Approximate Analysis of Blocking Queueing Networks with Temporal Dependence

Vittoria de Nitto Personé
University of Rome Tor Vergata, Italy
denitto@info.uniroma2.it

Giuliano Casale
Imperial College London, U.K.
g.casale@imperial.ac.uk

Evgenia Smirni
College of William and Mary, VA, U.S.A.
esmirni@cs.wm.edu

***Abstract*—In this paper we extend the class of MAP queueing networks to include blocking models, which are useful to describe the performance of service instances which have a limited concurrency level. We consider two different blocking mechanisms: Repetitive Service-Random Destination (RS-RD) and Blocking After Service (BAS). We propose a methodology to evaluate MAP queueing networks with blocking based on the recently proposed Quadratic Reduction (QR), a state space transformation that decreases the number of states in the Markov chain underlying the queueing network model. From this reduced state space, we obtain boundable approximations on average performance indexes such as throughput, response time, utilizations. The two approximations that dramatically enhance the QR bounds are based on maximum entropy and on a novel minimum mutual information principle, respectively. Stress cases of increasing complexity illustrate the excellent accuracy of the proposed approximations on several models of practical interest.**

I. Introduction

Blocking is the phenomenon where an IT service may not be available for a period of time, therefore any request for this service has to wait until the service becomes available again. This service unavailability can stem from a physical limit (e.g., memory or concurrency constraints) or it can relate to a system management decision in order to overcome an overload period and to guarantee QoS requirements. Consequently, blocking can affect system performance significantly. Despite its importance, blocking is a difficult phenomenon to model analytically, because it creates strong inter-dependencies in the system's components. The blocking concept can be summarized as follows: when a queue reaches its maximum capacity, then the flow of customers entering the queue is stopped. Queueing networks with blocking have been used to model telecommunication and computer systems with limited shared resources, such as interconnecting links or store-and-forward buffers, as well as production systems with finite storage buffers. We point the interested reader to [3], [16], [17], [18] for an extensive bibliography of different blocking mechanisms that model distinct behaviors of real systems including computer systems [9], communication systems and networks [1], [7], and software architectures [2].

Despite the practical applications of blocking queueing models, there is a lack of robust methodologies for their solution, which stems from the fact that general blocking queueing networks are not separable. The problem is worsened if the service processes of the various stations are non-renewal, a case of increasing importance to represent real systems such as multi-tier applications [13], [19]. Given the fact that blocking creates performance dependencies that are hard to understand without a sound methodology, there is a clear need for robust and general solutions.

In this paper, we provide a robust approximation methodology for various performance measures for MAP queueing networks with blocking, which relies on numerical optimization techniques and that enjoys errors bounds. In particular, we focus on the case of networks with a closed population of jobs that are the most important for sizing computer systems that have upper limits on the maximum number of concurrent users, and generalize the class of MAP queueing networks proposed in [5] to include blocking mechanisms. MAP queueing networks admit service processes that are described by Markovian Arrival Processes (MAPs), a class of Markov-modulated point processes that can model general distributions as well as the main features of non-renewal workloads, such as autocorrelation in service times or burstiness. Naturally, MAP queueing networks do not admit product form solutions and can be viewed as a generalization of non-product form networks with renewal service processes. In [5] the quadratic reduction (QR) bounding methodology for the solution of MAP queueing networks has been proposed. Applying the QR bounds to MAP blocking networks is a challenging problem because in presence of blocking the state space often differs significantly with respect to the original MAP queueing network state space, thus the QR characterizations obtained in [5] is not directly applicable anymore. The contributions of this paper can be summarized as follows:

- we provide a major extension to the *Quadratic Reduction* (QR) technique first introduced in [5] by including blocking.
- we introduce approximations based on maximum entropy [11] and a novel minimum mutual information principle that are shown to accurately predict model performance with only small error that dramatically improve the quality of the extension of the QR technique for MAP blocking networks.

Throughout this paper we consider a closed queueing network with routing matrix P such that jobs departing from queue i are directed to queue j with probability p_{ij}. If the capacity of queue j is F_j and n_j denotes the current population at queue j, then when $n_j = F_j$ queue j does not accept in its waiting buffer any *new* job before a departure occurs. Here, we consider the Blocking After Service (BAS) and the Repetitive Service-Random Destination (RS-RD) mechanisms [3].

- *Blocking After Service (BAS)* A queue i, if not empty, processes a job regardless of the job population at its destination j. When node i completes service and node j is full, node i suspends any activity (i.e., it is blocked) and the completed job waits until a departure occurs from node j. At that moment two simultaneous transitions take place: the completed/blocked job moves from i to j (since j can now accept a job, i "unblocks") and the job that leaves j (which effectively "unblocks" server i). In a general network topology where more than one queue compete for sending a job towards a full queue j, a policy regulating the order in which queues unblock has to be defined. Usually, the First Blocked First Unblocked (FBFU) policy is considered fair: first unblock the queue that was blocked first. In the remaining of this paper when we consider BAS we assume that it uses the FBFU policy. BAS models production systems and disk I/O subsystems [20].
- *Repetitive Service-Random Destination (RS-RD)* A queue i, if not empty, processes a job regardless of the job population at its destination j. If node j is full, the completed job is rerouted to node i where it receives a new service. During the new service, the job may select a destination that is independent from its previous one. Note that according to RS-RD blocking a node is never actually blocked, but it "wastes" its service by repeating it. RS-RD blocking is used to model congestion control in telecommunications systems [1].

The above two blocking mechanisms introduce complexity in the underlying Markov chains of MAP queueing networks. On one hand, RS-RD restricts the original state space, while preserving its regular structure, on the other hand BAS introduces new states describing the order in which queues progressively block once the capacity of the destination node becomes full, this information is needed to implement the FBFU rule. At a higher level, flows in a MAP network with blocking are harder to understand than in a MAP network without blocking because of the additional routing complication that is introduced.

Here, we incorporate additional information in the QR marginal probabilities and obtain a new class of specialized conditions that allows to represent the simultaneous unblocking and departure events that happen upon completion from a node that is full. Such conditions can accomodate Markov-modulated service rates, thus integrating within blocking models complex features such as higher-order moments and temporal dependence yet are not sufficient to result in tight bounds for blocking systems. The two approximation techniques that we introduce for blocking networks here, i.e., the maximum entropy method (MEM) and the minimal mutual information (MIM), can be used to "correct" the QR bounds by driving the estimation of the equilibrium probabilities of the model using nonlinear optimization. This correction shows to be very effective to address simultaneously (and very effectively) two difficult problems: the complex features of blocking networks *and* the complexities introduced by temporal dependence.

