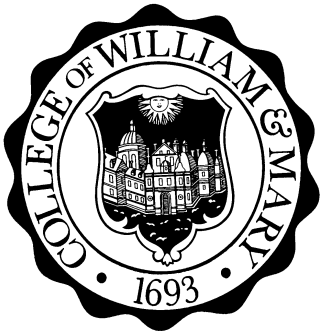


College of
William & Mary
Department of Computer Science

WM-CS-2008-03

Bound Analysis of Closed Queueing Networks with Nonrenewal Workloads

Giuliano Casale, Ningfang Mi, Evgenia Smirni

Bound Analysis of Closed Queueing Networks with Nonrenewal Workloads

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

ABSTRACT

Burstiness and temporal dependence in service processes are often found in multi-tier architectures and storage devices and must be captured accurately in capacity planning models as these features are responsible of significant performance degradations. However, existing models and approximations for networks of first-come first-served (FCFS) queues with general independent (GI) service are unable to predict performance of systems with temporal dependence in workloads.

To overcome this difficulty, we define and study a class of closed queueing networks where service times are represented by Markovian Arrival Processes (MAPs), a class of point processes that can model general distributions, but also temporal dependent features such as burstiness in service times. We call these models MAP queueing networks. We introduce provable upper and lower bounds for arbitrary performance indexes (e.g., throughput, response time, utilization) that we call Linear Reduction (LR) bounds. Numerical experiments indicate that LR bounds achieve a mean accuracy error of 2%. The result promotes LR bounds as a versatile and reliable bounding methodology of the performance of modern computer systems.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Modeling techniques

General Terms

Algorithms, Performance, Theory

Keywords

Queueing networks, closed systems, bound analysis, burstiness, nonrenewal service, temporal dependence, Markovian arrival processes

1. 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 [22–24, 31, 33]. A typical example of temporal dependence is workload burstiness, where the jobs processed by the system are not independent, e.g., the arrival of a short job is much more likely to be followed by the arrival of another short job (and vice versa for long 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 [28]. We address the current lack of more general modeling techniques for systems with nonrenewal workloads by introducing and analyzing a new class of closed queueing networks which can account for temporal dependence in their 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 [5] has been extensively used in the past, since these models enjoy simple solution formulas and low computational cost of exact and approximate algorithms [9, 21]. Queueing networks with general independent (GI) service [8, 15, 29, 32] have been proposed as a solution, but although much more accurate than product-form networks, they remain insufficient for robust performance predictions if the service process is nonrenewal. That is, because they completely ignore temporal dependence between service times, they *cannot* be used to predict performance correctly in systems with nonrenewal workloads.

This paper overcomes the limitations of existing modeling techniques by providing a bound analysis methodology for queueing networks with nonrenewal workloads. We study a class of closed queueing networks where service times are modeled by Markovian Arrival Processes (MAPs). We call these models *MAP queueing networks*. MAPs are a family of point processes which can easily model general distributions and nonrenewal features such as autocorrelation in service times [28]. Efficient fitting schemes for MAP parameterization from measurements are available, e.g., [1, 10, 19], and the resulting MAPs can approximate effectively long-range dependence [1].

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 [26] 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 usually needed to prove the bounding property. In this

paper, we show that it is possible to obtain provable bounds on performance indexes also in non-product-form models.

Our nonrenewal bounds derive from the analysis of the Markov process underlying the MAP queueing network. Because of the state space explosion, its 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*). This state space transformation captures the behavior of the network conditioned on a given queue being busy or idle. The number of states in these marginal spaces grows linearly with the number of jobs in the network; thus, the proposed approach remains computationally tractable also on models with large populations. We derive *exact* balance equations for the equilibrium behavior of the reduced state spaces and illustrate how these formulas can be combined with linear programming [6, 25] for the computation of bounds on mean value indexes. Because the number of reduced states grows linearly with the number of jobs in the network, we call these bounds Linear Reduction (LR) bounds.

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 provide evidence that existing GI approximations and decomposition methods show unacceptably large errors on queueing network models with temporal dependence in the service process (Section 2).
- We define MAP queueing networks as a generalization of existing queueing networks that can model nonrenewal workloads (Section 3).
- We develop the LR bounds on performance indexes for nonrenewal MAP queueing networks that are based on a new marginal state space reduction (Sections 4 and 5).
- We present an extensive set of representative case studies showing that the LR bounds capture very well mean performance indexes such as response times or utilizations (Section 6).

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 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. We conclude the paper by outlining model generalizations and extensions in Section 7. The AMPL specification [17] of the LR bounds is available at `http://www.cs.wm.edu/MAPQN/`.

2. PREVIOUS WORK

In Section 2.1, we review previous work on non-product-form queueing network models with FCFS queues and GI service [7]. These models are the renewal specialization of the MAP queueing networks introduced in Section 3. In Section 2.2, we evaluate the applicability of approximation algorithms for models with GI service to models with nonrenewal service. Due to limited space, we point the reader to [7, 12, 21] for general background on queueing network modeling and Markov processes. Throughout this paper we assume that service time distributions are modeled by the method of phases [7, 13].

2.1 Analysis of Models with Renewal Service

Closed networks of FCFS queues enjoy a product-form solution if all service times are exponentially distributed [5]. If one or more servers have renewal (also called general independent (GI)) service, such as hyperexponential or Coxian [13], the product-form theory does not apply and approximate methods are used for evaluating performance [7].

An approximation based on Markov renewal theory is developed by Reiser in [29]. For each queue, the MVA arrival theorem [30] is generalized to include the coefficient of variation (CV) of the GI service process. Experiments in [8, 15] show that this approach, although simple, is prone to large approximation errors.

In [32], Zahorjan et al. obtain an approximate mean value analysis (AMVA) by decomposition-aggregation [12]. 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 [30] and the results are weighted to approximate the GI network. Validation results of the AMVA decomposition-aggregation show good accuracy.

In [15], Eager et al. improve the results in [29] and [32]. 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. [15] also improves the decomposition method in [32] and makes it compatible with the iterative AMVA framework to achieve lower computational costs on networks with several queues.

Marie's method, the diffusion approximation (DA) 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 [7]. DA and MEM rely 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 properties have not been assessed [8]; DA and MEM are typically less accurate.

Finally, the Chandy-Herzog-Woo (CHW) method [11, 21] 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 [8].

2.2 Applicability to Nonrenewal Service

To the best of our knowledge, no results are available for analyzing closed networks with nonrenewal service, see [28] for related work in single queue systems. In this section, we establish the applicability of the methods described in Section 2.1 to closed networks with nonrenewal service.

We consider identically distributed service processes which are characterized by temporal dependence. The temporal dependence of a nonrenewal processes can be approximately modeled by the autocorrelation function ρ_k which captures the similarity in magnitude of samples spaced by k lags [13]. As an example, a service process can have hyperexponentially distributed samples without being necessarily renewal; that is, the usual terminology "hyperexponential service" implicitly refers to the renewal version of the process, but in general nonrenewal processes with hyperexponential distribution can be defined. These are immediately obtained by changing the order of the samples without altering their distribution, which results in temporal dependence.

