAP queueing networks to allow the inclusion of queues with load-dependent service rates. This feature is useful to describe resources, such as delay stations or flow-equivalent servers [13], which change speed dynamically as a function of the number of locally enqueued jobs. In particular, delay servers are often fundamental in capacity planning models to describe user think times between submission of consecutive requests to the system [20]. Motivated by the observation that LR bounds cannot be applied to load-dependent MAP queueing networks, we derive a more general state space transformation, called Quadratic Reduction (QR), that can be used to bound the performance of load-dependent models. QR bounds are observed to provide similar accuracy to LR bounds, but since the number of states of QR marginal state spaces grows quadratically with the number of jobs in the network, they should be used only to evaluate models with load-dependent servers where LR bounds do not apply. For instance, on the state space with 10^{12} states discussed above, the QR transformation considers $370,000$ states which, although more expensive that the LR transformation which uses only $3,600$ states, still remain much less than in the exact state space.

Outline and summary of contributions. The main contribution of this paper is to present a new methodology for the efficient analytic solution of queueing networks with nonrenewal workloads. This methodology automatically applies to queueing networks with renewal workloads as well. The stated contributions and outline of this work are as follows.

- We review existing approximations and decomposition methods for product-form and GI queueing networks and discuss their applicability to models with nonrenewal workloads (Section II).
- We define MAP queueing networks as a generalization of existing queueing networks that can model nonrenewal workloads (Section III).
- We develop the LR transformation and the related LR bounds on performance indexes of MAP queueing networks (Sections IV and V).
- We present the QR transformation and QR bounds for the analysis of MAP queueing networks with delays and/or load-dependent queues (Sections VI).
- We show validation results on random models and representative case studies proving that the LR and QR bounds capture very well mean performance indexes of MAP queueing networks (Section VII).
- Finally, Section VIII shows an example of performance analysis, based on the QR bounds, of a real multi-tier architecture subject to TPC-W workloads [20].

We stress that MAP queueing networks are a superset of existing non-product-form networks with GI workloads. Therefore, the presented analytic methodology has a wide applicability. The LR and QR bounds are corroborated by extensive numerical validation, where we show that they achieve a mean accuracy error of approximately 2% on a set of $10,000$ random models, promoting MAP queueing networks as versatile models of modern computer systems. Specifically, the QR bounds are indispensable for the frequent case of capacity planning models with delay servers that represent user think times between consecutive download requests. The AMPL specification [19] of the LR and QR bounds is available for download at `http://www.cs.wm.edu/MAPQN/`.

II. Previous Work

We review previous work on non-product-form queueing network models with FCFS queues and general independent (GI or renewal) service [6]; we point the reader to [6], [14], [25] for general background on queueing network modeling and Markov processes.

Closed networks of FCFS queues enjoy a product-form solution if all service times are exponentially distributed [4]. If one or more servers have renewal service, such as hyperexponential or Coxian [15], the product-form theory does not apply and approximate methods are used for evaluating performance [6].

An approximation based on Markov renewal theory is developed by Reiser in [34]. For each queue, the MVA arrival theorem [35] is generalized to include the coefficient of variation (CV) of the GI service process. Experiments in [7], [17] show that this approach, although simple, is prone to large approximation errors.

In [38], Zahorjan et al. obtain an approximate mean value analysis (AMVA) by decomposition-aggregation [14]. The underlying Markov process of the network is decomposed according to the active phases at the GI servers. Each partition is evaluated in isolation by Mean Value Analysis [35] and the results are weighted to approximate the GI network. Validation results of the AMVA decomposition-aggregation show good accuracy.

In [17], Eager et al. improve the results in [34] and [38]. The response time at the GI queue used in Reiser's method is replaced by a more effective interpolation which also accounts for the response time at the other queues. [17] also improves the decomposition method in [38] and makes it compatible with the iterative AMVA framework to achieve lower computational costs on networks with several queues.

Marie's method and the maximum entropy method (MEM) assume a product-form for the equilibrium state probabilities of the GI network and approximate the model accordingly [6]. MEM relies on formulas involving only the mean and the coefficient of variation; Marie's method is more general and uses specialized relations for Coxian distributions. Marie's method provides good accuracy in models with GI servers although its convergence has not been investigated [7].

The diffusion approximation (DA) method has been successfully applied in the analysis of single queues with non-renewal service times [37]. However, in queueing networks it is much harder to determine the equilibrium of the underlying Brownian motion without numerical techniques because product-form formulas exist only under restrictive assumptions [22]. As an additional difficulty, the results of DA hold under heavy-load assumptions that make them accurate only when all queues are in heavy-usage, which may be unrealistic in real systems. Since Brownian motion is a second-order model, it is also impossible to evaluate the impact of higher-order moments of service times.

The Chandy-Herzog-Woo (CHW) method [13], [25] replaces an arbitrary subsystem by a flow equivalent server which preserves the mean throughput of the original subsystem in each feasible state. If the subsystem includes GI servers, CHW is known to be less accurate than Marie's method [7].

Applicability to Nonrenewal Models: To the best of our knowledge, the only analysis of the accuracy of GI approximation methods when applied to closed networks with nonrenewal service has been recently provided in [9]. The discussion in [9] proves that GI approximations either do no apply to networks with nonrenewal workloads, because they completely ignore the temporal dependence between service times, or they are very inaccurate and unable to capture the trend of performance indexes such as utilizations or response times as in the case of decomposition/aggregation [14] and Asymptotic Bound Analysis (ABA) [30]. Related remarks are also given in Section VIII where we illustrate modeling inaccuracies of product-form and GI queueing networks in the evaluation of an e-Commerce system with bursty workloads.

TABLE I
SUMMARY OF MAIN NOTATION

$\alpha_i(n_i)$	service rate scaling for queue i with n_i enqueued jobs
B_j^k	states $(\vec{n},\vec{k})$ where j is busy in phase k
$C_j^k(i)$	mean queue-length of queue i within B_j^k
$\vec{e}_i$	vector of zeros with a one in the i-th position
h,k,u,k^*	phase indexes
i,j,m	queue indexes
I_j^k	states $(\vec{n},\vec{k})$ where j is idle in phase k
$J_j^k(i,h)$	utilization of queue i in phase h within $B_j^k \cup I_j^k$
k_i	active phase at queue i in $\vec{k}$
K_i	number of phases in queue i's MAP
K_{max}	maximum K_i, $1 \leq i \leq M$
$\vec{k}$	phase vector, i.e., active phases
M	number of queues in the network
μ_i	mean service rate of queue i
$\mu_i^{k,h}$	completion rate of queue i, phase $k \to h$
N	number of jobs in the network
n_i	number of jobs at queue i in $\vec{n}$
$\vec{n}$	population vector, i.e., job allocation
$p_{i,j}$	routing prob. from queue i to queue j
$\pi(\vec{n},\vec{k})$	prob. of state $(\vec{n},\vec{k})$
$\pi_j^k(n_i,h)$	prob. of n_i jobs in queue i in phase h within B_j^k
$\pi(n_i,h,n_j,k)$	prob. of n_i jobs in queue i in phase h and n_j jobs in queue j in phase k
$\bar{\pi}_j^k(n_i,h)$	prob. of n_i jobs in i in phase h within I_j^k
$q_{i,j}^{k,h}$	rate $(\vec{n},\vec{k}) \to (\vec{n}-\vec{e}_i+\vec{e}_j,\vec{k}')$, $k_i=k$, $k_i'=h$
$q_{i,j}^{k,h}(n_i)$	rate $q_{i,j}^{k,h}$ from states where queue i has n_i jobs
Q_i	mean queue-length at queue i
Q_i^k	mean queue-length at queue i in phase k
U_i	mean utilization of queue i
U_i^k	mean utilization of queue i in phase k
$v_i^{k,h}$	background trans. rate of queue i, phase $k \to h$
V_i	mean visit ratio at queue i ($V_1=1$)
X	mean throughput (measured at queue $i=1$)

III. MAP Queueing Networks

We introduce the class of MAP queueing networks supporting nonrenewal service which is studied in the rest of the paper. A summary of the main notation is given in Table I.

A. Model Definition

We consider a closed network with single-server queues, which serve jobs according to a MAP service time process and under work-conserving FCFS scheduling. The service process is independent of both the job allocation across the queues and the state of other service processes. The network is composed by M queues and populated by N statistically indistinguishable jobs (single class model), which proceed through the queues according to a state-independent routing scheme. That is, upon departure from a server i, a job joins queue j with fixed probability $p_{i,j}$. Without loss of generality, the average visit ratio at j with respect to the number of visits at queue 1 is V_j, thus $V_1 = 1$.

The service process at queue i is modeled by a MAP with $K_i \geq 1$ phases. General service can be approximated accurately by a MAP [39]. If $K_i = 1$, then the MAP reduces to an exponential distribution, otherwise it generates service time samples that are phase-type (PH) distributed [31]. That is, hyperexponential, hypoexponential, Erlang, and Coxian are all allowed service time distributions; nonrenewal service is also supported, e.g., Markov Modulated Poisson Process (MMPP), Interrupted Poisson Process (IPP) [18]. It should be nevertheless remarked that MAP fitting can be still a challenging problem if the data has an irregular temporal dependence structure, see [23] for a review. We point to [12] for a new technique, called Kronecker Product Composition (KPC), that can provide MAP fitting of higher-order moments and temporal dependence structure of arbitrary processes.

The transition from phase k to phase h for the MAP service process of queue i has rate $\phi_i^{k,h}$ and produces a service completion with probability $t_i^{k,h}$; if $h = k$ then $t_i^{k,k} = 1$ according to the MAP definition. We define $\mu_i^{k,h} = t_i^{k,h}\phi_i^{k,h}$ to be the rate of job completions in phase k that leave the MAP in phase h; $v_i^{k,h} = (1-t_i^{k,h})\phi_i^{k,h}$, $k \neq h$ is the complementary rate of transitions not associated with job completions that only change the MAP active phase (background transitions). In this representation of queue i's MAP, $\mu_i^{k,h}$ is the element in row k

and column h of the $\boldsymbol{D_1}$ matrix; $v_i^{k,h}$ is in row k and column h of $\boldsymbol{D_0}$. We point the reader to [23] and references therein for background on MAPs and MAP fitting.