TABLE I
SUMMARY OF MAIN NOTATION FOR MAP QUEUEING NETWORKS WITH BAS BLOCKING

b	cardinality of the list of blocked queues $\boldsymbol{m}$
b_i	blocking state of node i
B	maximum number of queues that can block on f
f	finite capacity queue
F_i	capacity of queue i
K_i	phases in the MAP service process of queue i
k_i	phase of the MAP service process of queue i
$\boldsymbol{m}$	list of queues blocked by f
$Add(\boldsymbol{m}, j)$	list obtained by adding j to the tail of $\boldsymbol{m}$
$Head(\boldsymbol{m})$	first queue to unblock after a departure from f that is not self-routed
M	number of queues in the network
N	number of jobs in the network
n_i	number of jobs at queue i
$\pi(n_i, k_i, n_j, k_j, \boldsymbol{m})$	prob. of n_i jobs in queue i in phase k_i and n_j jobs in queue j in phase k_j and $\boldsymbol{m}$ blocked
$q_{i,j}^{k_i,k'_i}$	rate of job departures from i to j when i's MAP is in phase k_i leaving it in phase k'_i

The rest of this paper is organized as follows. In Section II we define MAP queueing networks with BAS blocking and develop their analytical characterization by means of the QR state space reduction. Section III extends the analysis to RS-RD blocking. Section IV discusses the performance approximations and bounds following from the characterization of the QR marginal probabilities and illustrate them on a set of models with BAS and RS-RD blocking. Section V presents a set of experiments that illustrate the proximity of the two approximations to exact solutions. Finally, Section VI gives conclusions and outlines future work.

II. MAP QUEUEING NETWORKS WITH BAS BLOCKING

We introduce the class of MAP queueing networks supporting temporal dependent service. We first present the case of the BAS blocking mechanism. The RS-RD blocking mechanism is simpler and it is discussed in Section III.

We consider a closed MAP queueing network with N jobs visiting M single-server queues having first-come first-serve scheduling. For each queue i, its service time process is a load-independent Markovian Arrival Process [6]. To reduce the complexity of the notation, we present for the BAS case only where a single queue f has finite capacity $F_f < N$ and its sending nodes behave according to BAS blocking, while all other queues $i \neq f$ have infinite capacity (i.e., $F_i = N$) such that they can accommodate all jobs in the network. The generalization to networks with several finite capacity queues is not difficult and follows an identical argument as shown in [8]. RS-RD blocking is instead discussed in the general case as it does not complicate notation significantly.

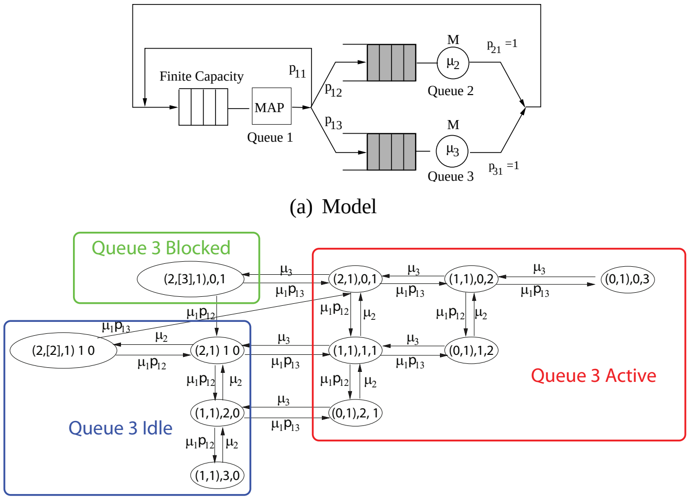
A. State Space for BAS blocking

A summary of the main notation is given in Table I. This is consistent with the original notation defined for MAP queueing networks [5] that it is here briefly reviewed.The service process at queue i is modeled by a MAP with $K_i \geq 1$ phases. It is worth noting that if a queue is blocked, it completely stops its activity including MAP phase transitions. Note that this holds for BAS blocking, but not for RS-RD where a queue is never effectively blocked. Furthermore, 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 $\mathbf{s} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_M)$, where for queue $i \neq f$ the local state $\mathbf{s}_i = (n_i, b_i, k_i)$ is defined as follows: n_i is the current queue-length including the job in service; b_i is the blocking state of node i (1=blocked, 0=active); $k \in K_i$ is the phase of queue i. Conversely, for the finite capacity queue f the state is $\mathbf{s}_f = (n_f, \boldsymbol{m}, k_f)$ where $\boldsymbol{m} = (m_1, m_2, \ldots, m_b)$ is a list that holds the sequence of $b = \sum_{i \neq f} b_i$ queues that can be unblocked by a departure from queue f in state s. The index b thus denotes the current number of blocked queues and ranges in $0, \ldots, B$, where $B \leq M - 1$ is the number of queues $i \neq f$ that can send jobs to f. Note that b_f is not needed since it is often assumed in the literature that a finite capacity queue is never blocked by itself. The case $\boldsymbol{m} = \emptyset$ denotes that no queue is blocked by f. We denote with $Head(\boldsymbol{m})$ the head of the list, i.e., the first queue to unblock upon a departure from f that is not self-routed, and with $Add(\boldsymbol{m}, j)$ the list resulting from the addition of element j to the tail of $\boldsymbol{m}$. Finally, let E_{BAS} be the state space of the queueing network assuming BAS blocking at each node $i \neq f$ that can send jobs to f ($p_{i,f} > 0$). In this state space, the Markov process transitions have rates from state $\mathbf{s} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_M)$ to $\mathbf{s}' = (\mathbf{s}'_1, \mathbf{s}'_2, \ldots, \mathbf{s}'_M)$ that are uniquely defined by the rates $q_{i,j}^{k,h}$ of jobs flowing from i to j in phase k leaving i in phase h.

The size of the infinitesimal generator corresponds to the cardinality of the related global balance equations. By considering only the population components n_i, the state space of a blocking network is a subset of the state space of the same network but with infinite capacity queues. This is logical because all states with $n_f > F_f$ do not exist. On the other hand, the order in which queues block needs to be accounted for explicitly in $\boldsymbol{m}$, which increases the state space cardinality. Thus, the state space of a BAS queueing network can be smaller or bigger than in the non-blocking case depending on the number of queues and jobs being considered.

We now give examples of the state space underlying a MAP queueing network with BAS blocking. For the sake of simplicity, we omit the state component b_i since it can be simply derived, i.e., $b_i = 1$ if and only if $i \in \boldsymbol{m}$. We consider an example model with $M = 3$ queues where queue 1 is a finite capacity station with MAP service, queues 2 and 3 have exponential service and infinite capacities. Figure 1(a) illustrates the model with routing probabilities. The exponential queues have rates μ_2 and μ_3, the MAP completes jobs in phase 1 with rate μ_1.

The underlying Markov process for the case with $N = 3$, $F_1 = 2$ and assuming queue 1 in phase 1, is shown in Figure 1(b). For ease of illustration, MAP phase change transitions are omitted, thus this partition is similar to the state space where the service at the MAP is exponential with rate μ_1. We point to [5] for figures illustrating the effects of phase changes in the MAP queueing network state space. Figure 1(b) classifies the activity of queue 3 into "active" ($n_3 > 0$ and $b_3 = 0$), "idle" ($n_3 = 0$ and $b_3 = 0$), or "blocked" ($n_3 > 0$ and $b_3 = 1$). This classification is useful to understand the different rates of departure from queue 3 across the state space. The states where queue 3 is active are the only states that contribute to the standard departure transitions out from queue 3. The state $((2, [3], 1), 0, 1)$ in the blocked subspace denotes the case where queue 3 is blocked ($\boldsymbol{m} = [3]$) since queue 3 has previously completed a job to be sent to queue 1 while this was full. As soon as queue 1 completes a job, two simultaneous transitions take place moving the current state to $((2, 1), 1, 0)$ in the idle subspace of queue 3 if the job completed by queue 1 is routed to queue 2. The current state becomes $((2, 1), 0, 1)$ in the active subspace of queue 3 if the completed job is routed to queue 3 which thus restarts immediately service after unblocking. Such simultaneous transitions are a distinctive characteristic of the state space due to BAS blocking.