A simple case of nonrenewal closed network is shown in Figure 1. We use this simple model to evaluate the applicability of methods for GI queueing networks to models with nonrenewal service. Queue 1 is exponential with rate $\mu_1 = 1$; queue 2 has MAP(2)

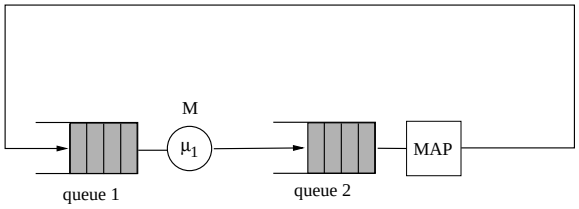


Figure 1: **Example model with nonrenewal service.** The service is an exponential process with rate μ_1 at queue 1, at queue 2 is a possibly nonrenewal two-phase Markovian Arrival Process (MAP(2)) [28].

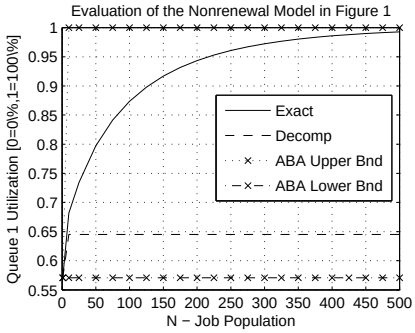


Figure 2: **Exact global balance solution of the nonrenewal model in Figure 1 compared with the ABA bounds [21, 26] and the decomposition-aggregation approximation [12].** Although accurate in renewal models, in nonrenewal models the decomposition is often unable to capture the trend of performance indexes such as the utilization shown here. The saturation of the approximation is due to the saturation of the underlying product-form models used in the decomposition. Also the ABA bounds, which apply to general networks, are inaccurate.

service [16] which can exhibit autocorrelation in service time samples and thus be a nonrenewal process. In a MAP(2), the lag-k autocorrelation ρ_k geometrically decays to zero with rate[1] γ_2 according to the relation $\rho_k = \gamma_2^k \left(1 - 1/\mathrm{CV}^2\right)/2$, [10]. In this example we choose $\gamma_2 = 0.75$, the MAP(2) is 33% faster than the exponential queue (i.e., mean rate $\mu_2 = 1.33$), has $\mathrm{CV} = 5$, and skewness 15; the process is obtained by the moment and autocorrelation matching algorithm in [10]. The results discussed below are qualitatively similar for other models.

Using the model in Figure 1, we have observed that several of the methods considered in Section 2.1 cannot be applied to models with nonrenewal workloads; we distinguish two groups:

Non-applicable methods. This group includes Reiser's approximation, the AMVA methods in [15, 32], Marie's method, DA, and MEM. Intuitively, these methods cannot apply for the following reasons. For the considered example, all these methods depend *only* on the mean, the CV, and the probability of starting service in one of the two phases. For instance, Marie's method applies corrections based on the two-phase Coxian renewal process model that is completely specified by these three parameters [7]. The information about the order of sampling, that is fundamental to nonrenewal service, is given only by the underlying Markov process which is not directly evaluated by these methods. Since these techniques ignore the order of sampling of the service times, they cannot account for the temporal dependence and therefore produce identical results if $\rho_k \equiv 0$, $k \geq 1$, or in the nonrenewal case $\rho_k \neq 0$, for some $k \geq 1$. Yet nonrenewal models can have performance that is extremely different from their renewal counterpart [23, 24]; therefore these methods are unfit for the analysis of nonrenewal models.

Applicable methods. Decomposition-aggregation [12] can instead be used for the analysis of models with nonrenewal service. Since this method requires to evaluate all or part of the underlying Markov process, it is not limited to statistical moments of the service time distribution, but can account for changes in the phase transition rates of the MAP which result in autocorrelated samples. Decomposition-aggregation can be easily applied to the underlying process by aggregating states with identical active MAP phases; methods similar to those in [15, 32] can be defined based on this partitioning.

However, we have found that decomposition-aggregation can frequently exhibit severe errors if used in networks with nonrenewal service. Figure 2 show the predicted utilization of decomposition (dashed line) versus the actual utilization (solid line) for the bottleneck queue 1 in Figure 1. The actual utilization is obtained by solving the underlying Markov process by global balance, therefore it is exact. The ABA bounds, which apply to general models [21, 26], are also depicted. We recall that other classes of bounds such as the popular balanced job bounds (BJB) apply only to product-form networks [21]. Although decomposition is very accurate in the renewal case, its application to the nonrenewal case results in increasingly large approximation errors for larger populations. Cases similar to the one plotted in Figure 2 are easy to find for different values of the model parameters. The low quality of the results in the nonrenewal case is due to the quick saturation of the product-form models used in these approximations, which reach maximum utilization for lower loads than the nonrenewal model.

The observations of this section indicate that the analysis of nonrenewal workloads cannot be performed accurately with existing techniques for models with GI service times. Due to the large approximation uncertainty and the lack of an exact product-form solution, bounding techniques are desirable. In order to address this limitation, in the following sections we introduce a bound analysis methodology for nonrenewal networks.

3. 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 1.

3.1 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 [4]. If $K_i = 1$, then the MAP reduces to an exponential distribution, otherwise it generates service time samples that

[1]The value γ_2 is in a MAP(2) the second largest eigenvalue of the Markov chain embedded at arrival instants [10].

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
$\bar{\pi}_j^k(n_i, h)$	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	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 $i = 1$)

Table 1: Summary of main notation

are phase-type (PH) distributed [28]. 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) [16]. It should be nevertheless remarked that MAP fitting can be still a challenging problem if the data has an irregular temporal dependence structure, see [19] for a review. We point to [10] 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 [19] and references therein for background on MAPs and MAP fitting.

3.2 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 computed as

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

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 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.

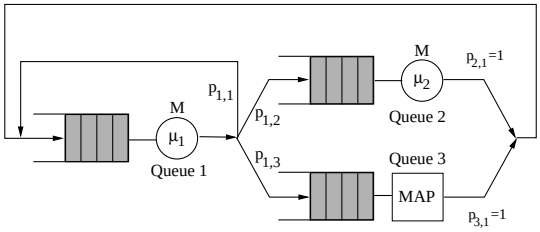


Figure 3: **Example network composed by two exponential queues and a MAP queue.** In the case where the MAP is a renewal two-phase hyperexponential process, this reduces to Balbo's model considered in the validation of approximations for GI service [8].

As a summarizing example, the MAP network in Figure 3 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 4. The state space description is given in the caption. 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 [8]; throughout the paper we illustrate some of the proposed techniques using this model.

4. STATE SPACE REDUCTION

General approximation techniques for non-product-form models, such as decomposition, are reviewed in Section 2. 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 [7].