B. Underlying Markov Process

General MAP service requires to maintain information at the process level on the current service phase at each queue. A feasible network state in the queueing network underlying Markov process is a tuple $(\vec{n}, \vec{k})$, where $\vec{n} = (n_1, n_2, \ldots, n_M)$, $0 \leq n_i \leq N$, $\sum_{i=1}^{M} n_i = N$, describes the number of jobs in each queue, and $\vec{k} = (k_1, k_2, \ldots, k_M)$, $1 \leq k_i \leq K_i$, specifies the active phase for each service process. According to this space, the Markov process transitions have rate $q_{i,j}^{k,h}$ from state $(\vec{n}, \vec{k})$ to $(\vec{n} - \vec{e}_i + \vec{e}_j, \vec{k}')$, $k_i = k$, $k'_i = h$, where $\vec{e}_t$ is a vector of zeros with a one in the t-th position; the rate is given by

$$q_{i,j}^{k,h} = \begin{cases} p_{i,j}\mu_i^{k,h}, & i \neq j, \\ v_i^{k,h} + p_{i,i}\mu_i^{k,h}, & i = j \text{ and } k \neq h. \end{cases} \quad (1)$$

In (1), $q_{i,j}^{k,h}$ is for $i \neq j$ the rate of departures from i to j triggering a phase transition in i's service process from phase k to h; otherwise it accounts for the background transitions $v_i^{k,h}$ and the rate of the self-looping jobs $p_{i,i}\mu_i^{k,h}$. Note that the case for $i = j$ and $k = h$ is not explicitly accounted since it corresponds to the diagonal of the infinitesimal generator of the Markov process. This diagonal is computed to make each row sum to zero.

The size of the infinitesimal generator corresponds to the cardinality of the related global balance equations and it is of the order of $\binom{N+M-1}{N}\binom{K_{max}+M-1}{K_{max}}$, where K_{max} is the maximum of K_i, $1 \leq i \leq M$; this size quickly becomes computationally prohibitive.

As a summarizing example, the MAP network in Figure 1 with routing probabilities $p_{1,1}$, $p_{1,2}$, $p_{1,3} = 1 - p_{1,1} - p_{1,2}$ at the first queue and $p_{2,1} = 1$, $p_{3,1} = 1$, at the remaining queues has underlying Markov process as shown in Figure 2. In the last figure, two queues are exponential with rates $\mu_1 \equiv \mu_1^{1,1}$ and $\mu_2 \equiv \mu_2^{1,1}$, respectively; the third queue is a MAP with $K_3 = 2$ phases having $\mu_3^{k,h} = 0$ for $k \neq h$, that is a MMPP(2) process. The notation, e.g., $(002, 1)$ indicates that the exponential queues are idle and the MAP queue has two jobs and is in phase 1; in $(110, 2)$, the phase 2 is the phase left active by the last served job. For $p_{1,1} = 0.1$ and $p_{1,2} = 0.7$ the network reduces to Balbo's model used in the numerical experiments in [7]; throughout the paper we illustrate some of the proposed techniques using this model.

IV. State Space Reduction

General approximation techniques for non-product-form models, such as decomposition, are reviewed in Section II. These approaches often start from the idea of applying a state space transformation to reduce model complexity. For instance, approximate lumping is used in decomposition to partition the state space into macrostates that can be evaluated in isolation [6].

However, existing state space reductions introduce approximation errors that cannot be bounded in sign or in magnitude.

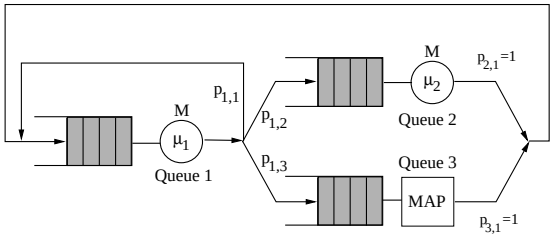

Fig. 1. Example network composed by two exponential queues and a MAP queue.

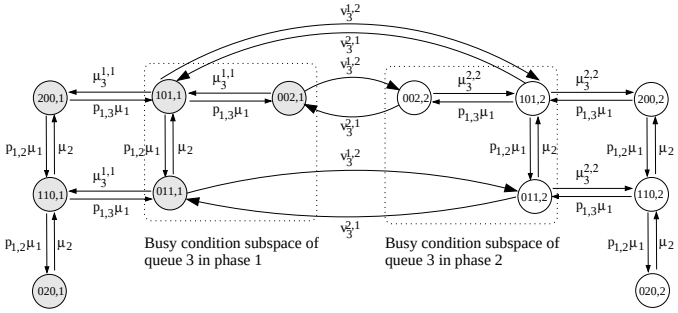

Fig. 2. Underlying Markov process of the network in Figure 1 in the simple case when the MAP is a MMPP(2) process; the job population is $N = 2$.

This leaves a high degree of uncertainty on the final approximation accuracy. In this section we develop a new family of state space reductions that does *not* introduce any degree of approximation, while still simplifies model analysis. The proposed reduction is therefore exact, but because of several differences from exact lumping, the transformation cannot be reduced to lumping or to any method presented in previous work.

A. Busy Condition Reduction

We introduce a state space reduction that scales linearly with the population size. We use the term "busy condition" to identify *the set of states* where a given queue is busy in a certain phase, which is intuitively similar to a conditional state space. For each model we generate the following $O(K_{max}^2 M^2)$ reduced state spaces with dimension $O(N)$ as follows.

Definition 1 (Marginal State Spaces): Let the busy condition subspace $B_j^k = \{(\vec{n}', \vec{k}') : n'_j \geq 1, k'_j = k\}$ be the set of states of the MAP network where queue j is busy and in phase k. The marginal state space *of queue i in phase h within B_j^k is the state space describing the observation within B_j^k of queue i's queue-length while its phase is h, $1 \leq h \leq K_i$, (the cases $i = j$ and $h = k$ are both considered).*

Since in a non-product-form network the state of a queue implicitly depends also on the activity of the rest of the network, the marginal state spaces allow to explore in a compact way the mutual relations between any two queues i and j. A probabilistic definition of marginal state space is given later in Section IV-A1. Two example marginal spaces for the model in Figure 2 obtained for the busy condition subspace B_3^1 are shown in Figure 3. The dashed ovals indicate states in the original state space in Figure 2 that are implicitly

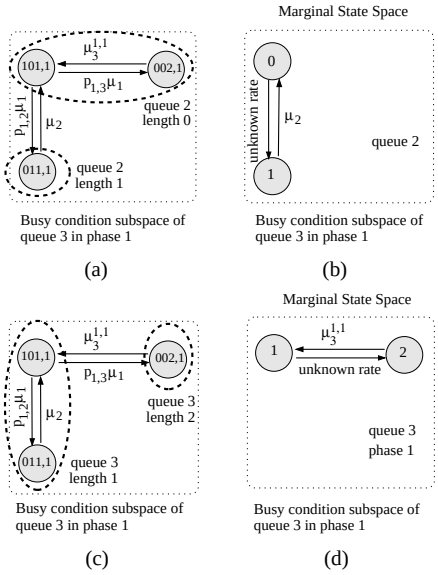

Fig. 3. Example of marginal state spaces for the model in Figure 2

accounted for in the marginal state spaces for queue 2 and for queue 3 in phase 1, depicted in Figures 2(b) and 2(d), respectively. Figures 3(a)-(b) are obtained by observing the exponential queue $i = 2$ in its only phase $h = 1$ within B_3^1. Since queue 3 is always busy in B_3^1, it has queue-length $n_3 \geq 1$ and the queue-length of queue 2 can only be $n_2 = 0$ or $n_2 = 1$. Note that the rate of transitions from $n_2 = 1$ to $n_2 = 0$ depends only on queue 2's service rate μ_2; the rate from $n_2 = 0$ to $n_2 = 1$ depends instead on job completions at the other queues and in the original state space is equal to $\pi(101,1)p_{1,2}\mu_1$ which is unknown[1] without the equilibrium probability $\pi(101,1)$. Figure 3(b) similarly describes the queue-lengths of queue 3 in phase 1 within B_3^1, which can be only $n_3 = 1$ or $n_3 = 2$ since queue 3 is busy. The unknown transition rate is in this case $\pi(101,1)p_{1,3}\mu_1$.

Figure 3 clearly shows that our approach *is not equivalent to* an exact lumping or a decomposition-aggregation for at least three reasons: the latter techniques are applied to the entire state space and not to busy subspaces only, the aggregates are always non-overlapping (two busy subspaces instead can overlap, e.g., $B_{3,1}$ and $B_{2,1}$), and the aggregates result in a reduced state space where *all* rates are known so that it is later analyzed by other techniques (e.g., decomposition solves each macrostate in isolation by global balance or mean value analysis).

The main idea motivating the busy condition reduction is as follows. Even if some rates are unknown, we can obtain balance equations both for the equilibrium inside each marginal space or between the probabilities of multiple marginal spaces. We show in Section V how the busy condition reduction can be used to define the LR performance bounds.

1) Marginal Probabilities: The marginal probability $\pi_j^k(n_i,h)$ of having n_i jobs in queue i during phase h, $1 \leq h \leq K_i$, while queue j is busy in phase k, $1 \leq k \leq K_j$, completely characterizes the marginal state spaces. Each marginal probability can be computed as

$$\pi_j^k(n_i,h) = \sum_{\{(\vec{n}',\vec{k}') \in B_j^k : n_i' = n_i, k_i' = h\}} \pi(\vec{n}',\vec{k}'),$$

where B_j^k is the busy condition subspace of queue j in phase k. By definition, it is $\pi_j^k(n_i = N,h) \equiv 0$ for $i \neq j$, $\pi_j^k(n_j = 0,k) \equiv 0$, $\pi_j^k(n_j,h) \equiv 0$ for $h \neq k$, and $\pi_j^k(n_j,k) \geq \sum_{h=1}^{K_i} \pi_i^h(n_j,k)$ for $j \neq i$, $n_j \geq 1$. The last inequality follows immediately by observing that $\pi_j^k(n_j,k)$ accounts for all states in $\sum_{h=1}^{K_i} \pi_i^h(n_j,k)$ plus the states within B_j^k where i is idle.