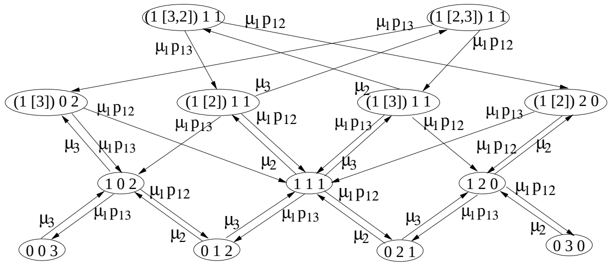


(a) Model

(b) State space for $N = 3$ and $F_1 = 2$ assuming queue 1 in phase 1. State notation is $((n_1, \boldsymbol{m}, k_1), n_2, n_3)$, $\boldsymbol{m} = \emptyset$ is omitted.

Fig. 1. **Example network with 2 infinite capacity exponential queues and a MAP queue with finite capacity**

To further appreciate the complexity of bound analysis for the BAS state space, Figure 2 illustrates a case where two queues can be blocked. Observe the changes in the BAS state space level compared to Figure 1. Let us now assume $F_1 = 1$, infinite capacities for queues 2 and 3, and $N = 3$ jobs in the network. For simplicity of graphical representation, the phase k_1 is omitted being always equal to 1. Figure 2 shows the totally different structure of the state space. When the system is in the state $(1, 1, 1)$ all queues are active. If queue 2 completes a job, the current state becomes $((1, [2], 1), 1, 1)$ where queue 2 is blocked ($\boldsymbol{m} = [2]$). If from this state queue 3 completes a job, the transition leads to state $((1, [2, 3], 1), 1, 1)$, where *both* queues 2 and 3 are blocked



(a) State space for $N = 3$ and $F_1 = 1$ assuming queue 1 in phase 1. State notation is $((n_1, \boldsymbol{m}, k_1), n_2, n_3)$, $\boldsymbol{m} = \emptyset$ and all phases $k_1 = 1$ are omitted.

Fig. 2. **Example model 1 when several queues are blocked**

($\boldsymbol{m} = [2, 3]$). According to the FBFU unblocking rule, when queue 1 completes a job, queue 2 is unblocked first with a transition to the state $((1, [3], 1), 1, 1)$ if the completed job is routed to queue 2, or to state $((1, [3], 1), 0, 2)$ if the completed job is routed to queue 3. This illustrates transitions that do not exist in non-blocking queueing networks and thus which require specialized characterization for bounding purposes. To obtain such a characterization, we develop in Theorem 5 a new class of balance conditions that is able to describe also the state space illustrated in Figure 2.

B. Quadratic Reduction (QR) of BAS State Space

Denote with $\pi(\mathbf{s})$ the equilibrium probability of state $\mathbf{s} \in E_{BAS}$ in the blocking MAP queueing network. We formulate the quadratic reduction for the BAS case as follows. We consider the following marginal probability

$$\pi(n_i, k_i, n_j, k_j, \boldsymbol{m}) \tag{1}$$

which is called *QR marginal probability* and describes the joint state of queues i and j in phases k_i and k_j while the queues in $\boldsymbol{m}$ are blocked by f. This formula is immediately obtained by summing $\pi(\mathbf{s})$ over the states with the considered values of n_i, k_i, n_j, k_j, and $\boldsymbol{m}$. The main advantage of the QR marginal probability over the original state space representation is that is scales only quadratically with the total population size, which is by far the largest parameter of the queueing network model specification. Thus, this provides substantial savings with respect to a direct state space evaluation by global balance that involves $O(N^M)$ unknown probabilities.

The goal of the next sections is to develop a characterization of the balance conditions that relate different values of the QR marginal probabilities. Previous work has shown that relations between marginal probabilities can be insufficient for an exact solution of the queueing network model, but they can be exploited to determine bounds on performance indexes using linear programming [5]. We show later in Section IV that this holds true also for MAP queueing networks both with BAS and RS-RD blocking.

Common metrics such as utilization, throughput, response times, and queue-lengths can be immediately computed from the QR marginal probabilities. For example, the utilization of queue i is

$$U_i = \sum_{(n_j, k_j, k_i, \boldsymbol{m})} \sum_{n_i \geq 1} \pi(n_i, k_i, n_j, k_j, \boldsymbol{m})$$

whereas the effective utilization that describes the activity of a queue *excluding its blocking time* is

$$E_i = \sum_{k_i=1}^{K_i} E_i^{k_i}$$

where the effective utilization [3] of phase k_i is

$$E_i^{k_i} = \sum_{(n_j, k_j, \boldsymbol{m})} \sum_{n_i \geq 1} \sum_{i \notin \boldsymbol{m}} \pi(n_i, k_i, n_j, k_j, \boldsymbol{m})$$

Note that $E_i = U_i$ if and only if $i = f$ or i cannot be blocked by f due to the network topology. The effective utilization takes into account the productive utilization of a queue, that is the period of time the queue is busy and it is not blocked, so it can produce useful work. Other measures such as mean queue-lengths, throughput, or response times are similarly defined. For example, the throughput may be obtained as an effective utilization divided by the product of mean number of visits and mean service times [3].

C. Basic Characterization Results

The first basic characterization result for QR marginal probabilities in a BAS setting follows by the equilibrium of the MAP service processes. During the period where queue i is actively serving a job, the MAP service process behaves at equilibrium in the same way of the same MAP considered in isolation, since we are assuming that the queue is never idle nor blocked. This equivalence introduces a balance between QR marginal probabilities relative to different phases.

Theorem 1: The effective utilization of queue i for phase k satisfies at equilibrium

$$\sum_{j=1}^{M} \sum_{\substack{h=1 \\ h \neq k \text{ if } j=i}}^{K_i} q_{i,j}^{k,h} E_i^k = \sum_{j=1}^{M} \sum_{\substack{h=1 \\ h \neq k \text{ if } j=i}}^{K_i} q_{i,j}^{h,k} E_i^h, \tag{2}$$

for $i = 1, \ldots, M$ and $k = 1, \ldots, K_i$.

Proof: Consider a partitioning of the state space into two subsets: $G_{i,k}$ where queue i is in phase k and its complementary set of states $\bar{G}_{i,k}$ where queue i is in phase $h \neq k$. By basic properties of Markov processes, the equilibrium probability flux exchanged by $G_{i,k}$ and $\bar{G}_{i,k}$ at equilibrium must be balanced. However, this is only due to phase changes that occur in the MAP, with or without an associated departure from i. The left hand side represents phase changes moving the current phase from k to any h, whereas the right hand side is the probability flux due to phase changes that move the active phase into k. Note that the condition $h \neq k$ if $j \neq i$ ignores phase self-routing of jobs that do not change the active phase. ■

Another characterization result follows by observing that the total population of jobs in each state of the underlying Markov chain sums to N. This implies that the sum of the conditional first moment of queue lengths is constant.