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

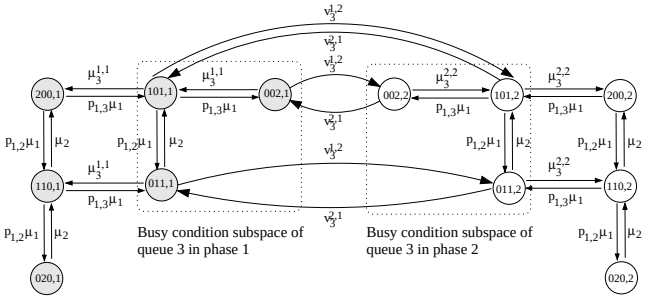


Figure 4: **Underlying Markov process of the network in Figure 3 in the simple case when the MAP is a MMPP(2) process; the job population is $N = 2$.** Two queues are exponential with rates $\mu_1 \equiv \mu_1^{1,1}$ and $\mu_2 \equiv \mu_2^{1,1}$, respectively; the third queue is a MAP with $K_3 = 2$ phases having $\mu_3^{k,h} = 0$ for $k \neq h$, that is a MMPP(2) process. $(002, 1)$ indicates that the exponential queues are idle and the MAP queue has two jobs and is in phase 1; in $(110, 2)$, the phase 2 is the phase left active by the last served job.

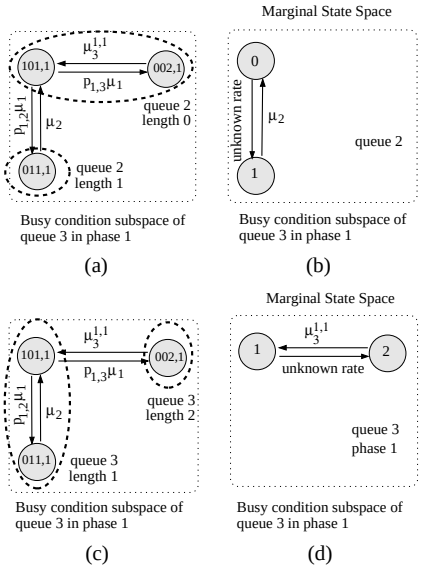


Figure 5: **Example of marginal state spaces for the model in Figure 4.** Figures (a) and (c) show the observation of either queue 2 or queue 3 in phase 1 during the busy subspace B_3^1 of queue 3 in phase 1. The dashed ovals indicate states in the original state space in Figure 4 that are implicitly accounted for in the marginal state spaces for queue 2 and for queue 3 in phase 1 depicted in figure (b) and (d), respectively. Note that the reduction is not a lumping or a decomposition-aggregation since we are completely ignoring the transition from/to the busy condition subspace that are present in Figure 4; also some rates would require the equilibrium probability of the state $(101, 1)$ in Figure 4 which is unknown, and therefore the marginal state space cannot be solved in isolation.

4.1 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 4.1.1. Two example marginal spaces for the model in Figure 4 obtained for the busy condition subspace B_3^1 are shown in Figure 5. Figures 5(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[2] without the equilibrium probability $\pi(101, 1)$. Figure 5(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 5 clearly shows that our approach differs from 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 5 how the busy condition reduction can be used to define the LR performance bounds.

4.1.1 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

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

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

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

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 [21] 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

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

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 [21]. In particular, the response time is $R = N/X$.

4.1.2 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*

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

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.,

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

$1 \leq k \leq K_j$.

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

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

□

4.1.3 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. *(Sketch of the proof, see appendix 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.

4.1.4 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.

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 $n_i + 1$, that is,*

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

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. *(Sketch of the proof, see appendix 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.

□

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, $1 \leq n_i \leq N-2$,

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

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^*. □

4.2 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 $B_j^k \cup I_j^k$. □

THEOREM 5. *The performance indexes in busy and idle subspaces are related by the following equation*

$$\begin{aligned}
&\textstyle\sum_{\substack{h=1,\\ h\neq k}}^{K_i}\sum_{j=1}^{M} q_{i,j}^{k,h}Q_i^k+\sum_{\substack{j=1,\\ j\neq i}}^{M}\sum_{h=1}^{K_i} q_{i,j}^{h,k}U_i^h\\
&=\textstyle\sum_{\substack{j=1,\\ j\neq i}}^{M}\sum_{h=1}^{K_j}\sum_{u=1}^{K_j} q_{j,i}^{h,u}J_i^k(j,h)+\sum_{\substack{h=1,\\ h\neq k}}^{K_i}\sum_{j=1}^{M} q_{i,j}^{h,k}Q_i^h,
\end{aligned} \quad (18)$$

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

PROOF. *(Sketch of the proof, see appendix 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. □

5. 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

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

$f_{min} = \min f(\boldsymbol{\pi})$
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 $\bar{\pi}_j^k(n_i,h) \in \boldsymbol{\pi}$.

Figure 6: **Linear program determining a lower bound on an arbitrary linear performance index $f_{exact} = f(\boldsymbol{\pi}_{exact})$.** For instance, f_{exact} can be either a mean index such as a throughput or a more detailed descriptor such as a marginal probability $\pi_j^k(n_i,h)$.

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 [6] 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

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

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 accuracy of the LR bounds is validated in the Section 6. 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 10 MAP(2) queues and $N = 50$ jobs using an interior point solver in approximately four minutes; for $N = 100$ the solution of the same model is found in approximately ten minutes suggesting good scalability. In general, the complexity of computing bounds with the linear program in Figure 6 grows as $O(\text{lp}(M^2K_{max} + MN, K_{max}^2M^2N))$, where $\text{lp}(r,c)$ 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 $2K_{max}^2M^2N$.

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

Table 2: **State Space Reduction Effectiveness.** Comparison of the total number of states in the marginal state spaces with the original state space of the queueing network. All queues have MAP service times with K_{max} phases. The number of states in the marginal state spaces grows linearly in the population size, whereas the growth for the original state space is combinatorial.

To appreciate the reduction of the state space, Table 2 compares the number of states in the marginal state spaces with the original state space size in models with larger population and number of queues. The reduced spaces have cardinality that can be several orders of magnitude smaller than the original state space.

5.1 Discussion

The balances obtained in the Section 4 provide a rich 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 underdetermined and describes a *family* of possible equilibria for the underlying Markov process, among which we cannot distinguish the real one. The linear programming approach allows to select 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 [6] may be used as a tool for investigating the relative importance of a certain balance for the model under study in order to minimize the size of the linear program.

6. ACCURACY VALIDATION

We assess the accuracy of the LR bounds using the following methodology. We use both randomly-generated models and representative case studies, see Table 3 for a description of the employed input parameters. In order to assess the accuracy of the LR bounds, we evaluate their maximal relative error with respect to the exact solution of the MAP network computed by global balance. Due to the state space explosion, the experimentation using exact global balance solutions is often prohibitive for MAP networks with more than three 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 [10].

Network Param.	*Case Studies*	*Random Models*
M	$2-3$	3
$p_{i,j}$	variable	random $[0,1]$
N	$10, 25, 100$	all in $[10, 1000]$
# of nonren. MAPs	$1, 2, 3$	1
MAP(2) Param.	*Case Studies*	*Random Models*
mean	variable	random $[0,1]$
CV	$0.5, 2, 4, 8$	random $[0.5, 10]$
skewness	30	random $[2, 250]$
γ_2	$[.00, .99]$	random $[.00, .99]$

Table 3: Input parameters used in the validation study.