Because any event in the underlying Markov process involves at most two-phases and two queues, that is, source and destination queues for job departures with a possible phase transition at the source queue, the marginal probabilities $\pi_j^k(n_i,h)$ still capture all departures and phase changes in the model. Therefore, the knowledge of all $\pi_j^k(n_i,h)$'s is sufficient to compute all mean performance indexes of interest in the original state space, including: the utilization of queue i, i.e., $U_i = \sum_{k=1}^{K_i} U_i^k$, where we denote by U_i^k the utilization of i in phase k, that is

$$U_i^k = \sum_{n_t=0}^{N} \sum_{h=1}^{K_t} \pi_i^k(n_t,h) \tag{2}$$

where t, $1 \leq t \leq M$, is an arbitrary queue since the summation is always equal to the probability of the busy subspace B_i^k; the throughput which by the Utilization Law [25] is

$$X = \sum_{k=1}^{K_1} \sum_{h=1}^{K_1} \sum_{j=1}^{M} q_{1,j}^{k,h} U_1^k = U_1\mu_1/V_1,$$

that is the mean rate of jobs flowing out of queue 1 assumed as reference for network completions and where μ_1 denotes the mean rate of the MAP service process at queue 1; the mean queue-length of queue i is $Q_i = \sum_{k=1}^{K_i} Q_i^k$, with

$$Q_i^k = \sum_{n_i=1}^{N} n_i\, \pi_i^k(n_i,k) \tag{3}$$

being the mean queue-length of i in phase k. Note that these indexes are also sufficient to compute response and residence times by Little's Law, see [25]. In particular, the response time is $R = N/X$.

2) Single Busy Subspace of a Single Queue: We characterize the equilibrium reached at steady state by marginal spaces. We focus on the marginal state spaces which describe a single busy subspace B_j^k and use the population constraint $\sum_{i=1}^{M} n_i = N$. Although an obvious condition, it is impossible to impose it if the state space is transformed in such a way to hide some of the n_i's, as in the marginal state spaces. We therefore define a new population constraint for the busy condition subspace.

Theorem 1: Define

$$C_j^k(i) = \sum_{n_i=1}^{N} \sum_{h=1}^{K_i} n_i \pi_j^k(n_i,h), \tag{4}$$

as the mean queue-length of i in the busy condition subspace B_j^k, thus $C_j^k(j) = Q_j^k$. Then within B_j^k the $C_j^k(i)$ sum to NU_j^k, i.e.,

$$\sum_{i=1}^{M} C_j^k(i) = NU_j^k, \tag{5}$$

$1 \leq k \leq K_j$.

[1]We henceforth assume that global balance solutions for MAP network is prohibitively expensive, therefore the equilibrium probabilities are all unknown.

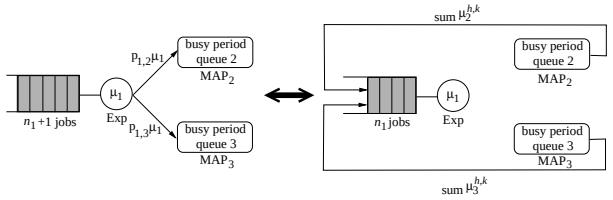

Fig. 4. **Example of marginal balance for a network with an exponential queue with rate μ_1 and two MAP queues.** Job leaving queue 1 enter in the busy period of either queue 2 or queue 3. At steady state, the departing flow is immediately balanced by the incoming flow when queue 1 has n_i jobs.

Proof: Using (2) and the population constraint we have

$$NU_j^k = \sum_{i=1}^{M} n_i \sum_{n_t=0}^{N} \sum_{h=1}^{K_t} \pi_j^k(n_t, h)$$

and choosing the arbitrary queue t equal to i

$$NU_j^k = \sum_{i=1}^{M} \sum_{n_i=1}^{N} \sum_{h=1}^{K_i} n_i \pi_j^k(n_i, h) = \sum_{i=1}^{M} C_j^k(i).$$

■

3) Multiple Busy Subspaces of a Single Queue: We obtain a constraint for multiple busy subspaces which resembles the global balance equations of the MAP service process considered in isolation.

Theorem 2: The utilizations of queue i in its K_i phases are in equilibrium, i.e.,

$$\sum_{j=1}^{M} \sum_{\substack{h=1 \\ h \neq k \text{ if } j=i}}^{K_i} q_{i,j}^{k,h} U_i^k = \sum_{j=1}^{M} \sum_{\substack{h=1 \\ h \neq k \text{ if } j=i}}^{K_i} q_{i,j}^{h,k} U_i^h, \quad (6)$$

for all $1 \leq i \leq M$, $1 \leq k \leq K_i$.

Proof: (Outline of the proof, see [8] for a complete derivation.) Consider the cut separating the group of states $\mathcal{G}_k^i$ where queue i is in phase k from the complementary set of states $\mathcal{C}_k^i$ where queue i is in phase $h \neq k$. The outgoing probability flux from $\mathcal{G}_k^i$ is the left hand side of (6) and must be balanced at steady state by an equal incoming flow generated by the phase change transitions in $\mathcal{C}_k^i$. This probability flux is exactly the right hand side of (6), which completes the proof. ■

The derived equation imposes that the MAP in isolation and the MAP observed in the busy subspaces of queue i have the same stochastic properties, which is expected if the service process of queue i is independent of the job allocation across the network and of the service processes of the other queues.

4) Marginal Balance Conditions: Compared to the previous balances which only involve means such as queue-lengths or utilizations, the balances described in this section, called *marginal balances*, are more informative as they relate individual marginal probabilities.

We have found that there exists a form of partial balance between marginal state spaces, although the class of models considered in this paper is non-product-form. This new class of balances, called *marginal balances*, shows that MAP service imposes an equilibrium between the departure and the arrival process of queue i in groups of states belonging to different busy subspaces. Marginal balance derives from global balance, but characterizes only the set of marginal queue-length probabilities which makes it always computationally tractable. The balance is expressed as follows.

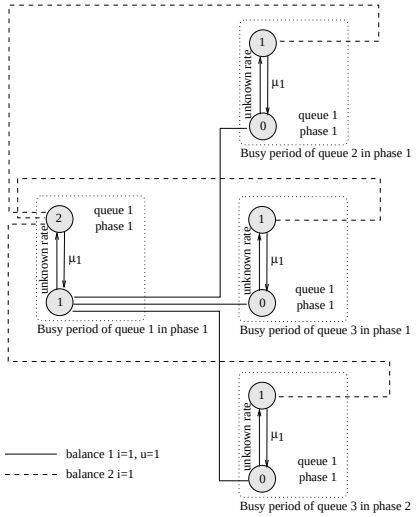

Fig. 5. **Example of marginal balances (7) for the model in Figure 2 and queue $i = 1$.** The illustration for a MAP queue is similar. The departure rate from queue i when its queue-length is $n_i + 1$ is balanced by the arrival rate with queue-length n_i. States connected by identical lines appear in the same marginal balance.

Theorem 3 (Marginal Balance): The arrival rate at queue i when its queue-length is n_i jobs, $1 \leq n_i \leq N-1$, is balanced by the rate of departures when the queue-length is n_i+1, i.e.,

$$\sum_{\substack{j=1 \\ j \neq i}}^{M} \sum_{k=1}^{K_j} \sum_{h=1}^{K_j} \sum_{u=1}^{K_i} q_{j,i}^{k,h} \pi_j^k(n_i, u) = \sum_{\substack{j=1 \\ j \neq i}}^{M} \sum_{k=1}^{K_i} \sum_{h=1}^{K_i} q_{i,j}^{k,h} \pi_i^k(n_i + 1, k), \quad (7)$$

for all $1 \leq i \leq M$. In the case $n_i = 0$ the marginal balance specializes to the more informative relation

$$\sum_{\substack{j=1 \\ j \neq i}}^{M} \sum_{k=1}^{K_j} \sum_{h=1}^{K_j} q_{j,i}^{k,h} \pi_j^k(n_i = 0, u) = \sum_{\substack{j=1 \\ j \neq i}}^{M} \sum_{k=1}^{K_i} q_{i,j}^{k,u} \pi_i^k(n_i = 1, k), \quad (8)$$

which holds for each phase u, $1 \leq u \leq K_i$, with $1 \leq i \leq M$.

Proof: (Outline of the proof, see [8] for a complete derivation.) The statement is a consequence of the state partitioning that separates the states where i has no more than n_i enqueued jobs from the states where the queue-length is at least $n_i + 1$ jobs. Their exchanged probability flux must be balanced at steady state. The flux from the partition for states n_i to the partition for state $n_i + 1$ is equal to the rate of a job completed anywhere in the network being routed to queue i. This is the left hand side of (7), which also accounts for all possible phases of the job's departing queue j and the destination queue i. The opposite flux from $n_i + 1$ to n_i has rate equal to the right hand side of (7), which is the set of all possible departures from i that are not routed to i itself. ■

A general illustration of the marginal balance equations is given in Figure 4. Figure 5 shows the two marginal balance equations for queue 1 in the model of Figure 2; in the case $n_i = 0$ there exists a single balance since the phase of the exponential queue can be only $u = 1$. We recall that product-form local balance does *not* hold in this model, since the MMPP(2) at queue 3 is not an allowed FCFS service process

in the BCMP theorem [4]; therefore, marginal balance holds regardless of local balance. States connected by identical lines appear in the same marginal balance. The busy period B_1^1 corresponds to the right-hand side of (7); the remaining busy periods appear in the left hand side of (7). Although some rates are unknown, it is found that balance exists between the probabilities of the four marginal state spaces. For larger population values, the figure remains qualitatively similar.

Following the proof of the marginal balance conditions, we obtain an additional balance between marginal probabilities.