Theorem 2: The QR marginal probabilities for BAS blocking satisfy the following constraints

$$\sum_{i=1}^{M}\sum_{k_i=1}^{K_i}\sum_{n_i=1}^{F_i} n_i\pi(n_i,k_i,n_j,k_j,\boldsymbol{m}) = N\pi(n_j,k_j,\boldsymbol{m}) \qquad (3)$$

for all $j = 1,\dots,M$, $n_j = 0,\dots,F_j$, $k_j = 1,\dots,K_j$, and for all lists of blocked queues $\boldsymbol{m}$.

Proof: Since the sum of the total population in a state is constant, we can write

$$\sum_{\boldsymbol{s}\in S}\sum_{i=1}^{M} n_i\pi(\boldsymbol{s}) = N\sum_{\boldsymbol{s}\in S}\pi(\boldsymbol{s}) \qquad (4)$$

for any partition of states $S \subseteq E_{BAS}$, where we omit $n_i = 0$ since the corresponding term in the summation is zero. Define S as the set of states where the blocked queue list is $\mathbf{m}$ and queue j has population n_j in phase k_j, thus the right hand side becomes $N\pi(n_j,k_j,\boldsymbol{m})$. Denote by $\boldsymbol{s}_k$ the components of $\boldsymbol{s}$ different from n_i, k_i, n_j, k_j, $\boldsymbol{m}$. We can equivalently rewrite the above expression as

$$\sum_{i=1}^{M}\sum_{k_i=1}^{K_i}\sum_{n_i=1}^{F_i} n_i\sum_{\boldsymbol{s}_k}\pi(n_i,k_i,n_j,k_j,\boldsymbol{s}_k,\boldsymbol{m}) = N\pi(n_j,k_j,\boldsymbol{m}) \qquad (5)$$

However, the inner summation on $\boldsymbol{s}_k$ gives the QR marginal $\pi(n_i,k_i,n_j,k_j,\boldsymbol{m})$ which proves the theorem. ■

The above characterization generalizes in a weaker form also to second-order queue-length moments.

Corollary 1: The second-order joint moments of queue-lengths in a MAP network with BAS blocking satisfy

$$\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k_i=1}^{K_i}\sum_{k_j=1}^{K_j}\sum_{n_i=1}^{F_i}\sum_{n_j=1}^{F_j}\sum_{\boldsymbol{m}} n_in_j\pi(n_i,k_i,n_j,k_j,\boldsymbol{m}) = N^2 \qquad (6)$$

Proof: Using the same argument of Theorem 2 we have

$$\sum_{\boldsymbol{s}\in S}\left(\sum_{i=1}^{M} n_i\right)^2\pi(\boldsymbol{s}) = N^2\sum_{\boldsymbol{s}\in S}\pi(\boldsymbol{s}) \qquad (7)$$

thus the result follows immediately setting $S = E_{BAS}$. Note that a similar formula holds also without blocking [5]. ■

We remark that a higher-order extension of the above theorem holds as well, but it cannot be represented explicitly using the QR marginal probabilities since a order-k formula requires the joint probability of k queue-length terms, and QR can express such relations only for $k \leq 2$.

The next theorem can be seen as an extension of Theorem 2 as it defines a relation between the sum of mean queue-lengths of all queues and the utilization of queue i when a given queue j is in phase k_j.

Theorem 3: The sum of the mean queue-lengths of all queues conditioned on queue j being in phase k_j satisfies

$$\sum_{\boldsymbol{m}}\sum_{n_j=0}^{F_j}\sum_{w=1}^{M}\sum_{n_w=1}^{F_w}\sum_{k_w=1}^{K_w} n_w\pi(n_w,k_w,n_j,k_j,\boldsymbol{m}) \geq N\sum_{\boldsymbol{m}}\sum_{n_j=0}^{F_j}\sum_{n_i=1}^{F_i}\sum_{k_i=1}^{K_i}\pi(n_i,k_i,n_j,k_j,\boldsymbol{m}) \qquad (8)$$

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

The proof is qualitatively similar to the one used for non-blocking MAP queueing networks, we point the interested reader to [8, Thm. 4] for a complete derivation.

D. Marginal Balance Conditions

The theorems in the previous section provide a characterization of basic properties of utilization and queue-lengths in the QR marginal representation. However, these properties depend very loosely on the inter-dependencies between stations, such as the flows of jobs between queues and the rules of BAS blocking. A strong characterization of BAS blocking and job flows is provided by the following marginal balance conditions. Such conditions express (by the QR marginals) the probability flux balance resulting from cuts of the Markov chain that separate states with a marginal population n_i from those where queue i has population $n_i + 1$.

Theorem 4 (Marginal balance): The arrival flow of queue i when the local queue-length is of n_i jobs, $0 < n_i \leq F_i - 1$, is in equilibrium with the departure flow when the queue-length is $n_i + 1$, i.e.,

$$\begin{aligned}
&\sum_{\substack{j=1\\ j\neq i\\ j\neq f}}^{M}\sum_{n_j=1}^{F_j}\sum_{k=1}^{K_j}\sum_{h=1}^{K_j}\sum_{v=1}^{K_i}\sum_{\boldsymbol{m}:j\notin\boldsymbol{m}} q_{j,i}^{k,h}\pi(n_j,k,n_i,v,\boldsymbol{m})\\
&+\sum_{n_f=1}^{F_f}\sum_{k=1}^{K_f}\sum_{h=1}^{K_f}\sum_{v=1}^{K_i}\sum_{\boldsymbol{m}:Head(\boldsymbol{m})\neq i} q_{f,i}^{k,h}\pi(n_f,k,n_i,v,\boldsymbol{m})\\
&=\sum_{\substack{j=1\\ j\neq i}}^{M}\sum_{n_f=0}^{F_f-1}\sum_{k=1}^{K_i}\sum_{h=1}^{K_i}\sum_{v=1}^{K_f} q_{i,j}^{k,h}\pi(n_f,v,n_i+1,k,\emptyset)\\
&+\sum_{\substack{j=1\\ j\neq i\neq f}}^{M}\sum_{k=1}^{K_i}\sum_{h=1}^{K_i}\sum_{v=1}^{K_f}\sum_{\boldsymbol{m}:i\notin\boldsymbol{m}} q_{i,j}^{k,h}\pi(F_f,v,n_i+1,k,\boldsymbol{m})\\
&+\sum_{k=1}^{K_i}\sum_{v=1}^{K_f}\sum_{p=1}^{K_f}\sum_{\boldsymbol{m}:Head(\boldsymbol{m})=i}\sum_{\substack{w=1\\ w\neq f\\ w\neq i}}^{M} q_{f,w}^{v,p}\pi(F_f,v,n_i+1,k,\boldsymbol{m}),
\end{aligned} \qquad (9)$$

for $i = 1,\dots,M$ and $i \neq f$. When $i = f$ the expression becomes

$$\begin{aligned}
&\sum_{\substack{j=1\\ j\neq f}}^{M}\sum_{n_j=1}^{F_j}\sum_{k=1}^{K_j}\sum_{h=1}^{K_j}\sum_{v=1}^{K_f} q_{j,f}^{k,h}\pi(n_j,k,n_f,v,\emptyset)\\
&=\sum_{\substack{j=1\\ j\neq f}}^{M}\sum_{n_j=0}^{F_j-1}\sum_{k=1}^{K_f}\sum_{h=1}^{K_f}\sum_{v=1}^{K_j} q_{f,j}^{k,h}\pi(n_j,v,n_f+1,k,\emptyset),
\end{aligned} \qquad (10)$$

for each $n_f = 0,\dots,F_f - 1$. Furthermore, for $n_i = 0$ (equiv. $n_f = 0$ when $i = f$) the above balances admit a stronger form where they hold true for each phase $k = 1,\dots,K_i$ considered in isolation.