For each model, we use the linear program in Figure 6 to compute upper and lower limits 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 from the exact response time R. We do not report errors on other measures due to lack of space, but we remark that they are typically in the same range as those of response time. We used the GNU Linear Programming Kit [18] to solve the linear program on a Intel Xeon 3.73Ghz using an AMPL specification [17] of the linear program in Figure 6. The AMPL specification is available for download at [14].

6.1 Random Models

In order to evaluate the general quality of the LR bounds, we evaluate $10,000$ random models. The models are generated according to the specifications in Table 3. Each random model is solved for all feasible populations 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 error function is a conservative estimator since it returns the *maximum* error of R_{bnd} over all evaluated populations. The converge of the bounds to the exact asymptotic value is not accounted by this metric and only the worst case error is measured.

We see that the variability in the routing matrix makes it possible to evaluate different levels of balancing in the mean service demand at the queues. Table 4 indicates that the proposed bounds perform extremely well also for this class of models. The mean error is $1-2\%$ for both bounds with a standard deviation of 0.02; the median is less than the mean, indicating that the asymmetry of the error distribution is more concentrated on small errors. The maximum error is found to be 14.2% for the 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 the 1% of the total number of experiments. The burstiness in these cases increases the response times at the autocorrelated station in a way that cannot be easily captured. Furthermore, the lower bound seems to be more sensitive to increased variability and autocorrelation than the upper bound, where the worst case error is for a MAP with moderate burstiness. The difference in sensitivity to MAP parameters is a positive property of the LR bounds, because large inaccuracies in one bound can be compensated by the relative accuracy of the other. Detailed numerical sensitivity of the two bounds with respect to the model parameters supporting these intuitions is discussed in the next subsection.

		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 4: **Results of Random Experiments** Absolute maximal relative error ($0 \equiv 0\%$, $1 \equiv 100\%$) over $10,000$ random queueing networks for the response time $R = N/X$ (R_{min}=lower LR bound, R_{max}=upper LR bound). E.g., the mean of the error Δ is 1.3% for R_{max} and 2.2% for R_{min}.

6.2 Representative Case Studies

We consider six representative case studies illustrating the accuracy properties of the LR bounds; the results presented in this sections are typical of the actual bound accuracy as we have shown in the random model validation section.

Depending on the experiment, the MAP can be either an Erlang-2 (E_2), a renewal two-phase hyperexponential (H_2), a Poisson process (M), or a nonrenewal MAP(2) (MAP). As we show in Case 3, the bounding is more difficult for increasing values of CV; therefore we consider the E_2 process with $\mathrm{CV} < 1$ only in Case 1. To focus on the effects on accuracy of the most important moments (i.e., mean and CV), we also fix in the case studies the skewness to 30. In the random experiments, we have spanned all feasible skewness values for the considered MAPs (range $[2, 250]$).

6.2.1 Case 1: Sensitivity to Renewal and Nonrenewal Service Processes

The network is composed by two queues in series. The service process can be E_2, M, H_2 or MAP. For all processes, the mean rate is $\mu_1 = 1$ at the bottleneck queue 1, $\mu_2 = 2$ at the non-bottleneck queue 2. The MAP has $\mathrm{CV} = 5$ and autocorrelation decay rate ($\gamma_2 = 0.5$); the H_2 has the same moments of the MAP, but being renewal $\rho_k \equiv 0$, for all lags $k \geq 1$.

Results. Table 5 reports the maximal relative error on response times for all possible combinations of service processes. It is found that: (1) nonrenewal models are significantly more difficult to evaluate than renewal models, e.g., on the most difficult renewal case, the H_2/H_2 model, the LR bounds progressively converge to the exact as N increases, while on the MAP/MAP case the error remains at 10% also for $N = 100$; (2) hyper-exponential CVs are typically more difficult to approximate than hypo-exponential CVs; (3) the error of R_{min} is no greater than 4% and is quite insensitive to the service process type; (4) R_{max} is more sensitive to nonrenewal service where it achieves a worst-case error of 11%.

6.2.2 Case 2: Sensitivity to Network Routing

We evaluate the impact on accuracy of network routing, showing the counter-intuitive fact that nonrenewal balanced networks are difficult to approximate. We consider the three queue model in Figure 3 with the mean service rates considered in [8], i.e., $\mu_1 = 1/0.028$, $\mu_2 = 1/0.04$, and $\mu_3 = 1/0.28$. We evaluate the accuracy of the LR response time bounds in the case where the network is perfectly *balanced* ($p_{1,1} = 0.2$, $p_{1,2} = 0.7$, $p_{1,1} = 0.1$), *partially unbalanced* (i.e., queue 3 is bottleneck, queue 1 and queue 2 are balanced, $p_{1,1} = 0.1$, $p_{1,2} = 0.7$, $p_{1,1} = 0.2$, this case corresponds to Balbo's model in [8]), or *unbalanced* (queue 3 is bottleneck, queue 2 is slower than queue 1, $p_{1,1} = 0.33$, $p_{1,2} = 0.33$, $p_{1,3} = 0.34$). The MAP queue 3 has $\mathrm{CV} = 4$ and $\gamma_2 = 0.5$.

Service		$N=10$		$N=25$		$N=100$	
renewal service processes only							
bnk	*nonbnk*	Δ_{min}	Δ_{max}	Δ_{min}	Δ_{max}	Δ_{min}	Δ_{max}
E_2	E_2	0.01	0.00	0.01	0.00	0.00	0.00
E_2	H_2	0.01	0.08	0.02	0.05	0.02	0.00
H_2	E_2	0.01	0.00	0.00	0.00	0.00	0.00
H_2	H_2	0.01	0.09	0.01	0.06	0.01	0.01
at least one nonrenewal service process							
bnk	*nonbnk*	Δ_{min}	Δ_{max}	Δ_{min}	Δ_{max}	Δ_{min}	Δ_{max}
MAP	E_2	0.01	0.00	0.01	0.00	0.00	0.00
MAP	H_2	0.01	0.09	0.01	0.06	0.01	0.01
E_2	MAP	0.00	0.11	0.00	0.11	0.00	0.10
H_2	MAP	0.00	0.05	0.00	0.05	0.00	0.06
MAP	MAP	0.00	0.11	0.00	0.11	0.00	0.10

Table 5: **Case 1 - Sensitivity to Renewal/Nonrenewal Service Processes.** Absolute value of the maximal relative error ($0 \equiv 0\%$, $1 \equiv 100\%$) on the response time $R = N/X$ of two queues networks (Δ_{min}=lower bound error, Δ_{max}=upper bound error).

Results. Figure 7 shows the LR bounds on response times and queue 3 utilization in the balanced and partially unbalanced cases; the utilization bounds follows immediately from the response time bounds by Little's Law and the Utilization Law [21] and are useful to evaluate the LR bound accuracy as a function of the bottleneck queue congestion level. We have found that the unbalanced case is extremely similar to the partially unbalanced and therefore is not plotted. It is found that: (1) the LR bounds of both utilization and response times are very close to the exact value on most populations; (2) both bounds progressively converge to the asymptotic exact, a feature that is not always found in standard bounds for queueing networks (e.g., the ABA lower utilization bound never converges asymptotically if $M \geq 2$); (3) the slower asymptotic convergence in the balanced case makes the approximation more challenging, but the maximal error remains less than 11% of response time and 12% of bottleneck utilization.