Corollary 1:
Let k^*, $1 \leq k^* \leq K_i$, be a phase of queue i; the following balance holds for each queue-length n_i, $0 \leq n_i \leq N-2$,

$$\begin{aligned}
&\textstyle\sum_{\substack{j=1\\ j\neq i}}^{M}\sum_{k=1}^{K_j}\sum_{h=1}^{K_j}(q_{j,i}^{k,h}\pi_j^k(n_i+1,k^*)\\
&+\textstyle\sum_{\substack{u=1\\ u\neq k^*}}^{K_i} q_{i,j}^{k,h}\pi_j^k(n_i,u))+\sum_{\substack{k=1\\ k\neq k^*}}^{K_i} q_{i,i}^{k^*,k}\pi_i^{k^*}(n_i+1,k^*)\\
&=\textstyle\sum_{\substack{j=1\\ j\neq i}}^{M}(q_{i,j}^{k^*,k^*}\pi_i^{k^*}(n_i+2,k^*)+\sum_{\substack{k=1\\ k\neq k^*}}^{K_i}(q_{i,j}^{k,k}\pi_i^k(n_i+1,k)\\
&\quad+q_{i,j}^{k,k^*}\pi_i^k(n_i+2,k)+\textstyle\sum_{\substack{h=1\\ h\neq k}}^{K_i} q_{i,j}^{k,h}\pi_i^k(n_i+1,k)))\\
&\qquad+\textstyle\sum_{\substack{k=1\\ k\neq k^*}}^{K_i} q_{i,i}^{k,k^*}\pi_i^k(n_i+1,k), \quad (9)
\end{aligned}$$

for all $1 \leq i \leq M$. For $n_i = N-1$ the balance reduces to

$$\begin{aligned}
&\textstyle\sum_{\substack{k=1\\ k\neq k^*}}^{K_i} q_{i,i}^{k^*,k}\pi_i^{k^*}(n_i+1,k^*)\\
&\qquad+\textstyle\sum_{\substack{j=1\\ j\neq i}}^{M}\sum_{k=1}^{K_j}\sum_{h=1}^{K_j}\sum_{\substack{u=1\\ u\neq k^*}}^{K_i} q_{i,j}^{k,h}\pi_j^k(n_i,u)\\
&=\textstyle\sum_{\substack{j=1\\ j\neq i}}^{M}\sum_{\substack{k=1\\ k\neq k^*}}^{K_i}(q_{i,j}^{k,k}\pi_i^k(n_i+1,k)+\sum_{\substack{h=1\\ h\neq k}}^{K_i} q_{i,j}^{k,h}\pi_i^k(n_i+1,k))\\
&\qquad+\textstyle\sum_{\substack{k=1\\ k\neq k^*}}^{K_i} q_{i,i}^{k,k^*}\pi_i^k(n_i+1,k), \quad (10)
\end{aligned}$$

for all $1 \leq i \leq M$.

Proof: The proof follows similarly to the proof of Theorem 3 by now considering the set of states where i has no more than n_i enqueued jobs except for phase k^*, $1 \leq k^* \leq K_i$, where its population can be no more than n_i+1. The theorem follows imposing the equilibrium at the interface with the set of states where the marginal queue-length is at least n_i+1 and in phase $k \neq k^*$ and at least n_i+2 and in phase k^*. ■

B. Idle Condition Reduction

This state space reduction can be regarded as the complementary of the busy condition reduction described in the previous section. We consider the idle condition subspace I_j^k where queue j is empty and the last served job has left the MAP process at j in phase k, $1 \leq k \leq K_j$. We obtain a set of $O(K_{max}M^2)$ reduced state spaces with dimension $O(N)$ by describing the evolution within I_j^k of the queue-length of i during phase h, $1 \leq h \leq K_i$. The related marginal probability function is

$$\bar{\pi}_j^k(n_i,h) = \textstyle\sum_{(\vec{n}',\vec{k}')\in\bar{S}_j^k(n_i,h)}\pi(\vec{n}',\vec{k}'), \quad (11)$$

where the marginal space is $\bar{S}_j^k(n_i,h) = \{(\vec{n}',\vec{k}') \in I_j^k : n_i' = n_i, k_i' = h\}$, the idle subspace is $I_j^k = \{(\vec{n},\vec{k}) : n_j = 0, k_j = k\}$. Further, by the given definitions, $\bar{\pi}_j^k(n_j,h) \equiv 0$ if $n_j \geq 1$ or $h \neq k$ and similarly to the busy condition reduction $\pi_j^k(n_j,k) \geq \sum_{h=1}^{K_i}\bar{\pi}_i^h(n_j,k)$ for $j \neq i$, $n_j \geq 1$. Note that from the complementarity of $\pi_j^k(n_i,h)$ and $\bar{\pi}_j^k(n_i,h)$, the total state space probability is immediately obtained as

$$\textstyle\sum_{h=1}^{K_i}\sum_{n_i=0}^{N}(\pi_j^k(n_i,h)+\bar{\pi}_j^k(n_i,h)) = 1, \quad (12)$$

for all $1 \leq i \leq M$. Moreover, let the utilization of queue i in phase h within $B_j^k \cup I_j^k$ be

$$J_j^k(i,h) = \textstyle\sum_{n_i=1}^{N}(\pi_j^k(n_i,h)+\bar{\pi}_j^k(n_i,h)). \quad (13)$$

where by definition the second term in the summation may be rewritten as

$$\textstyle\sum_{n_i=1}^{N}\bar{\pi}_j^k(n_i,h) = \pi_i^h(n_j=0,k), \quad (14)$$

which similarly to (12) relates the busy and idle reductions.

Balances similar to those given for the busy condition reduction can be derived for the idle time reduction. For instance, following the proof of (5) one immediately obtains the population constraint

$$\textstyle\sum_{i=1}^{M}\bar{C}_j^k(i) = N\bar{\pi}_j^k(n_j=0,k), \quad (15)$$

where $\bar{\pi}_j^k(n_j=0,k)$ is the probability of I_j^k and

$$\bar{C}_j^k(i) = \textstyle\sum_{n_i=1}^{N}\sum_{h=1}^{K_i} n_i\bar{\pi}_j^k(n_i,h) \quad (16)$$

is the mean queue-length of i in phase h within I_j^k.

The balance equations obtained for the idle reduction are often redundant with the balances of the busy ones. Therefore, we are not interested in developing a comprehensive characterization of this reduction. We point out two relations deriving from manipulations of the global balance equations which characterize $B_j^k \cup I_j^k$ where j is in phase k; these formulas cannot be expressed within the probability space of the busy subspace only.

Theorem 4: The sum of mean queue-lengths during the subspace $B_i^k \cup I_i^k$ satisfies

$$\textstyle\sum_{t=1}^{M}(C_k^j(t)+\bar{C}_k^j(t)) \geq N\sum_{h=1}^{K_i} J_j^k(i,h), \quad (17)$$

for all $1 \leq i \leq M$, $1 \leq j \leq M$, $1 \leq k \leq K_j$.

Proof: Letting $\sum_{B_j^k\cup I_j^k} \equiv \sum_{(\vec{n},\vec{k})\in B_j^k\cup I_j^k}$, we have

$$\begin{aligned}
N\textstyle\sum_{B_j^k\cup I_j^k}\pi(\vec{n},\vec{k}) &= \textstyle\sum_{t=1}^{M}\sum_{B_j^k\cup I_j^k} n_t\pi(\vec{n},\vec{k})\\
&= \textstyle\sum_{t=1}^{M}(\sum_{B_j^k} n_t\pi(\vec{n},\vec{k})+\sum_{I_j^k} n_t\pi(\vec{n},\vec{k}))\\
&= \textstyle\sum_{t=1}^{M}(C_k^j(t)+\bar{C}_k^j(t)),
\end{aligned}$$

where the last passage follows by definition of $C_k^j(t)$ and $\bar{C}_k^j(t)$ as mean queue-lengths in B_j^k and I_j^k. Starting from the same term we also have

$$N\textstyle\sum_{B_j^k\cup I_j^k}\pi(\vec{n},\vec{k}) \geq N\sum_{h=1}^{K_i} J_j^k(i,h)$$

since the utilization of any queue i, $1 \leq i \leq M$, during $B_j^k \cup I_j^k$ cannot be greater than the sum of the probabilities of all states of $B_j^k \cup I_j^k$. ■

LR lower bound

$f_{min} = \min\ f(\boldsymbol{\pi})$
subject to:
/* *preliminary definitions* */
eq. (2),(3),(4),(13),(16);
$C_j^k(j) = Q_j^k$;
$\pi_j^k(n_j,k) \geq \sum_{h=1}^{K_i} \pi_i^h(n_j,k)$, if $n_j \geq 1$, $i \neq j$;
$\pi_j^k(n_j,k) \geq \sum_{h=1}^{K_i} \bar{\pi}_i^h(n_j,k)$, if $n_j \geq 1$, $i \neq j$;
$\pi_j^k(n_j,h) = 0$, if $n_j = 0$;
$\pi_j^k(n_j,h) = 0$, if $h \neq k$;
$\pi_j^k(n_i,h) = 0$, if $n_i = N$, $i \neq j$;
$\bar{\pi}_j^k(n_j,h) = 0$, if $n_j \geq 1$;
/* *exact characterization* */
eq. (5), (6), (7), (8), (9), (10), (15), (17), (18);
/* *reduction constraints* */
eq. (12), (14);
/* *feasibility of results* */
$\pi_j^k(n_i,h) \geq 0$, for all $\pi_j^k(n_i,h) \in \boldsymbol{\pi}$.
$\bar{\pi}_j^k(n_i,h) \geq 0$, for all $\bar{\pi}_j^k(n_i,h) \in \boldsymbol{\pi}$.

Fig. 6. Linear program determining a lower bound on an arbitrary linear performance index $f_{exact} = f(\boldsymbol{\pi}_{exact})$.

Theorem 5: The performance indexes in busy and idle subspaces are related by the following equation

$$\sum_{\substack{h=1,\\ h\neq k}}^{K_i} \sum_{j=1}^{M} q_{i,j}^{k,h} Q_i^k + \sum_{\substack{j=1,\\ j\neq i}}^{M} \sum_{h=1}^{K_i} q_{i,j}^{h,k} U_i^h = \sum_{\substack{j=1,\\ j\neq i}}^{M} \sum_{h=1}^{K_j} \sum_{u=1}^{K_j} q_{j,i}^{h,u} J_i^k(j,h) + \sum_{\substack{h=1,\\ h\neq k}}^{K_i} \sum_{j=1}^{M} q_{i,j}^{h,k} Q_i^h, \tag{18}$$

for all $1 \leq i \leq M$, $1 \leq k \leq K_i$.