Proof: The proof is based on the definition of the equilibrium probability flux exchanged between states with n_i and with $n_i + 1$ jobs in queue i. First, consider $i \neq f$. The left hand side of equation (9) includes all departures from any non-empty queue j (i.e., $n_j > 0$) toward queue i. After these departures, the population of i becomes $n_i + 1$, except in the case where $j = f$ and $Head(\mathbf{m}) = i$, i.e., queue i is unblocked by the departure from f. In this case, queue i is waiting for free space in f and, because of the simultaneous transitions, the population in i remains equal to n_i. As a consequence, when $j = f$, the condition $Head(\mathbf{m}) \neq i$ must be also true, this corresponds to the second term of the left side of (9).

The right hand side of the equation considers all departures from queue i with population equal to $n_i + 1$. After these departures, i's population becomes n_i. These departures include:

- Case a: Transitions from i towards any queue j, $j \neq i$. Note that these transitions are always possible because queue j does not have finite capacity, and for queue f this transition can occur when $n_f < F_f$; this is the first term of the right side. When queue f is full, a transition from i is still possible if queue i is not blocked, that is i is not in the $\mathbf{m}$ list; this case corresponds to the second term of the right side of (9).
- Case b: Transitions from node f to any other node w, $w \neq f$, $w \neq i$ when f is full and node i is the first blocked one, that is $Head(\mathbf{m}) = i$. These transitions trigger a simultaneous transition from queue i, thus decrease its population to n_i. This is the third term on the right side of (9).

Let $S(k, n_i) \equiv \{\mathbf{s} = (\mathbf{s_1}, \mathbf{s_2}, \ldots \mathbf{s_M}) | \mathbf{s_i} : n'_i \leq n_i, k'_i = k\}$. Since the theorem requires $n_i \leq F_i - 1$, there always exists the related set $\bar{S}(k, n_i) \equiv \{\mathbf{s} = (\mathbf{s_1}, \mathbf{s_2}, \ldots \mathbf{s_M}) | \mathbf{s_i} : n'_i \geq n_i + 1, k'_i = k\}$. 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. We seek for a representation of the exchanged probability flux using the QR marginal probabilities. The flux 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. By considering that batch completions are not allowed, these transitions correspond to the two cases described above. Therefore, F is the following flux of completions:

$$\begin{aligned} F \equiv & \sum_{j=1, j\neq i}^{M} \sum_{n_f=0}^{F_f - 1} \sum_{k=1}^{K_i} \sum_{\mathbf{s}': n'_i = n_i + 1, n'_f = n_f} \sum_{h=1}^{K_i} q_{i,j}^{k,h} \pi(\mathbf{s}') \\ & + \sum_{j=1, j\neq i, j \neq f}^{M} \sum_{k=1}^{K_i} \sum_{\mathbf{s}': n'_i = n_i + 1, n'_f = F_f, i \notin \mathbf{m}} \sum_{h=1}^{K_i} q_{i,j}^{k,h} \pi(\mathbf{s}') \\ & + \sum_{\nu=1}^{K_f} \sum_{p=1}^{K_f} \sum_{\substack{\mathbf{s}': n'_i = n_i + 1, n'_f = F_f, \\ Head(\mathbf{m}) = i}} \sum_{w=1, w \neq i, w \neq f}^{M} q_{f,w}^{\nu,p} \pi(\mathbf{s}') \qquad (11) \end{aligned}$$

which excludes the self-routed jobs (i.e., $j = i$) that naturally do not decrease $n_i + 1$ to n_i. The opposite flux G needs to increase the queue-length of queue i to $n_i + 1$. Transitions towards states where i has $n_i + 1$ are allowed provided that the following conditions hold: the sending queue j is not empty and if $j = f$, $Head(\mathbf{m}) \neq i$ so that a simultaneous transition does not happen. The flux G represents all transitions from queue j to i, $j \neq i$.

$$\begin{aligned} G \equiv & \sum_{j=1, j\neq i, j\neq f}^{M} \sum_{k=1}^{K_j} \sum_{h=1}^{K_j} q_{j,i}^{k,h} \sum_{\mathbf{s}': n'_j > 0, n'_i = n_i} \pi(\mathbf{s}') \\ & + \sum_{k=1}^{K_f} \sum_{h=1}^{K_f} q_{f,i}^{k,h} \sum_{\mathbf{s}': n'_f > 0, n'_i = n_i, i \notin \mathbf{m}} \pi(\mathbf{s}') \end{aligned}$$

Consider now the special case $i = f$. The flux F (since it includes all possible transitions from queue f to any queue j) can be simplified as follows:

$$F \equiv \sum_{j=1, j\neq i}^{M} \sum_{n_f=0}^{F_f - 1} \sum_{k=1}^{K_f} \sum_{\mathbf{s}': n'_f = n_f + 1, n'_j = n_j} \sum_{h=1}^{K_f} q_{f,j}^{k,h} \pi(\mathbf{s}')$$

Similarly, the opposite flux G that describes all transitions that bring a job from queue j to queue f, $j \neq f$, is simplified as follows:

$$G \equiv \sum_{j=1, j\neq f}^{M} \sum_{k=1}^{K_j} \sum_{h=1}^{K_j} q_{j,f}^{k,h} \sum_{\mathbf{s}': n'_j > 0, n'_f < F_f} \pi(\mathbf{s}')$$

Note that the (9) and (10) would hold also if instantiated for $n_i = 0$ or $n_f = 0$ only, respectively. In fact, we can give a more detailed condition by recalling that if $n_i = 0$, then phase transitions in i are not possible, hence the balance $F = G$ 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 $S(k, n_i = 0) \equiv \{\mathbf{s} = (\mathbf{s_1}, \mathbf{s_2}, \ldots \mathbf{s_M}) | \mathbf{s_i} : n'_i = 0, k'_i = k\}$ and $\bar{S}(k, n_i = 1)$. A similar argument holds for $n_f = 0$. The proof continues by imposing the equilibrium balance $F = G$ and by rewriting the flux equations in terms of the QR marginal probability. For additional details, we refer the interested reader to [8, Thm. 3]. ■

The above equations show several differences compared to the marginal balances developed for MAP queueing networks with infinite capacity [5]. In addition to the obvious condition on the stations contributing to the throughput flow being active, i.e., $j \notin \boldsymbol{m}$, the last term in the right hand side of (9) describes the departures from station i following an unblocking event in station f that frees capacity which i has priority to use because of $Head(\mathbf{m}) = i$. Thus, this term captures the fundamental behavior of a departure from f that unblocks queue i. Interestingly, the departure flow from i is regulated in this case by the rate of departure of f, thus showing a case of non-product-form behavior where the throughput of a station depends on the rate of another station.