6.2.3 Case 3: Sensitivity to Service Variability and Autocorrelations

We consider the nonrenewal MAP/MAP model in the last row of Table 5 and we vary the CV and the autocorrelation decay rate γ_2 for the two identical MAPs.

Results. Table 6 reports the maximal relative error on response times. It is found that: (1) the error of R_{min} is loosely sensitive to CV and γ_2; (2) the error of R_{max} is proportional to both CV and γ_2, but decreases with the population size; (3) the maximum error of R_{max} (13%) is compensated by the minimum error of R_{min} (0%). Note that the errors are higher than in the random models since we are now considering two MAPs instead of one.

6.2.4 Case 4: Applicability to Real Workloads

We illustrate the applicability to real workloads evaluating the model in Figure 3 using Balbo's partially unbalanced configuration ($p_{1,1} = 0.1$, $p_{1,2} = 0.7$, $p_{1,1} = 0.2$) and considering at queue 3 the nonrenewal MMPP(16) fitted in [1] from the classic long-range-dependent (LRD) Bellcore-pAug89 trace of [22]. This trace is often considered in the literature as representative of many long-range dependent processes found in modern computer, communication, and multimedia systems. We scale the mean of the MMPP(16) so that $\mu_3 = 1/0.28$; CV, skewness and autocorrelations are unchanged (see [1] for details on this MMPP(16) and its autocorrelations ρ_k).

Results. Table 7 illustrates results for different populations. We consider $N = 50$ instead of $N = 100$ because global balance is prohibitively expensive in the second case. The results indicate that: (1) the maximal error is of 5% and the accuracy improves with the population size. This result is consistent with the accuracy levels in the other case studies and indicates the applicability of the LR bounds also with models of real workloads; (2) because of the large order of the MMPP, the experiment also illustrates the low sensitivity of accuracy to changes in the number of phases.

		$N=10$		$N=25$		$N=100$	
CV	γ_2	Δ_{min}	Δ_{max}	Δ_{min}	Δ_{max}	Δ_{min}	Δ_{max}
2	0.000	0.00	0.01	0.00	0.01	0.00	0.00
2	0.250	0.00	0.02	0.00	0.01	0.00	0.00
2	0.500	0.00	0.02	0.00	0.01	0.00	0.00
2	0.750	0.00	0.02	0.00	0.02	0.00	0.01
2	0.990	0.00	0.02	0.00	0.02	0.00	0.02
4	0.000	0.00	0.05	0.01	0.03	0.01	0.00
4	0.250	0.00	0.05	0.01	0.04	0.01	0.01
4	0.500	0.00	0.06	0.00	0.04	0.01	0.02
4	0.750	0.00	0.06	0.00	0.05	0.01	0.03
4	0.990	0.00	0.06	0.00	0.06	0.00	0.06
8	0.000	0.00	0.12	0.01	0.10	0.02	0.04
8	0.250	0.00	0.12	0.01	0.10	0.02	0.05
8	0.500	0.00	0.12	0.01	0.11	0.02	0.07
8	0.750	0.00	0.12	0.00	0.12	0.01	0.09
8	0.990	0.00	0.13	0.00	0.13	0.00	0.13

Table 6: **Case 3 - Sensitivity to Burstiness and Autocorrelations.** Absolute value of the maximal relative error ($0 \equiv 0\%$, $1 \equiv 100\%$) on the response time $R = N/X$ for a MAP/MAP network (Δ_{min}=lower bound error, Δ_{max}=upper bound error).

N	Δ_{min}	Δ_{max}
1	0.00	0.00
2	0.05	0.03
10	0.03	0.02
25	0.01	0.01
50	0.00	0.01

Table 7: **Case 4 - Applicability to Real LRD Workloads.** We use the Bellcore-Aug89 trace [22] fitted in [1] by a MMPP(16). Absolute value of the maximal relative error ($0 \equiv 0\%$, $1 \equiv 100\%$) on the response time $R = N/X$ for the network in Figure 3 with the MMPP(16) at queue 3 (Δ_{min}=lower bound error, Δ_{max}=upper bound error).

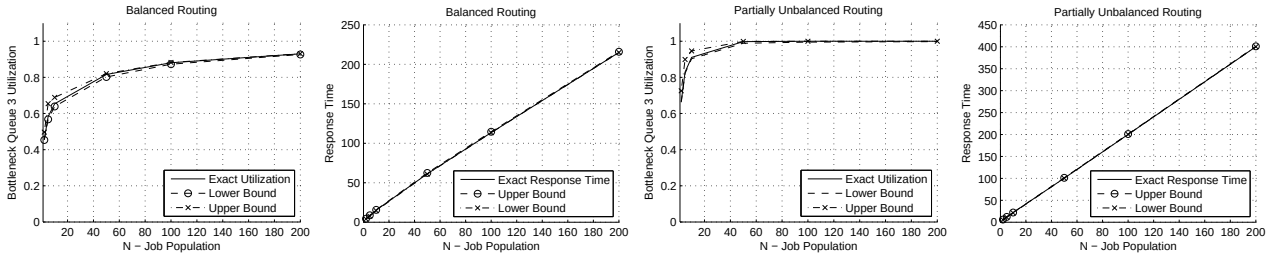
6.2.5 Case 5: Sensitivity to Multiple MAP Queues

We consider a closed network with three queues in series. The mean service rate at queue i is $\mu_i = i$. Service is either exponential or MAP(2) with $\mathrm{CV} = 4$ and $\gamma_2 = 0.5$. During the i-th experiment, $0 \leq i \leq 3$, the first i queues are exponential, while the remaining $M - i$ are MAP(2).

Results. Table 8 reports results of the four experiments. It is found that: (1) the worst case error of 11% is achieved when all three queues are MAP(2); (2) both LR bounds are sensitive to the increase in the number of MAP queues; (3) with zero or one MAP(2) the bounds yet converge asymptotically to the exact value; (4) the same conclusion is not immediate for the other cases, but computing R_{min} and R_{max} for $N = 250, 500$ reveals (not shown in the table) that at $N = 250$ the gap between the bounds is $R_{max}/R_{min} - 1 \approx 11\%$, while for $N = 500$ it drops to $\approx 6\%$ suggesting convergence.

6.2.6 Case 6: Sensitivity to Network Size

In general, we have observed that for large models with dense routing matrices the linear constraining of the inevitably very small marginal probabilities may lead to numerically difficult problems.



(a) Utilization (b) Response Time (c) Utilization (d) Response Time

Figure 7: **Case 2 - Sensitivity to Network Routing.** Analysis of utilization and response time in the balanced and partially unbalanced cases of the network in Figure 3. The unbalanced case (not depicted) is qualitatively similar to the partially unbalanced, but shows a quicker convergence to the saturation.