Proof: (Outline of the proof, see [8] for a complete derivation.) The proof follows similarly to that of Theorem 2 by weighting the contribution of each group of states by n_i. We point to the technical report [8] for an extensive derivation. ■

V. Linear Reduction Bounds

We obtain the LR bounds using the results for the busy and the idle condition reductions. We determine the values of the marginal probabilities

$$\boldsymbol{\pi} = \{\pi_j^k(n_i,h),\ \forall\ i,j,k,h,n_i\} \cup \{\bar{\pi}_j^k(n_i,h),\ \forall\ i,j,k,h,n_i\}$$

so that the linear function $f(\boldsymbol{\pi})$ is a bound on a performance metric $f_{exact} \equiv f(\boldsymbol{\pi}_{exact})$, where $\boldsymbol{\pi}_{exact}$ is the set of exact equilibrium probabilities of the MAP network. In the case of lower bounds $f_{min} \leq f_{exact}$, the values of the marginal probabilities in $\boldsymbol{\pi}$ can be determined using linear programming [5] as follows.

Proposition 1 (LR Lower Bound): The program in Figure 6 returns a lower bound $f_{min} \leq f(\boldsymbol{\pi}_{exact})$.

Proof: All the relations in the linear program are exact as we have proved in the previous sections; therefore $\boldsymbol{\pi} = \boldsymbol{\pi}^{exact}$ is a feasible solution. Since linear programming always returns an optimum

$$\min\ f(\boldsymbol{\pi}) = \min\{f(\boldsymbol{\pi}) |\ \text{feasible}\ \boldsymbol{\pi}\},$$

we conclude that $\min\ f(\boldsymbol{\pi}) \leq f(\boldsymbol{\pi}^{exact})$ because $\boldsymbol{\pi}^{exact}$ is a feasible value of $\boldsymbol{\pi}$. ■

The last proposition generalizes immediately if the linear program is reformulated to compute an upper bound (LR Upper Bound) $f_{max} = \max f(\boldsymbol{\pi}) \geq f(\boldsymbol{\pi}_{exact})$; therefore the same constraints in Figure 6 can be used both for upper and lower bounds and only the objective function has to be modified.

The computational costs of the LR technique are indeed feasible for practical applications, e.g., we have solved the linear program for a model with ten MAP(2) queues and $N = 50$ jobs using an interior point solver in approximately four minutes; for $N = 100$ the solution of the same model is found in approximately ten minutes suggesting good scalability. In general, the complexity of computing bounds with the linear program in Figure 6 grows as $O(\mathrm{lp}(M^2K_{max} + MN, K_{max}^2M^2N))$, where $\mathrm{lp}(r,c)$ is the computational cost of solving a linear program with r rows and c columns. The number of rows is either dominated by the number of possible marginal balances for the case $n_i \geq 1$ that is $O(MN)$ or by the number of inequalities (17) which grows as $O(M^2K_{max})$; the number of columns is $O(K_{max}^2M^2N)$ because the cardinality of $\boldsymbol{\pi}$ is upper bounded by $2K_{max}^2M^2N$.

TABLE II
Linear and Quadratic Reduction Effectiveness

M	N	K_{max}	total states marginal spaces LR	total states marginal spaces QR	total states original space
3	50	2	$1.84 \cdot 10^3$	$9.36 \cdot 10^4$	$5.30 \cdot 10^3$
3	100	2	$3.64 \cdot 10^3$	$3.67 \cdot 10^5$	$2.06 \cdot 10^4$
3	200	2	$7.24 \cdot 10^3$	$1.45 \cdot 10^6$	$8.12 \cdot 10^4$
5	50	2	$5.10 \cdot 10^3$	$2.60 \cdot 10^5$	$1.27 \cdot 10^6$
5	100	2	$1.01 \cdot 10^4$	$1.02 \cdot 10^6$	$1.84 \cdot 10^7$
5	200	2	$2.01 \cdot 10^4$	$4.04 \cdot 10^6$	$2.80 \cdot 10^8$
10	50	2	$2.04 \cdot 10^4$	$1.04 \cdot 10^6$	$5.03 \cdot 10^{10}$
10	100	2	$4.04 \cdot 10^4$	$4.08 \cdot 10^6$	$1.71 \cdot 10^{13}$
10	200	2	$8.04 \cdot 10^4$	$1.62 \cdot 10^7$	$7.04 \cdot 10^{15}$

To appreciate the reduction of the state space, Table II compares the number of states in the LR marginal state spaces with the original state space size in models with larger population and number of queues. The column with results for the QR transformation is discussed in Section VI-A. In the table, all queues have MAP service times with K_{max} phases. The number of states in the LR marginal state spaces grows linearly in the population size, whereas the growth for the original state space is combinatorial. Here, the reduced spaces have cardinality that can be several orders of magnitude smaller than the original state space.

A. Discussion

The balances obtained in Section IV provide a quite accurate characterization of the underlying Markov process of the MAP

network. However, the number of exact relations remains much smaller than the number of the marginal probabilities $\pi_j^k(n_i, h)$ and $\bar{\pi}_j^k(n_i, h)$. We stress that our exact characterization is in general under-determined and describes a *family* of possible equilibria for the underlying Markov process, among which we cannot distinguish the correct one. The linear programming approach we have adopted selects the equilibrium that provides a worst-case or best-case bound on a given performance metric.

Because of the complexity of the feasibility region described by (2)-(18), it is also very hard to establish the relative importance of each equation with respect to the others, as well as determining analytical linear independence conditions among the balance equations. In our experiments, we have frequently observed that removing either equation (7) or (18) reduces significantly the quality of the bounds. Conversely, we have found that (9) and (10) improve accuracy only on certain models. Standard sensitivity analysis of linear programs [5] may be used as a tool for investigating the relative importance of a certain balance equation for the model under study.

VI. Analysis of Load-Dependent Models

An important generalization of the LR bounding technique is the analysis of models including load-dependent servers, i.e., resources that dynamically change their rate of service according to the number of enqueued jobs. This extension is fundamental to use in MAP queueing networks the following types of resources:

- *delay servers*, i.e., resources where jobs receive immediately service without queueing. Delay servers are important to model a customer's *think time* before submitting a new request to the system. It is a classic result that a delay server with exponential service times can be modeled as a queue where the service rate grows linearly with the number of enqueued jobs [25]. We use this property to support the inclusion of delay servers with exponential service times in MAP queueing networks.
- *load-dependent devices*, i.e., resources that model certain physical devices where service times are affected by the current queue-length size. For instance, disk drives can be approximated as load-dependent stations [32].
- *flow equivalent servers*, i.e., service centers that abstract a sub-network of several queues by reproducing its mean throughput as a function of the sub-network population [13]. It is easy to find examples showing that, for MAP queueing networks, flow equivalent aggregation is in general an approximate method, i.e., replacing some exponential queues of a MAP queueing network with a flow equivalent server affects the value of the mean performance indexes.

The remainder of this section is organized as follows. In Section VI-A we discuss how the exact balance equations found in Section IV generalize to the load-dependent case. In particular, we argue that a more general reduction of the state space, where the number of states grows quadratically with the population N, is required to reduce the model size while preserving the exactness of the representation. This finding is consistent with previous work on product-form networks, where the computational costs of load-dependent models grow quadratically as $O(N^2)$ compared to the linear $O(N)$ complexity of the constant-rate case [25]. Therefore, we define the Quadratic Reduction (QR) bounds resulting from this new description of the state space and illustrate their accuracy on case studies.

A. Load-Dependent MAP Queueing Networks

MAP queueing networks immediately generalize to the load-dependent case by letting service rates to be a function of the queue-length. Define $q_{i,j}^{k,h}(n_i)$ as the rate of transitions from state $(\vec{n}, \vec{k})$ to $(\vec{n} - \vec{e}_i + \vec{e}_j, \vec{k}')$, where the current queue-length of station i is n_i and $\vec{k}$ and $\vec{k}'$ differ only for $k_i = k$, $k_i' = h$. This rate can be computed as

$$q_{i,j}^{k,h}(n_i) = \alpha_i(n_i) q_{i,j}^{k,h}, \tag{19}$$

where $q_{i,j}^{k,h}$ is defined as in (1) and $\alpha_i(n_i)$, $1 \leq n_i \leq N$, is a user-specified scaling function for the service rate of station i when its queue-length size is n_i. According to this definition, if $n_i = 1$, for all stations $1 \leq i \leq M$, then all service processes are constant-rate and this case reduces to the class of MAP networks considered in the previous sections; otherwise, the model is a load-dependent MAP queueing network. In particular, if $q_{i,j}^{k,h}(n_i) = n_i \mu_i$, and $K_i = 1$, then the service times are exponentially distributed with rate growing linearly with the queue-length size; thus, resource i becomes a delay server [25].

In order to show that the balance equations found in Section IV do *not* generalize to the load-dependent case without a major change of the state space reduction, consider for example the marginal balance (7). The left-hand side of (7) captures departures from queue j using the term $q_{j,i}^{k,h} \pi_j^k(n_i, u)$. Since the queue-length n_j does not appear explicitly in $\pi_j^k(n_i, u)$, in order to replace $q_{j,i}^{k,h}$ with $q_{j,i}^{k,h}(n_j)$ we first need to express also n_j. This is not possible with the linear reduction described in the previous sections which does not allow to jointly express n_i and n_j and thus requires a more general class of marginal state spaces.

Definition 2: The quadratic marginal state space *of queue i in phase h and queue j in phase k is the state space describing the* joint *observation of queue i's queue-length while its phase is h, $1 \leq h \leq K_i$, and queue j's queue-length while its phase is k, $1 \leq k \leq K_j$. The cardinality of the quadratic marginal state spaces grows as $O(N^2 K_{max}^2)$ and the underlying marginal probability function is*

$$\pi(n_j, k, n_i, u) = \sum_{\substack{(\vec{n}', \vec{k}'): n_j' = n_j, n_i' = n_i, \\ k_j' = k, k_i' = u}} \pi(\vec{n}', \vec{k}'), \tag{20}$$

subject to the symmetry constraint $\pi(n_i, k, n_j, u) = \pi(n_j, u, n_i, k)$, for all $0 \leq n_i \leq N$, $0 \leq n_j \leq N$, $1 \leq k \leq K_i$, $1 \leq u \leq K_j$.