We now introduce a new class of balance conditions that describe the behavior of throughput while queue f is full. These balances are related to cuts of the Markov chain underlying the queueing network that separate the states shown in Figure 2 into partitions where $\boldsymbol{m}$ has different length b. Intuitively, as queue f enters into an extended period of time during which it remains full, the queues feeding f progressively block leading to changes in the composition of the list $\mathbf{m}$. The next theorem summarizes the balance between the rate of change of $\mathbf{m}$ due to queue blocking and the corresponding rate of unblocking events due to departures from f.

Theorem 5: The QR marginal probabilities for states where the finite capacity queue f is full satisfy

$$\sum_{\substack{j=1\\ j\neq f}}^{M}\sum_{n_j=1}^{F_j}\sum_{k=1}^{K_j}\sum_{h=1}^{K_j}\sum_{v=1}^{K_f}\sum_{\substack{\boldsymbol{m}:j\notin\boldsymbol{m}\\ \sum_i b_i\,=\,b}} q_{j,f}^{k,h}\pi(n_j,k,F_f,v,\boldsymbol{m})$$
$$=\sum_{\substack{j=1\\ j\neq f}}^{M}\sum_{n_j=0}^{F_j}\sum_{k=1}^{K_f}\sum_{h=1}^{K_f}\sum_{v=1}^{K_j}\times$$
$$\times\sum_{\boldsymbol{m}:\,\sum_i b_i\,=\,b+1} q_{f,j}^{k,h}\pi(n_j,v,F_f,k,\boldsymbol{m}),\quad(12)$$

for all number of blocked queues $b=0,\ldots,B-2$. When the number of blocked queues is $b=B-1$ we can write the stronger condition

$$\sum_{\substack{j=1\\ j\neq f}}^{M}\sum_{n_j=1}^{F_j}\sum_{k=1}^{K_j}\sum_{h=1}^{K_j}\sum_{v=1}^{K_f} q_{j,f}^{k,h}\pi(n_j,k,F_f,v,\boldsymbol{m}_j)=\sum_{\substack{j=1\\ j\neq f}}^{M}\sum_{\substack{w=1\\ w\neq i}}^{M}\sum_{k=1}^{K_f}\times$$
$$\times\sum_{v=1}^{K_f}\sum_{\substack{\boldsymbol{m}_j\\ \sum_i b_i=b+1}} q_{f,w}^{k,v}\pi(F_f,k,F_f,k,Add(\boldsymbol{m}_j,j)),\quad(13)$$

where $\boldsymbol{m}_j$ is any blocking list with $b=B-1$ blocked queues satisfying $j\notin\boldsymbol{m}_j$, and $Add(\boldsymbol{m}_j,j)$ is a list obtained by adding queue j at the tail of $\boldsymbol{m}_j$.

Proof: Consider a partitioning of the state space E_{BAS} into the following two subsets: H_b where there are up to $b\geq 0$ blocked queues and H_{b+1}, where the finite capacity queue is full and there are $b+1$ or more blocked queues on f. The left hand side of (12) represents the probability flux flowing through the state space cut associated to departures from station j to f that block j thus adding an entry at the end of $\boldsymbol{m}$. Conversely, the right hand side of (12) is the probability flux of departures from f such that at least a station $j\neq f$ gets unblocked thus reducing by one entry the list $\boldsymbol{m}$. Since no more than one queue gets blocked or unblocked at a time, it follows that the balance fully characterizes the probability flux balance across the cut that separates H_b from H_{b+1} which proves the equation.

Equation (13) considers the case where only a single queue j (in addition to f) is left unblocked, for any feasible choice of j. In this case, we know that only a departure event from j can increase the blocked queue list $\boldsymbol{m}_j$ to $Add(\boldsymbol{m}_j,j)$. Thus, we can apply the same argument used to prove (12) by focusing on the cut that separates in E_{BAS} the partition having blocking list $Add(\boldsymbol{m}_j,j)$ from the rest of the chain, which completes the proof. ■

III. MAP Queueing Networks with RS-RD blocking

We now characterize MAP queueing networks with RS-RD blocking. For RS-RD, the notation is simpler than the one of the BAS case, with essentially no changes from the basic MAP networks without blocking. The main difference introduced by RS-RD into MAP queueing networks is a reduction of the cardinality of the state space due to the removal of all states where $n_i>F_i$. Differently from BAS, there is no need for tracking the order of blocking by the list $\boldsymbol{m}$, since a job that cannot be delivered is simply re-executed without blocking the sender queue activity. Thus, the QR marginal probabilities are immediately expressed in the RS-RD case as $\pi(n_i,k_i,n_j,k_j)$, where the $\boldsymbol{m}$ list is no longer used being always $\boldsymbol{m}=\emptyset$. As stated earlier, we consider throughout this section the general case where several queues may have finite capacity, i.e., $F_i<N$ for any subset of indexes i. We denote by E_{RS-RD} the state space in the RS-RD case[1]. Since QR marginals are a restriction of those used in the BAS case, most performance indexes including queue length Q_i and utilization U_i are defined similarly to the BAS case, where we simply substitute the QR marginal probabilities for BAS with those used in RS-RD and summations on $\boldsymbol{m}$ consider only $\boldsymbol{m}=\emptyset$. A different definition is instead used for the effective utilization of queue i in phase k which is given by

$$E_i^{k_i}=\textstyle\sum_{n_i=1}^{F_i}\Big(\pi(n_i,k_i,n_i,k_i)$$
$$-\textstyle\sum_{j=1,j\neq i,p_{ij}>0}^{M}\sum_{k_j=1}^{K_j}p_{ij}\pi(n_i,k,F_j,k_j)\Big)\quad(14)$$

where the first term sums to the utilization of queue i in phase k_i, while the other summations represent the probability of observing the destination station j full. The basic characterization of the RS-RD state space holds similarly for the BAS case except for the formulas where the effective utilization is involved, i.e., Theorem 1 that is here extended to the RS-RD case. Due to limited space we report only proof outlines since the general ideas behind the RS-RD proofs are qualitatively similar to the ones used in [5] for the non-blocking case.

Theorem 6: The utilization levels of queue i in its K_i phases are in equilibrium, i.e., for each phase k, $1\leq k\leq K_i$,

$$\textstyle\sum_{j=1,j\neq i}^{M}\sum_{h=1}^{K_i}q_{i,j}^{k,h}E_i^k+\sum_{h=1,h\neq k}^{K_i}q_{i,i}^{k,h}U_i^k=$$
$$\textstyle\sum_{j=1,j\neq i}^{M}\sum_{h=1}^{K_i}q_{i,j}^{h,k}E_i^h+\sum_{h=1,h\neq k}^{K_i}q_{i,i}^{h,k}U_i^h\quad(15)$$

Proof: (Outline) The proof follows the same steps of the BAS case. However, (15) differs from the BAS case because in RS-RD a queue is never effectively blocked. As a consequence, for self-routed jobs the classical utilization should be taken into account. A complete derivation can be found in [8, Thm. 2′]. ■

The following theorem shows that a balance holds between the marginal probabilities similarly to the one developed in the BAS case. This theorem differs from the one for non-blocking MAP networks in [5] only in the fact that it involves a subset of the original state space.