MAPs	N	Δ_{min}	Δ_{max}	MAPs	N	Δ_{min}	Δ_{max}
0/3	2	0.00	0.00	2/3	2	0.00	0.10
0/3	10	0.03	0.00	2/3	10	0.04	0.11
0/3	25	0.01	0.00	2/3	25	0.05	0.09
0/3	100	0.00	0.00	2/3	100	0.09	0.03
1/3	2	0.00	0.05	3/3	2	0.00	0.13
1/3	10	0.03	0.02	3/3	10	0.04	0.15
1/3	25	0.02	0.00	3/3	25	0.07	0.12
1/3	100	0.00	0.00	3/3	100	0.11	0.04

Table 8: **Case 5 - Sensitivity to Multiple MAP Queues.** Absolute value of the maximal relative error ($0 \equiv 0\%$, $1 \equiv 100\%$) on the response time $R = N/X$ for the MAP/MAP network in Table 5 (Δ_{min}=lower bound error, Δ_{max}=upper bound error). The notation $k/3$, $k = 0, 1, 2, 3$, used in the MAPs column indicates the number of nonexponential MAPs used in the model; the remaining $M - k$ queues have exponential service.

In these cases, interior point algorithms perform better than the simplex algorithm [6], but the numerical corrections of the linear solver can inflate computational times. However, if the model is large and the routing is sparse, the numerical conditioning is usually much improved, and models up to 15-20 queues and hundreds of jobs can be evaluated efficiently.

In this example, we study a distributed system composed by a farm of J Web servers. Each Web server serves directly static object requests (e.g., pictures) and communicates with a local database for the generation of HTML pages. Each local subsystem of Web server and database server is modeled as shown in Figure 8 similarly to the model of a real J2EE Web application defined and validated in [20]. We use the values found in [20] for the parameterization of the service times at the Web servers ($\mu_{WS}^{-1} = 12.98ms$), at the DB servers ($\mu_{DB}^{-1} = 10.64ms$), and for the local communication ($\mu_{WS-DB-Comm}^{-1} = 1.12ms$). The ratio between static object requests and Web pages requests is set to $9.15 : 1$ as measured in [3]. We fix the routing probability from the communication link to each subsystem to $1/J$. We use the median file size measured in [3] which is approximately three packets. The mean, variance, skewness and lag-1 autocorrelation of inter-arrival times between packets is set as in the real workload used in Section 6.2.4. With these parameters, the fitted MAP at the communication links is a MAP(2) with rates $v_{Net}^{1,2} = 0.8074$, $v_{Net}^{2,1} = 0.2107$, $\mu_{Net}^{1,1} = 14.5809$, $\mu_{Net}^{1,2} = 0.0$, $\mu_{Net}^{2,1} = 0.0$, and $\mu_{Net}^{2,2} = 131.7814$.

Results. Since, due to the state space explosion, exact global balance solutions are prohibitive for $M \geq 4$, we report in Table 9 the relative gap $R_{max}/R_{min} - 1$ between LR upper and lower response time bounds for different value of J. According to the

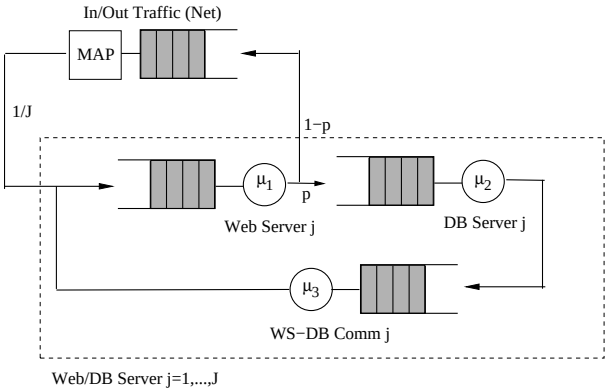


Figure 8: **Case 6: Distributed Web System.** Web farm composed by J Web servers. Bursty inbound/outbound traffic is modeled by a MAP server with temporal dependent service. Incoming requests are dispatched to a given Web server with fixed probability $1/J$.

model definition, the number of queues M is equal to $3J+1$, where the additional queue is representative of inbound/outbound traffic. As J increases, there is little impact on the accuracy of the method which remains acceptable on all populations. In particular, as the population increases the error decreases below 10% on models with large populations.

To summarize, we have provided numerical evidence that the LR bounds provide accurate estimates using an extensive set of experiments. Sensitivity analysis to the different parameters has been conducted to identify the major sources of inaccuracy for the LR bounds.

7. CONCLUSIONS AND FUTURE WORK

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 [23, 24, 31]. 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, 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

M	N	bnd gap	M	N	bnd gap
10	10	0.187	13	10	0.189
10	25	0.141	13	25	0.143
10	100	0.098	13	100	0.097
10	250	0.089	13	250	0.088
16	10	0.181	19	10	0.180
16	25	0.137	19	25	0.138
16	100	0.089	19	100	0.089
16	250	0.085	19	250	0.080

Table 9: **Case 6 - Sensitivity to Network Size.** Relative gap $R_{max}/R_{min} - 1$ of upper and lower bounds ($0 \equiv 0\%$, $1 \equiv 100\%$) on the response time $R = N/X$ (R_{min}=lower bound, R_{max}=upper bound).

computationally tractable and allow the efficient computation of upper and lower linear reduction (LR) bounds on arbitrary MAP network performance indexes, such as utilizations, throughputs, response times, and queue-lengths. To the best of our knowledge, this is the first time that bounds for queueing networks with MAP service are obtained. The LR 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 bounds are extremely accurate, showing on average a 2% relative error on the response time with a maximum error of 14% (for 99% of the random models the maximum error is also found to be less than 10%). We also remark that if the objective function f used in the program in Figure 6 is a nonlinear function of $\boldsymbol{\pi}$, our result immediately generalizes to the nonlinear case provided that the results are global optima.

An important extension of this work is the optional inclusion of delay servers into the MAP network. Since the service rates of a delay are load-dependent, this requires a generalization of our work to the load-dependent case. It may be also interesting to extend the presented class of MAP networks to include additional scheduling disciplines or multiclass workloads in order to fully subdue the class of product-form queueing networks. Finally, it may be interesting to compare our bounds with standard diffusion or asymptotic approximations.

Acknowledgement

This work was supported by the National Science Foundation under grants ITR-0428330 and CNS-0720699. The authors thank Gianfranco Balbo, Jeff Buzen, Larry Dowdy, Giuseppe Serazzi, Murray Woodside, Qi Zhang, and the SIGMETRICS reviewers for detailed comments which greatly helped in improving the quality of this paper.

8. 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] S. Asmussen and F. Koole. Marked point processes as limits of Markovian arrival streams. *J. App. Prob.*, 30:365–372, 1993.

[5] 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.

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

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

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

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

[10] 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.

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

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

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

[14] *MAP Queueing Networks Webpage.* `http://www.cs.wm.edu/MAPQN/`.

[15] 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.

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

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

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

[19] 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.

[20] 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.

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

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] R. D. Nelson. *Probability, Stochastic Processes and Queueing Theory*. Springer-Verlag, 1995.

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

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

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

[31] 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.

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

[33] 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.