A comparison of the total number of states in the marginal state spaces generated by the quadratic reduction with the original state space and the linear reduction is given in Table II. The table indicates that the QR transformation, although more expensive than the LR transformation, is still highly scalable

with respect to the number of states in the original Markov process.

The probabilities $\pi(n_j, k, n_i, u)$ allow the generalization of the exact balance equations found in Section IV to the load-dependent case. The resulting formulas are obtained using replacement rules of the type:

$$q_{i,j}^{k,h}\pi_i^k(n_j,u) \quad \rightarrow \quad \sum_{n_i=1}^{N} q_{i,j}^{k,h}(n_i)\pi(n_i,k,n_j,u), \quad (21)$$

$$q_{j,i}^{k,h}\pi_i^u(n_j,k) \quad \rightarrow \quad \sum_{n_j=1}^{N} q_{j,i}^{k,h}(n_j)\pi(n_i,u,n_j,k), \quad (22)$$

$$q_{j,i}^{k,h}\bar{\pi}_i^u(n_j,k) \quad \rightarrow \quad q_{j,i}^{k,h}(n_j)\pi(n_i=0,u,n_j,k), \quad (23)$$

$1 \leq k, h, u \leq K_i$, which derive from the fact that in theorem proofs the rates $q_{i,j}^{k,h}$ cannot be factored out of the summations if these rates are load-dependent. For example, using the replacement rules, the marginal balance (7) becomes in the load-dependent case

$$\sum_{\substack{j=1\\ j\neq i}}^{M}\sum_{k=1}^{K_j}\sum_{h=1}^{K_j}\sum_{u=1}^{K_i}\sum_{n_i=1}^{N} q_{j,i}^{k,h}(n_j)\pi(n_j,k,n_i,u)$$
$$= \sum_{\substack{j=1\\ j\neq i}}^{M}\sum_{k=1}^{K_i}\sum_{h=1}^{K_i} q_{i,j}^{k,h}(n_i+1)\pi(n_i+1,k,n_i+1,k),$$

for $1 \leq i \leq M$, $1 \leq n_i \leq N-1$. Using the same approach, the balance (18) generalizes similarly as

$$\sum_{\substack{h=1,\\ h\neq k}}^{K_i}\sum_{j=1}^{M}\sum_{n_i=1}^{N} q_{i,j}^{k,h}(n_i)n_i\pi(n_i,k,n_i,k) +$$
$$\sum_{\substack{j=1,\\ j\neq i}}^{M}\sum_{h=1}^{K_i}\sum_{n_i=1}^{N} q_{i,j}^{h,k}(n_i)\pi(n_i,h,n_i,h)$$
$$= \sum_{\substack{j=1,\\ j\neq i}}^{M}\sum_{h=1}^{K_j}\sum_{u=1}^{K_j}\sum_{n_j=1}^{N} q_{j,i}^{h,u}(n_j)\sum_{ni=0}^{N}\pi(n_i,k,n_j,h)$$
$$+ \sum_{\substack{h=1,\\ h\neq k}}^{K_i}\sum_{j=1}^{M}\sum_{n_i=1}^{N} q_{i,j}^{h,k}(n_i)n_i\pi(n_i,h,n_i,h),$$

for all $1 \leq i \leq M$, $1 \leq k \leq K_i$. The expression of the other load-dependent balance equations are obtained similarly to the two extensions illustrated above.

The replacement rules also show that in the quadratic state space we do not need to distinguish between busy and idle condition, since these are now immediately determined by the state of queue i in $\pi(n_i, k, n_j, u)$. That is, the subset of values $n_i = 0$ refers to the idle condition, the subset $n_i \geq 1$ refers to the busy condition of queue i. Therefore, some of the equations used in the LR bounds to impose consistency between the busy and idle condition equations are also unnecessary in the load-dependent analysis, e.g., (14) is no longer needed.

A further advantage of the quadratic state space reduction is that the increased detail allows to formulate balance equations that cannot be expressed with the linear state space reduction. These additional equations make bounds defined on the quadratic marginal state spaces always slightly more accurate than the LR bounds.

Theorem 6: The quadratic state space reduction satisfies the following second-order population constraint:

$$\sum_{i,j=1}^{M}\sum_{n_i,n_j=1}^{N}\sum_{h=1}^{K_i}\sum_{k=1}^{K_j} n_i n_j \pi(n_i,h,n_j,k) = N^2. \quad (24)$$

Proof: Consider the summation

$$S = \sum_{(\vec{n},\vec{k})}(n_1 + n_2 + \ldots, +n_M)^2\pi(\vec{n},\vec{k}).$$

QR lower bound

$f_{min} = \min\ f(\boldsymbol{\pi})$
subject to:
/* *preliminary definitions* */
eq. (2),(3),(4),(13),(16);
$\sum_{n_j=0}^{N}\sum_{k=1}^{K_j}\pi(n_j,k,n_j,k) = 1$, for all $1 \leq j \leq M$;
$\pi(n_j,k,n_i,h) = 0$, if $i = j$, $n_i = n_j$, $h \neq k$;
$\pi(n_j,k,n_i,h) = 0$, if $i = j$, $n_i \neq n_j$;
$\pi(n_j,k,n_i,h) = 0$, if $i \neq j$, $n_i + n_j > N$;
$\pi(n_j,k,n_i,h) = \pi(n_i,h,n_j,k)$, for all n_i, n_j, h, k;
$\pi(n_j,k,n_j,k) = \sum_{n_i=0}^{N-n_j}\sum_{h=1}^{K_i}\pi(n_j,k,n_i,h)$, if $i \neq j$;
/* *exact characterization* */
eq. (24), (5), (6), (7), (8), (9), (10), (15), (17), (18);
/* *feasibility of results* */
$\pi(n_j,k,n_i,h) \geq 0$, for all $(n_j,k,n_i,h) \in \boldsymbol{\pi}$.

Fig. 7. Linear program defining the QR lower bound

Since for all states $n_1 + n_2 + \ldots, +n_M = N$, we have immediately

$$S = \sum_{(\vec{n},\vec{k})} N^2\pi(\vec{n},\vec{k}) = N^2\sum_{(\vec{n},\vec{k})}\pi(\vec{n},\vec{k}) = N^2.$$

However, expanding $(n_1 + n_2 + \ldots, +n_M)^2$ we have also

$$S = \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{(\vec{n},\vec{k})} n_i n_j \pi(\vec{n},\vec{k}),$$

which can be decomposed into the left hand side of (24). ■

The generalized balances developed above allow the definition of a new class of upper and lower performance bounds as shown in the next section.

B. Quadratic Reduction (QR) Bounds

Based on the exact characterization of load-dependent MAP queueing networks discussed above, we define the *Quadratic Reduction (QR) bounds* with a linear programming approach similar to the LR bounds. The linear program used for the QR bounds is shown in Figure 7; here, e.g., the reference to (7) refers to the generalization of (7) obtained by applying the replacement rules for the load-dependent case given in the previous section. The main changes with respect to the LR lower bound in Figure 6 are the different preliminary definitions, the addition of (24), and the removal of the reduction constraints (12) and (14). All equations numbered from (2) to (18) should be first generalized to the load-dependent case using the replacement rules given in Section VI-A. Due to the increased number of probability terms, the computational cost of the QR bounds is of the order of $O(lp(M^2K_{max}^2N^2, M^2K_{max}^2N^2))$, where the dominant cost is enforcing all symmetry constraints $\pi(n_i,k,n_j,u) = \pi(n_j,u,n_i,k)$.

VII. Accuracy Validation

We assess the accuracy of the LR and QR bounds using the following validation methodology. We consider both randomly-generated models and representative case studies, see Table III for a description of random model parameters. We evaluate the maximal relative error of response time

TABLE III
INPUT PARAMETERS USED IN THE GENERATION OF RANDOM MODELS

Network	*Value*	*MAP(2)*	*Value*
M	3	mean	random in $[0, 1]$
$p_{i,j}$	random in $[0, 1]$	CV	random in $[0.5, 10]$
N	all in $[10, 1000]$	skewness	random in $[2, 250]$
# of MAPs	1	γ_2	random in $[.00, .99]$

bounds with respect to the exact global balance solution of the MAP queueing network. Due to the state space explosion, the experimentation using exact global balance solutions is prohibitive for MAP networks with four or more queues and population $N \geq 100$. Given its mean, CV, skewness, and autocorrelation decay rate γ_2, a MAP(2) is generated using the exact moment and autocorrelation matching formulas in [12]. For each random queueing network model, we use the linear programs in Figures 6-7 to compute upper and lower LR and QR bounds X_{max} and X_{min} on the mean throughput $f(\boldsymbol{\pi}) = X$. Then, using Little's Law, we get the response time bounds $R_{min} = N/X_{max}$ and $R_{max} = N/X_{min}$ which are used to compute absolute relative errors with respect to the exact response time R. We do not report errors on other measures due to lack of space, but we remark that they are typically in the same range as those of response time. The validation has used the GNU Linear Programming Kit [21] to solve the linear programs of LR and QR bounds on an Intel Xeon 3.73GHz starting from an AMPL specification [19]. AMPL specifications of the proposed bounds are available for download at [16].

A. Random Models

LR Bounds. In order to evaluate the general quality of the LR bounds, we have evaluated $10,000$ random models. The models are generated drawing random numbers from a uniform probability distribution and according to the specifications in Table III. Each random model is solved for all possible populations values $N \in [10, 1000]$ and the following absolute value of the maximal relative error is computed

$$\Delta_{bnd} = \max_{N} \left| \frac{R_{bnd}(N) - R_{exact}(N)}{R_{exact}(N)} \right|,$$

where $R_{exact}(N)$ is the response time of the exact solution computed for the network considered with population N and $R_{bnd}(N)$ is the LR bound evaluated with the same population, either $R_{max}(N)$ or $R_{min}(N)$. We stress that the Δ_{bnd} error function is a conservative estimator since it returns the *maximum* error of R_{bnd} over all evaluated populations. Thus, although we have often observed the convergence of the LR bounds to the exact response time value for large N, this is accurate asymptotic behavior is not accounted by the Δ_{bnd} metric and only the worst case error is measured.