Theorem 7: The arrival rate at queue i when its queue length is n_i jobs, $0<n_i\leq F_i-1$, is balanced by the rate of

[1] The interested reader can refer to [3] for a recursive expression to compute the state space cardinality for a queueing network where all queues have the same capacity and RS-RD blocking.

departures when the queue length is $n_i + 1$, i.e.,

$$\begin{aligned}&\textstyle\sum_{j=1,j\neq i}^{M}\sum_{n_j=1}^{F_j}\sum_{k=1}^{K_j}\sum_{h=1}^{K_j}\sum_{u=1}^{K_i} q_{j,i}^{k,h}\pi(n_i,u,n_j,k)\\ &\textstyle= \sum_{j=1,j\neq i}^{M}\sum_{n_j=0}^{F_j-1}\sum_{u=1}^{K_j}\sum_{k=1}^{K_i}\sum_{h=1}^{K_i} q_{i,j}^{k,h}\pi(n_i{+}1,k,n_j,u)\end{aligned} \tag{16}$$

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

$$\begin{aligned}&\textstyle\sum_{j=1,j\neq i}^{M}\sum_{n_j=1}^{F_j}\sum_{k=1}^{K_j}\sum_{h=1}^{K_j} q_{j,i}^{k,h}\pi(n_i=0,u,n_j,k)\\ &\textstyle= \sum_{j=1,j\neq i}^{M}\sum_{n_j=0}^{F_j-1}\sum_{h=1}^{K_j}\sum_{k=1}^{K_i} q_{i,j}^{k,u}\pi(n_i=1,k,n_j,h)\end{aligned} \tag{17}$$

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

Proof: A complete derivation can be found in [8, Thm. 3′]. ■

Also Theorem 2, Corollary 1 and Theorem 3 of the BAS case still hold for the RS-RD case by setting $\boldsymbol{m} = \emptyset$, the proof is qualitatively identical to the BAS case [8].

Finally, as for the original MAP queueing networks, the queue-length of i in all its phases satisfies the follows balance.

Theorem 8: The states of queue i in phase k and in phase h are related by the balance

$$\begin{aligned}&\textstyle\sum_{h=1,h\neq k}^{K_i}\sum_{j=1,j\neq i}^{M}\sum_{n_j=0}^{F_j-1}\sum_{u=1}^{K_j}\sum_{n_i=1}^{F_i} q_{i,j}^{k,h} n_i\pi(n_i,k,n_j,u)\\ &\qquad\textstyle+ \sum_{h=1,h\neq k}^{K_i}\sum_{n_i=1}^{F_i} q_{i,i}^{k,h} n_i\pi(n_i,k,n_i,k)\\ &\textstyle+ \sum_{j=1,j\neq i}^{M}\sum_{h=1}^{K_i}\sum_{n_j=0}^{F_j-1}\sum_{u=1}^{K_j}\sum_{n_i=1}^{F_i} q_{i,j}^{h,k}\pi(n_i,h,n_j,u)\\ &\textstyle= \sum_{j=1,j\neq i}^{M}\sum_{h=1}^{K_j}\sum_{u=1}^{K_j}\sum_{n_j=1}^{F_j} q_{j,i}^{h,u}\sum_{n_i=0}^{F_i-1}\pi(n_i,k,n_j,h)\\ &\qquad\textstyle+ \sum_{h=1,h\neq k}^{K_i}\sum_{n_i=1}^{F_i} q_{i,i}^{h,k} n_i\pi(n_i,h,n_i,h)\\ &\textstyle+\sum_{h=1,h\neq k}^{K_i}\sum_{j=1,j\neq i}^{M}\sum_{n_j=0}^{F_j-1}\sum_{u=1}^{K_j}\sum_{n_i=1}^{F_i} q_{i,j}^{h,k} n_i\pi(n_i,h,n_j,u)\end{aligned}$$

Proof: A complete derivation is given in [8, Thm. 5′]. ■

IV. Boundable Approximations

The fundamental idea behind the proposed approximations and bounds is to use the exact characterization developed in Sections II and III to formulate an educated guess of the values of the QR marginal probabilities. We here describe our methodology for BAS networks, the application to RS-RD blocking follows easily by considering $\boldsymbol{m} = \emptyset$.

To determine an approximate marginal distribution for the model, we assume the values $\pi(n_i, k_i, n_j, k_j, \boldsymbol{m})$ as unknowns in an optimization program $\mathcal{O}$. This optimization program takes the form

$$\begin{aligned}\mathcal{O}: \qquad &\min f_{obj}(\boldsymbol{\pi}^G) \qquad \text{s.t.}\\ &\mathbf{A}\boldsymbol{\pi}^G \leq \boldsymbol{b}\\ &\mathbf{C}\boldsymbol{\pi}^G \leq \boldsymbol{d}\end{aligned}$$

where $\boldsymbol{\pi}^G$ is the vector of the current guesses for all the QR marginal probabilities $\pi(n_i, k_i, n_j, k_j, \boldsymbol{m})$, f_{obj} is a (possibly nonlinear) objective function to be optimized, and the constraints are of two types. A first group of constraints, $\mathbf{A}\boldsymbol{\pi}^G \leq \boldsymbol{b}$, is the set of all equations and inequalities developed in the BAS (or RS-RD) characterizations, including the specialized marginal balances for $n_i = 0$. Notice that such equations are all linear constraints, mainly equalities. A second group of linear constraints, $\mathbf{C}\boldsymbol{\pi}^G \leq \boldsymbol{d}$, imposes obvious conditions that describe in the optimization program the feasible values of the terms $\pi^G(n_i, k_i, n_j, k_j, \boldsymbol{m}) \in \boldsymbol{\pi}^G$ in order to specify a valid QR marginal distribution. These constraints impose, for instance, that the unknowns of the linear program are probabilities, hence numbers ranging in $[0, 1]$, or that a queue can only be in a single state at a time hence, e.g., $\pi^G(n_i, k_i, n_i{+}c, k_i, \boldsymbol{m}) = 0$, $\forall c \neq 0$. A summary of these basic conditions is given in Table II.

Let $\boldsymbol{\pi}^G_{opt}$ be the guess $\boldsymbol{\pi}^G$ which provides the optimal value for the objective function f_{obj}. The crucial property of the optimization program $\mathcal{O}$ is that its constraints are satisfied by the exact QR marginal distribution $\boldsymbol{\pi}$. It then follows that the exact solution $f_{obj}(\boldsymbol{\pi})$ is always a *feasible* solution for the optimization program $\mathcal{O}$, although it may not be necessarily the optimal one $f_{obj}(\boldsymbol{\pi}^G_{opt})$. This property leads to the following approximation and bounding techniques.