APPENDIX
Global Balance Equations

Let δ_m be a binary variable that is one if and only if queue m is busy in state $(\vec{n},\vec{k})$, i.e., $n_m \geq 1$. For the state $(\vec{n},\vec{k})$ of a closed queueing network with MAP servers, the associated global balance equation is

$$\sum_{m=1}^{M}\sum_{j=1}^{M} O_{m,j}(\vec{n},\vec{k}) = \sum_{m=1}^{M}\sum_{j=1}^{M} I_{m,j}(\vec{n},\vec{k}), \quad (19)$$

where the O flows are outgoing probability fluxes and the I flows are incoming fluxes. Denoting the m-th element of $\vec{k}$ by k, we have for $j=m$ that the contribution of the background transitions of queue m MAP is

$$O_{m,m}(\vec{n},\vec{k}) = \delta_m \sum_{\substack{h=1\\h\neq k}}^{K_m} q_{m,m}^{k,h}\pi(\vec{n},\vec{k}),$$
$$I_{m,m}(\vec{n},\vec{k}) = \delta_m \sum_{\substack{h=1\\h\neq k}}^{K_m} q_{m,m}^{h,k}\pi(\vec{n},\vec{k}+(h-k)\vec{e}_m),$$

and similarly for $j\neq m$ the contribution of the completion transitions of queue m MAP is

$$O_{m,j}(\vec{n},\vec{k}) = \delta_m \sum_{h=1}^{K_m} q_{m,j}^{k,h}\pi(\vec{n},\vec{k}),$$
$$I_{m,j}(\vec{n},\vec{k}) = \delta_j \sum_{h=1}^{K_m} q_{m,j}^{h,k}\pi(\vec{n}-\vec{e}_j+\vec{e}_m,\vec{k}+(h-k)\vec{e}_m),$$

where in the last definition δ_j assures that $\vec{n}-\vec{e}_j+\vec{e}_m$ is a feasible population vector.

Detailed proof of Theorem 2

Let k be the phase of i in $\vec{k}$; let also $B_i^k \cup I_i^k$ be the states where i is in phase k either busy (set of states B_i^k) or idle (set of states I_i^k). We evaluate the identity relation (see (19))

$$\sum_{B_i^k\cup I_i^k}\sum_{m=1}^{M}\sum_{j=1}^{M}(O_{m,j}(\vec{n},\vec{k}) - I_{m,j}(\vec{n},\vec{k})) = 0. \quad (20)$$

The value in the summation for the term $m=i$ and $j=i$ is obtained by observing that $\delta_i \equiv 1$ within B_i^k, thus $\sum_{B_i^k\cup I_i^k}\delta_i \equiv \sum_{B_i^k}$ and therefore

$$\sum_{B_i^k\cup I_i^k}\delta_i\Big(\sum_{\substack{h=1\\h\neq k}}^{K_i} q_{i,i}^{k,h}\pi(\vec{n},\vec{k}) - \sum_{\substack{h=1\\h\neq k}}^{K_i} q_{i,i}^{h,k}\pi(\vec{n},\vec{k}+(h-k)\vec{e}_i)\Big) = \sum_{\substack{h=1\\h\neq k}}^{K_i} q_{i,i}^{k,h}U_i^k - \sum_{\substack{h=1\\h\neq k}}^{K_i} q_{i,i}^{h,k}U_i^h.$$

Similarly, for $m=i$ and $j\neq i$, noting that summing $\delta_j q_{i,j}^{h,k}\pi(\vec{n}-\vec{e}_j+\vec{e}_i,\vec{k}+(h-k)\vec{e}_i))$ over all $B_i^k\cup I_i^k$ reduces to summing $\pi(\vec{n},\vec{k})$ on B_i^h we get

$$\sum_{B_i^k\cup I_i^k}\sum_{\substack{j=1\\j\neq i}}^{M}\sum_{h=1}^{K_i}\Big(q_{i,j}^{k,h}\delta_i\pi(\vec{n},\vec{k}) - q_{i,j}^{h,k}\delta_j\pi(\vec{n}-\vec{e}_j+\vec{e}_i,\vec{k}+(h-k)\vec{e}_i)\Big) = \sum_{\substack{j=1\\j\neq i}}^{M}\sum_{h=1}^{K_i}(q_{i,j}^{k,h}U_i^k - q_{i,j}^{h,k}U_i^h).$$

Let u be the phase of m in $\vec{k}$, then for $m\neq i$ and $1\leq j\leq M$

$$\sum_{B_i^k\cup I_i^k}\sum_{\substack{m=1\\m\neq i}}^{M}\sum_{j=1}^{M}\sum_{h=1}^{K_m}\Big(\delta_m q_{m,j}^{u,h}\pi(\vec{n},\vec{k}) - \delta_j q_{m,j}^{h,u}\pi(\vec{n}-\vec{e}_j+\vec{e}_m,\vec{k}+(h-u)\vec{e}_m)\Big) = \sum_{\substack{m=1\\m\neq i}}^{M}\sum_{j=1}^{M}\sum_{u=1}^{K_m}\sum_{h=1}^{K_m}\Big(q_{m,j}^{u,h}\sum_{n_m\geq 1}\pi_i^k(n_m,u) - q_{m,j}^{h,u}\sum_{n_m\geq 1}\pi_i^k(n_m,h)\Big) \equiv 0,$$

where the inner summation on u accounts for all possible phases of queue m within $B_i^k\cup I_i^k$. The terms sum to zero because the sums of $q_{m,j}^{h,u}\pi_i^k(n_m,h)$ and $q_{m,j}^{u,h}\pi_i^k(n_m,u)$ over all queue-lengths $n_m\geq 1$ and phases of m are identical. The final result follows immediately by inserting back into (20) the above formulas.

Detailed proof of Theorem 3

Let $S(k,n_i) \equiv \{(\vec{n}',\vec{k}'):n_i'\leq n_i,k_i'=k\}$, since the theorem requires $n_i\leq N-1$ there always exists the related set

$$\bar{S}(k,n_i) \equiv \{(\vec{n}',\vec{k}'):n_i'\geq n_i+1,k_i'=k\}.$$

The equilibrium probability flux exchanged by $\cup_{k=1}^{K_i}S(k,n_i)$ and $\cup_{k=1}^{K_i}\bar{S}(k,n_i)$ must be in balance because their union is the entire state space [27, p.351]. We seek for a representation of the exchanged probability flux using the marginal probabilities. The flux $\mathcal{F}$ from $\cup_{k=1}^{K_i}\bar{S}(k,n_i)$ to $\cup_{k=1}^{K_i}S(k,n_i)$ needs to decrease the queue-length of queue i to n_i-1. Only states where i has n_i+1 jobs can have transitions to states where i has n_i jobs (i.e., the MAP queueing network by definition does *not* allow batch completions); therefore $\mathcal{F}$ is the following flux of job completions

$$\mathcal{F} \equiv \sum_{\substack{j=1\\j\neq i}}^{M}\sum_{k=1}^{K_i}\sum_{h=1}^{K_i} q_{i,j}^{k,h}\sum_{\substack{(\vec{n}',\vec{k}')\in\\\{B_i^k\cap\bar{S}(k,n_i):\\n_i'=n_i+1\}}}\pi(\vec{n}',\vec{k}'),$$

which excludes the self-routed jobs (case $j=i$) that do not decrease n_i+1 to n_i. The opposite flux is