TABLE IV
RESULTS OF RANDOM EXPERIMENTS

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

Table IV indicates that the LR bounds perform extremely well. The table reports absolute maximal relative error ($0 \equiv 0\%$, $1 \equiv 100\%$) over $10,000$ random MAP queueing networks for the response time $R = N/X$ (R_{min}=lower LR bound, R_{max}=upper LR bound). The mean error is $1\% - 2\%$ for both bounds with a small standard deviation; the median is less than the mean, indicating that the asymmetry of the error distribution is more concentrated on small errors. The maximum error is found to be 14.2% for the response time upper bound and 12.6% for the lower bound. We have inspected carefully these cases and found that models with more than 10% error in at least one of the two bounds account for only 1% of the total number of experiments. Furthermore, in these models, the LR lower bound seems to deteriorate by high variability and burstiness, while the worst case error of the LR upper bound is found for MAPs with low or moderate burstiness. The difference in sensitivity to MAP parameters is a positive property of the LR bounds, as inaccuracies in one bound can be compensated by the accuracy of the other bound.

In a preliminary version of this paper [9], we have reported a complete sensitivity analysis of LR bounds with respect to the parameters considered in Table III. The results in [8], [9] indicate that the LR bounding methodology is very robust with respect to perturbations of the MAP distribution and burstiness characteristics. Changes in the routing probabilities have limited impact on the maximum error too and the worst case is found to be in models with balanced routing. Finally, [9] also reports sensitivity experiments on queueing network models with larger number of MAPs or larger number of queues.

QR Bounds. In order to validate the relative accuracy of the QR bounds with respect to the LR bounds, we have solved the same random models considered above by interpreting them as load-dependent models. This can be done easily by setting the scaling factor $\alpha_i(n) = 1$ for all queues i and queue-length values $1 \leq n \leq N$. The numerical results of these random experiments are essentially identical to those shown in Table IV: the maximum errors of the upper and lower QR bounds remain 14.1% and 12.6%, respectively. Also the other statistical indicators in Table IV change by less than 0.01%. These results suggest that using the QR bounds does not imply increased accuracy with respect to the LR bounds, although the computational costs of the former are much higher than those of the latter. Nevertheless, the QR transformation remains the only feasible technique for load-dependent queueing networks where the LR bounds cannot be applied. This justifies the use of QR bounds.

B. Balbo's Model

Following the last observation in the previous subsection, we now focus on the validation of QR bound accuracy on a case study of load-dependent MAP queueing networks. Here, we do not consider random load-dependent models because random scaling factors $\alpha_i(n_i)$ are hardly representative of

a real system behavior, i.e., load-dependency tends to have regular analytical shapes (e.g., "inverted U-shapes" in models of memory thrashing phenomena [33]).

We use the QR bounds to evaluate the example network given in Figure 1 when this is augmented with a delay server that is placed on the feedback loop of queue 1. Because of the presence of a delay server, the model is load-dependent with $\alpha_4(n_4) = n_4$, where the index associated to the delay server is 4. Similarly to Balbo's example model, the routing probability from queue 1 to queue 2 is $p_{1,2} = 0.7$, from queue 1 to queue 3 is $p_{1,3} = 0.2$ while the probability of entering into the feedback loop of the delay server is $p_{1,4} = 0.1$. Since the feedback loop now includes the delay server, it is always $p_{1,1} = 0$ and a job completed at the delay server re-enters the system by first joining queue 1. We use this configuration to evaluate the accuracy of the QR bounds as a function of the number of jobs N and of the mean think time Z at the delay server. Because of the prohibitive cost of the exact evaluation of a queueing network with four stations and tens of jobs, we use the relative gap between the upper and lower bounds as a descriptor of the QR bounds accuracy.

The accuracy results of QR bounds on Balbo's model augmented with the delay server are given in Table V. The metric "bnd gap" indicates the relative error $R_{max}/R_{min} - 1$, between upper and lower response time bounds. The table reports a sensitivity analysis with respect to N and Z: N ranges between 10 and 50, and Z is instead defined by its ratio to the mean service time μ_3^{-1} of the bottleneck resource, i.e., queue 3, and ranges between $0.5\mu_3^{-1}$ and $5\mu_3^{-1}$. In this model, low values of Z indicate networks where jobs tend to queue at the bottleneck resource, queue 3, and where the system behaves not too differently from a model with constant-rate servers only. Conversely, when Z is large, jobs tend to spend most of their cycle time at the delay server waiting to join the queues; this makes the model much more complex than constant-rate MAP queueing networks.

The results in Table V for increasing values of Z and N indicate that the good accuracy of the QR bounds is consistent with the results of the random model validation, showing a maximum gap between upper and lower QR bounds of about 12%. The results also indicate that QR bound accuracy is affected differently by changes in Z or N: for increasing values of the think time Z, the QR bound accuracy decreases, which is an issue that appears also in the approximation of product-form models and which is attributed to the difficulty of estimating the number of active jobs circulating outside the delay server. Conversely, for increasing values of the total population N, the QR bound accuracy increases, dropping to an error of just $1.5 - 2\%$ for medium congestion levels ($N = 50$).

Summarizing, this case study based on the analytical model in [7] indicates that QR bounds provide good accuracy levels for different choices of the think time Z and of the number of jobs N. The most important observation is that the QR bounds accuracy grows quickly as the number of jobs N increases. This is a nice property of the proposed bounds, because evaluating system performance under burstiness conditions is meaningful only if there are enough requests to create large fluctuations in the system utilization and queue-lengths: this often requires large populations.

TABLE V
BALBO'S MODEL: QR BOUND ACCURACY ON RESPONSE TIMES

Z/μ_3^{-1}	N	bnd gap	Z/μ_3^{-1}	N	bnd gap
0.5	10	5.1%	3	10	7.5%
0.5	25	2.3%	3	25	3.1%
0.5	50	1.4%	3	50	1.8%
1	10	6.2%	4	10	8.8%
1	25	2.7%	4	25	3.6%
1	50	1.6%	4	50	2.1%
2	10	6.8%	5	10	12.2%
2	25	2.8%	5	25	4.5%
2	50	1.7%	5	50	2.5%

VIII. ANALYSIS OF TPC-W E-COMMERCE SYSTEM

In this section, we present an application of MAP queueing networks to the performance analysis of an enterprise application running on a multi-tier architecture. We consider the e-commerce system studied in [10] subject to a TPC-W e-commerce benchmark [20], which simulates the operations of an online bookstore. The e-commerce system 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. We have used the following experimental setup: the Web server and the application server are installed on the same front server, a 1-way 3.2 GHz Pentium D; the database resides on a 2-way 3.2 GHz Pentium D, and the incoming TPC-W requests depart from two 2-way 3.2 GHz Pentium D machines. An illustration of the multi-tier architecture is given in Figure 8.

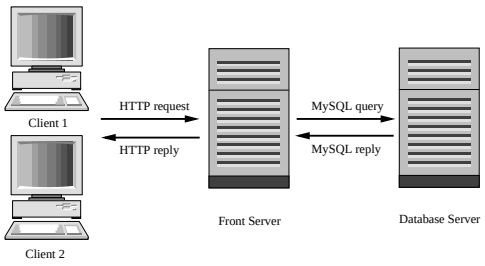

Fig. 8. TPC-W e-commerce System

In the TPC-W benchmark, the requests are directed to a set of HTML pages which include both static images served locally from the front server and dynamic content retrieved from the database server. The HTTP requests are generated by a set of N clients, called Emulated Browsers (EBs), which generate a new request in Z seconds after completing the download of the previously requested page (HTML and images). The distribution of the think times is negative exponential with rate Z^{-1}. Because of the closed-loop structure of the TPC-W workload, where EBs wait HTTP replies before delivering the next request, the number of simultaneous active sessions is upper bounded by N and the system can be modeled as a closed network. We model this closed system as a MAP queueing network composed by two servers representing the front and database servers, respectively, followed by a delay

Fig. 9. MAP queueing network model of TPC-W e-commerce system

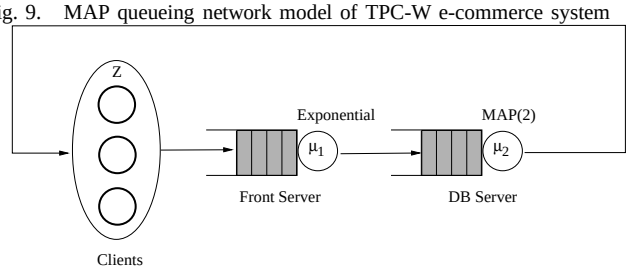

TABLE VI
MEAN SERVICE DEMANDS OF FRONT AND DATABASE SERVERS

	Mean service time [ms]	
N/EBs	front server	DB server
25	5.58	3.26
50	5.00	2.75
75	5.00	3.38
100	5.32	3.38
150	5.32	4.52

server that models the think times Z of EBs. An illustration of this MAP queueing network is given in Figure 9. We point to [10] for a discussion on why this queueing network provides a realistic model of the TPC-W system.

The service processes of the two queues in Figure 9 are parameterized consistently with the processes obtained in [10] from measurements of the browsing mix workload of the TPC-W benchmark. The service process at the front server is modeled as an exponential process with mean service time $\mu_1^{-1} = 5.58$ 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 with a two-phase MAP with mean $\mu_2^{-1} = 3.26$ ms, CV equal to 4, skewness equal to 8.58, and autocorrelation decay rate $\gamma_2 = 0.86$. For increasing values of N, the *measured* mean service demand of both servers changes as reported in Table VI; thus, also in the model we scale the mean service time at the two servers according to these values.

Since measurements on the real system are obtained in terms of server utilizations, in Table VII we focus on the performance analysis of the utilization $U_1(N)$ of the front server[2]. Here, we report the results for the MAP queueing network analysis (exact global balance, upper and lower QR bounds), for product-form queueing networks (PF QN) analyzed by the MVA algorithm, and for GI queueing networks (GI QN) solved exactly by global balance. The last column reports the mean 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 remark that this model can be analyzed *only* by the QR bounds, since the presence of think times between consecutive requests to the system has imposed the inclusion of a delay server in the queueing network.

We first make the obvious observation that modeling methods do not provide results that are identical to the measured

[2]Note that since the utilization is inversely proportional to the end-to-end response times, the considerations in the previous sections for the lower (resp. upper) QR response time bounds apply here to the upper (resp. lower) QR utilization bounds.

TABLE VII
PERFORMANCE ANALYSIS OF TPC-W E-COMMERCE SYSTEM. Bold entries are errors greater than 10% with respect to the measured utilization.

	Front Server Utilization $U_1(N)$					
	MAP QN			PF QN	GI QN	*Real System*
N	Lower QR	Exact	Upper QR	MVA	Exact	*Measured*
25	0.2591	0.2631	0.2730	0.2727	0.2659	*0.2733*
50	0.4457	0.4550	0.4883	0.4875	0.4578	*0.4602*
75	0.6262	0.6405	0.7091	**0.7194**	0.6466	*0.6495*
100	0.7572	0.7800	0.8066	**0.9479**	**0.8410**	*0.7445*
150	0.7269	0.7687	0.7829	**0.9997**	**0.9370**	*0.7379*

values, e.g., even if we solve MAP queueing networks using an exact evaluation of the underlying Markov chain, the predicted utilization is always slightly different from the real one because of the unavoidable parameterization errors deriving from estimating service characteristics from measured traces. The results in Table VII can be interpreted by first noting that the quality of the different modeling techniques changes radically in the jump from $N \leq 75$ to $N \geq 100$. As observed before, medium or large populations are needed to make the effects of burstiness significant for system performance. Table VII indicates that for $N \leq 75$ all methods are essentially accurate with an approximation error no more than 10.8% of the measured utilization for the MVA method when $N = 75$; in these cases, GI models are the most accurate, suggesting that the distribution, and not the burstiness, of the database service times is the main determinant of system performance. For $N \geq 100$, instead, the performance effects of burstiness are strong and the much increased accuracy of MAP queueing networks solutions is immediately visible compared to product-form and GI models, which suffer large 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.

Summarizing, the results of this section on the performance analysis of a real system indicate that MAP queueing networks are a much more robust performance analysis methodology than product-form and GI models. The proposed QR bounds introduce approximation errors on the global balance solution that are much lower than the inaccuracies of product-form and GI estimates. This performance analysis example on a real system provides a strong argument for the adoption of MAP queueing networks for capacity planning.

IX. CONCLUSIONS

Recent workload characterizations have shown that nonrenewal service processes are good abstractions of real systems' workloads, especially of those found in storage systems and Web servers [27], [28], [36]. We have shown that existing queueing network models, which always consider renewal service processes and do not account for nonrenewal features such as autocorrelation in service times, can grossly overestimate or underestimate actual system performance.

We have presented a solution to this problem by studying a new class of MAP closed networks that supports nonrenewal service. We have introduced a class of exact state space

reductions that are computationally tractable and allow the efficient computation of upper and lower linear reduction (LR) and quadratic reduction (QR) bounds on arbitrary MAP network performance indexes, such as utilizations, throughputs, response times, and queue-lengths. QR bounds are more expensive to evaluate than LR bounds, but their fundamental advantage is that they generalize also to the evaluation of models with delay servers and load-dependent service times. To the best of our knowledge, this is the first time that bounds for queueing networks with nonrenewal service are obtained. The LR and QR bounds AMPL specification together with additional resources on MAP queueing networks are available at `http://www.cs.wm.edu/MAPQN/`. Experiments indicate that the LR and QR bounds are extremely accurate, showing on average a 2% relative error on the response time.

Finally, we have shown the applicability of the proposed bounds to the capacity planning of a real TPC-W e-Commerce system. Numerical results indicate that MAP queueing networks evaluated either exactly or with bounds are always very close to the measured server utilization values, whereas traditional MVA and GI models show considerable errors up to 35% and 27% of the measured values, respectively, arguing for the adoption of MAP queueing networks for capacity planning of system with bursty workloads.

ACKNOWLEDGEMENT

This work was supported by the National Science Foundation under grants CNS-0720699 and CCF-0811417. The authors thank Gianfranco Balbo, Jeff Buzen, Larry Dowdy, Giuseppe Serazzi, Murray Woodside, and Qi Zhang who greatly helped in improving the quality of this paper.

REFERENCES

[1] A. T. Andersen and B. F. Nielsen. A Markovian approach for modeling packet traffic with long-range dependence. *IEEE JSAC*, 16(5):719–732, 1998.

[2] M. F. Arlitt and C. L. Williamson. Web server workload characterization: The search for invariants. In *Proc. of ACM SIGMETRICS*, pp. 126–137, 1996.

[3] M. Arlitt and T. Jin. Workload characterization of the 1998 world cup web site. TR HPL-1999-35R1, HP Labs, 1999.

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

[5] D. Bertsimas and J. Tsitsiklis. *Introduction to Linear Optimization*. Athena, 1997.

[6] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi. *Queueing Networks and Markov Chains*. Wiley, 1998.

[7] A. B. Bondi, W. Whitt. The influence of service-time variability in a closed network of queues. *Perf. Eval.*, 6:219–234, 1986.

[8] G. Casale, N. Mi, and E.Smirni. Bound analysis of closed queueing networks with nonrenewal workloads. TR WM-CS-2008-03, College of William and Mary, 2008.

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

[10] G. Casale, N. Mi, L. Cherkasova, and E.Smirni. How to parameterize models with bursty workloads. In *Proc. of the First Workshop on Hot Topics in Measurement and Modeling of Computer Systems, HOTMETRICS 2008*.

[11] G. Casale. An efficient algorithm for the exact analysis of multiclass queueing networks with large population sizes. In *Proc. of joint ACM SIGMETRICS/IFIP Performance*, 169–180, ACM Press, 2006.

[12] G. Casale, E.Z. Zhang, and E. Smirni. Interarrival Times Characterization and Fitting for Markovian Traffic Analysis. TR WM-CS-2008-02, College of William and Mary, 2008.

[13] K. M. Chandy, U. Herzog, and L. Woo. Parametric analysis of queueing networks. *IBM J. Res. Dev.*, 19(1):36–42, 1975.

[14] P.J. Courtois. Decomposability, instabilities, and saturation in multiprogramming systems. *CACM*, 18(7):371–377, 1975.

[15] D.R. Cox and P.A.W. Lewis. *The Statistical Analysis of Series of Events*. John Wiley and Sons, New York, 1966.

[16] MAP Queueing Networks Webpage, online resources at `http://www.cs.wm.edu/MAPQN/`.

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

[18] W. Fischer and K. S. Meier-Hellstern. The Markov- Modulated Poisson Process (MMPP) cookbook. *Perf. Eval.*, 18(2):149–171, 1993.

[19] R. Fourer and D.M. Gay and B.W. Kernighan. *AMPL – A Modeling Language for Mathematical Programming*. Springer-Verlag, 1995.

[20] D. Garcia and J. Garcia. TPC-W E-commerce benchmark evaluation. *IEEE Computer*, pages 42–48, Feb. 2003.

[21] *GNU GLPK 4.8*. `http://www.gnu.org/software/glpk/`.

[22] J.M. Harrison, R.J. Williams, and H. Chen. Brownian models of closed queueing networks with homogeneous customer populations. In *Stochastics and Stochoastics Reports*, 29:37–74, 1990, Taylor & Francis.

[23] A. Horváth and M. Telek. Markovian modeling of real data traffic: Heuristic phase type and MAP fitting of heavy tailed and fractal like samples. In *Performance Evaluation of Complex Systems: Techniques and Tools, IFIP Performance 2002, LNCS Tutorial Series Vol 2459*, pages 405–434, 2002.

[24] S. Kounev and A. Buchmann. Performance modeling and evaluation of large-scale J2EE applications. In *Proc. of the 29th International Conference of the Computer Measurement Group (CMG)*, pages 273–283, 2003.

[25] E. D. Lazowska, J. Zahorjan, G. Graham, and K. C. Sevcik. *Quantitative System Performance*. Prentice-Hall, 1984.

[26] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic. *IEEE/ACM T. Networking*, 2(1):1–15, 1994.

[27] Z. Liu. Long range dependence and heavy tail distributions (special issue). *Perf. Eval.*, 61(2-3):91–93, 2005.

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

[29] J. Morrison and P.R. Kumar. New linear program performance bounds for closed queueing networks. *Discrete Event Dynamic Systems: Theory and Applications*, 11:291–317, 2001.

[30] R. R. Muntz and J. W. Wong. Asymptotic properties of closed queueing network models. In *Proc. Ann. Princeton Conf. on Inf. Sci. and Sys.*, pp. 348–352, 1974.

[31] M. F. Neuts. *Structured Stochastic Matrices of M/G/1 Type and Their Applications*. Marcel Dekker, NY, 1989.

[32] J. Padhye, A. D. Rahatekar and L. W. Dowdy A Simple LAN File Placement Strategy. In *Proc. of CMG 1995*, pages 396–406.

[33] K. R. Pattipati, M. M. Kostreva and J. L. Teele. Approximate Mean Value Analysis Algorithms for Queueing Networks: Existence, Uniqueness, and Convergence Results. *JACM*, 37:643–673, 1980.

[34] M. Reiser. A queueing network analysis of computer communication networks with window flow control. *IEEE T. Comm.*, 27(8):1199–1209, 1979.

[35] M. Reiser and S. S. Lavenberg. Mean-value analysis of closed multichain queueing networks. *JACM*, 27(2):312–322, 1980.

[36] A. Riska and E. Riedel. Long-range dependence at the disk drive level. In *Proc. of 3rd Conf. on Quantitative Eval. of Systems (QEST)*, pp. 41–50, IEEE Press, 2006.

[37] W. Whitt *Stochastic process limits*. Springer, NY, 2002.

[38] J. Zahorjan, E. D. Lazowska, and R. L. Garner. A decomposition approach to modelling high service time variability. *Perf. Eval.*, 3:35–54, 1983.

[39] E.Z. Zhang, G. Casale, and E. Smirni. Interarrival Times Characterization and Fitting for Markovian Traffic Analysis. TR WM-CS-2008-02, College of William and Mary, 2008.

[40] Q. Zhang, N. Mi, A. Riska, and E. Smirni Performance-Guided Load (Un)Balancing Under Autocorrelated Flows. *IEEE T. on Parall. and Distrib. Sys.*, 19(5):652–665, May 2008.