$$\mathcal{G} \equiv \sum_{\substack{j=1\\j\neq i}}^{m}\sum_{k=1}^{K_j}\sum_{h=1}^{K_j} q_{j,i}^{k,h}\sum_{u=1}^{K_i}\sum_{\substack{(\vec{n}',\vec{k}')\in\\\{B_j^k\cap\bar{S}(u,n_i):\\n_i'=n_i\}}}\pi(\vec{n}',\vec{k}'),$$

and describes all possible transitions that bring a job from queue $j\neq i$ to i for all possible phases u of i in $\vec{k}$. To account for the single step behavior, we have imposed that the population in i is of n_i jobs. However, by the given definitions for the busy condition reduction

$$\sum_{(\vec{n}',\vec{k}')\in\{B_i^k\cap\bar{S}(k,n_i):n_i'=n_i+1\}}\pi(\vec{n}',\vec{k}') = \pi_i^k(n_i+1,k),$$
$$\sum_{(\vec{n}',\vec{k}')\in\{B_j^k\cap\bar{S}(u,n_i):n_i'=n_i\}}\pi(\vec{n}',\vec{k}') = \pi_j^k(n_i,u),$$

and imposing the equilibrium balance $\mathcal{F}=\mathcal{G}$ for $n_i\geq 1$, we find immediately (7). Note that (7) would hold also for $n_i=0$; nevertheless, in this case we can give the more detailed condition (8) by recalling that if $n_i=0$ phase transitions in i are impossible, hence the balance $\mathcal{F}=\mathcal{G}$ splits into a set of disjoint probability flux balances, one for each phase u of i. The proof in this case is almost identical by considering the interface between the sets $S(u,n_i=0)\equiv\{(\vec{n}',\vec{k}'):n_i'\leq 0,k_i'=u\}$ and $\cup_{k=1}^{K_i}\bar{S}(k,n_i=1)$.

Detailed proof of Theorem 5

Let k be the phase of i in $\vec{k}$ and consider the weighted sum of the global balance equations

$$\sum_{B_i^k\cup I_i^k} n_i\sum_{m=1}^{M}\sum_{j=1}^{M}(O_{m,j}(\vec{n},\vec{k}) - I_{m,j}(\vec{n},\vec{k})). \quad (21)$$

The high-level structure of the proof is analogous to the proof of (6), thus we first consider the term for $m=j=i$

$$\sum_{B_i^k\cup I_i^k} n_i\delta_i\Big(\sum_{\substack{h=1\\h\neq k}}^{K_i} q_{i,i}^{k,h}\pi(\vec{n},\vec{k}) - \sum_{\substack{h=1\\h\neq k}}^{K_i} q_{i,i}^{h,k}\pi(\vec{n},\vec{k}+(h-k)\vec{e}_i)\Big) = \sum_{\substack{h=1\\h\neq k}}^{K_i} q_{i,i}^{k,h}Q_i^k - \sum_{\substack{h=1\\h\neq k}}^{K_i} q_{i,i}^{h,k}Q_i^h.$$

Similarly, for $m = i$ and $j \neq i$,

$$\begin{aligned}
&\textstyle\sum_{B_i^k \cup I_i^k} n_i \sum_{\substack{j=1 \\ j\neq i}}^{M} \sum_{h=1}^{K_i} (q_{i,j}^{k,h} \delta_i \pi(\vec{n}, \vec{k}\,) \\
&\qquad - q_{i,j}^{h,k} \delta_j \pi(\vec{n} - \vec{e}_j + \vec{e}_i, \vec{k} + (h-k)\vec{e}_i\,)) \\
&\qquad = \textstyle\sum_{B_i^k \cup I_i^k} \sum_{\substack{j=1 \\ j\neq i}}^{M} \sum_{h=1}^{K_i} q_{i,j}^{k,h} \delta_i n_i \pi(\vec{n}, \vec{k}\,) \\
&\qquad\qquad - \textstyle\sum_{B_i^h} q_{i,j}^{h,k} (n_i - 1) \pi(\vec{n}, \vec{k}\,) \\
&\qquad = \textstyle\sum_{\substack{j=1 \\ j\neq i}}^{M} \sum_{h=1}^{K_i} (q_{i,j}^{k,h} Q_i^k + q_{i,j}^{h,k} U_i^h - q_{i,j}^{h,k} Q_i^h) \\
&= \textstyle\sum_{\substack{j=1 \\ j\neq i}}^{M} \sum_{h=1}^{K_i} q_{i,j}^{h,k} U_i^h + \sum_{\substack{j=1 \\ j\neq i}}^{M} \sum_{\substack{h=1 \\ h\neq k}}^{K_i} (q_{i,j}^{k,h} Q_i^k - q_{i,j}^{h,k} Q_i^h).
\end{aligned}$$

Let u be the phase of m in $\vec{k}$, then for $m \neq i$ and $1 \leq j \leq M$

$$\begin{aligned}
&\textstyle\sum_{B_i^k \cup I_i^k} n_i \sum_{\substack{m=1 \\ m\neq i}}^{M} \sum_{j=1}^{M} \sum_{h=1}^{K_m} (\delta_m q_{m,j}^{u,h} \pi(\vec{n}, \vec{k}\,) \\
&\qquad - \delta_j q_{m,j}^{h,u} \pi(\vec{n} - \vec{e}_j + \vec{e}_m, \vec{k} + (h-u)\vec{e}_m\,)) \\
&= \textstyle\sum_{\substack{m=1 \\ m\neq i}}^{M} \sum_{j=1}^{M} \sum_{u=1}^{K_m} \sum_{h=1}^{K_m} (q_{m,j}^{u,h} \sum_{B_m^u \cap \{B_i^k \cup I_i^k\}} n_i \pi_i^k(n_m, u) \\
&\qquad\qquad - q_{m,j}^{h,u} (n_i + \delta_{j=i}) \textstyle\sum_{B_m^u \cap \{B_i^k \cup I_i^k\}} \pi_i^k(n_m, h)),
\end{aligned}$$

where $\delta_{j=i}$ is one if $j = i$, zero otherwise. In the last summation, all terms for $j \neq i$ simplify to zero and thus the final value is

$$\begin{aligned}
&\textstyle\sum_{\substack{m=1 \\ m\neq i}}^{M} \sum_{u=1}^{K_m} \sum_{h=1}^{K_m} (q_{m,i}^{u,h} \sum_{B_m^u \cap \{B_i^k \cup I_i^k\}} n_i \pi_i^k(n_m, u) \\
&\qquad - q_{m,i}^{h,u} (n_i + 1) \textstyle\sum_{B_m^u \cap \{B_i^k \cup I_i^k\}} \pi_i^k(n_m, h)) \\
&= - \textstyle\sum_{\substack{m=1 \\ m\neq i}}^{M} \sum_{u=1}^{K_m} \sum_{h=1}^{K_m} (q_{m,i}^{h,u} \sum_{B_m^u \cap \{B_i^k \cup I_i^k\}} \pi_i^k(n_m, h)) \\
&\qquad = - \textstyle\sum_{\substack{m=1 \\ m\neq i}}^{M} \sum_{u=1}^{K_m} \sum_{h=1}^{K_m} (q_{m,i}^{h,u} J_i^k(m, h).
\end{aligned}$$

The final result follows immediately by inserting back into (21) the above formulas and replacing in the last formula, without loss of generality, the index m with j.