A. Performance Metric Bounds

First, suppose that f_{obj} defines a performance metric of interest, such as the utilization of station i

$$f_{obj}(\boldsymbol{\pi}^G) = \sum_{k_i=1}^{K_i}\sum_{n_i\geq 1} \pi^G(n_i, k_i, n_i, k_i, \boldsymbol{m})$$

or its average queue-length

$$f_{obj}(\boldsymbol{\pi}^G) = \sum_{k_i=1}^{K_i}\sum_{n_i\geq 1} n_i\pi^G(n_i, k_i, n_i, k_i, \boldsymbol{m})$$

Then, by construction, minimizing $\mathcal{O}$ returns a lower bound $f_{obj}(\boldsymbol{\pi}^G_{opt}) = \min f_{obj}(\boldsymbol{\pi}^G) \leq f_{obj}(\boldsymbol{\pi})$, since $\boldsymbol{\pi}^G = \boldsymbol{\pi}$ is a feasible solution of the optimization program. Similarly, solving $\mathcal{O}$ as a maximization problem returns an upper bound $f_{obj}(\boldsymbol{\pi}^G_{opt}) = \max f_{obj}(\boldsymbol{\pi}^G) \geq f_{obj}(\boldsymbol{\pi})$. Noting that utilizations and queue-lengths are linear functions of $\boldsymbol{\pi}^G$, it then follows that $\mathcal{O}$ can be solved efficiently as a linear optimization program. Such a solution provides upper and lower bounds on the performance metrics of a MAP queueing network[2]. Notice that other metrics, such as the effective utilization or the throughput, may be defined similarly to the utilization and queue-length in terms of a linear objective function. Conversely, response times need to be estimated using Little's law as ratios of average queue-length and average throughput. Hence, they can be solved as nonlinear global optimization programs or, more easily, estimated indirectly from the bounds on queue-length and throughput. This approach to bounding the performance of a MAP queueing network has been also investigated for models without blocking in [5] and we refer to it as *QR bounds*.

[2] We stress again that such values are guaranteed to be bounds by construction if the optimizer returns a *global* optimum for $\mathcal{O}$, as it is always the case for linear programs used for bounds computation.

B. Approximate Model Solution

The second main application of the optimization program $\mathcal{O}$ is in approximating the QR marginal probabilities. We define objective functions that allow one to obtain accurate approximations, noticeably also on cases where the QR bounds are not tight. We here introduce two approximation techniques: a *maximum entropy method* (MEM) for MAP queueing networks and a new principle of *minimal mutual information* (MMI). It is important to remark that, since the QR bounds can always be generated regardless of these approximations, the gap between upper and lower bounds provides an independent assessment on the maximum inaccuracy in using MEM or MMI in place of an exact solution. Thus, such approximations are always bounded, meaning that the maximum error of MEM or MMI is the maximum distance from a point lying in between the upper and lower QR bounds. Furthermore, the objective functions are non-linear, hence one should consider a local optimum obtained by a nonlinear solver[3].

MEM searches for a set of QR marginal probabilities that maximizes the information content of the distribution as defined by the entropy function H. To simplify notation, for the rest of this section let

$$\pi^G(n_i, n_j) \equiv \pi^G(n_i, k_i, n_j, k_j, \boldsymbol{m}).$$

MEM optimizes in $\mathcal{O}$ the objective function

$$\max H = \max \left(- \sum_{n_i,n_j,k_i,k_j,\boldsymbol{m}} \pi^G(n_i, n_j) \log_2 \pi^G(n_i, n_j) \right)$$

The values of performance indexes such as utilizations and queue-lengths are then obtained directly from the QR marginal distribution that maximizes H. The rationale behind a maximum entropy solution is that it is known to be exact in a number of queueing models, noticeably in exponential single-class closed queueing networks [10]. Notice that a well-known maximum entropy method for queueing networks has already been developed in [11] based on the analysis of the $GI/GI/1$ queue. However, the MEM technique we propose differs substantially from the one in [11]. First, the method is able for the first time to consider the state of *all* the queues in the network simultaneously, instead of recursively evaluating queues one at a time as in [11]. Importantly, our technique is also able to consider the impact of autocorrelation in job flows introduced by MAPs, which is ignored in the analysis of the $GI/GI/1$ queue. Indeed, this is a critical aspect of a MAP queueing network that cannot be ignored, being responsible of dynamic bottleneck switch effects, even at equilibrium, that significantly affect the model solution [4]. Finally, and perhaps most importantly, our MEM solution is subject to satisfying the very large set of constraints developed in Sections II and III, whereas the one in [11] considers a small subset of such constraints. Therefore, our approximation is more heavily constrained to be representative of the model under study. Stemming from this last point, we remark that the main limitation of the proposed MEM compared to the one in [11] is that our method requires numerical optimization, whereas [11] is based on simple closed-form formulas.

In addition to the MEM method, we introduce the MMI criterion as a new technique for approximating an unknown probability distribution of a queueing network. For a QR marginal probability distribution, MMI considers the following objective function

$$\min \left(\sum_{n_i,n_j,k_i,k_j,\boldsymbol{m}} \pi^G(n_i, n_j) \log_2 \frac{\pi^G(n_i, n_j)}{\pi^G(n_i, n_i)\pi^G(n_j, n_j)} \right)$$

Following standard information theory, the argument of the minimization is the mutual information of $\pi^G(n_i, n_j)$, which quantifies how much the knowledge on the state of queue i reduces our uncertainty about the state of station j. However, by noting that for a product-form model the knowledge of the state of a queue provides little information on the state of the other stations (for a closed model it only provides an upper bound $n_j \leq N - n_i$ that becomes progressively looser as N and M increase), we conclude that the MMI solution may be interpreted as a product-form-type approximation for a MAP queueing network. That is, when the mutual information is minimal, the corresponding marginal probability distribution finds the description in which queues i and j are maximally independent. Clearly, in networks with blocking the mutual information is not in general minimal, since blocking yields a strong dependence between the behavior of two (or even more than two) queues. However, the fundamental idea of our proposed method is that the blocking is already strongly characterized by our QR marginal balances, hence MMI deals only with allocating the portion of the probability mass that remains unconstrained. We illustrate this concept below in a "toy" example. Notice also that the MMI approach is expected to be accurate especially in heavy load, where closed networks progressively approach the behavior of open models due to the formation of bottleneck stations whose service process, being continuously busy, acts as an "arrival process" for the rest of the network. Open networks are typically less inter-dependent than their closed counterparts.

C. Toy Example

To better understand the properties of MEM and MMI, consider the following illustrating example. The model is composed by three queues with exponential service rate $\mu_1 = \mu_2 = 1$, $\mu_3 = 2$. The routing matrix is

$$P = \begin{bmatrix} 0 & 0.50 & 0.50 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \tag{18}$$

which is a special case for the topology shown in Figure 1. Buffer capacities are $F_1 = 1$, $F_2 = F_3 = N$, with $N = 3$ being the job population; the blocking mechanism is RS-RD. Despite its apparent simplicity, for this model the QR bounds

[3] We stress that since here the focus is on approximation, rather than bounds, one does not need to ensure global optimality of the final result in order to have a usable solution. As usual, the gap between primal and dual formulations of the optimization program can be used as a measure of the relative quality of $f_{obj}(\boldsymbol{\pi}^G)$ with respect to its global optimum.

provide the following estimates of upper bounds (U_k^{max}) and lower bounds (U_k^{min}) on the exact utilizations (U_k) of queue k:

	queue 1	queue 2	queue 3
U_k^{max}	0.5000	0.7500	0.9524
U_k	0.4828	0.4483	0.8966
U_k^{min}	0.4762	0.3333	0.7500

In this example, the utilization of queue 2 is loosely captured by the QR bounds that leave a gap of about 42% between the upper and lower limits. That is, the solver is allowed to allocate the probability mass in ways that vary significantly with respect to the performance of queue 2, in other words, queue 2 is not sufficiently constrained by the characterization in Section III. A closer investigation reveals inconsistencies on the solution with respect to the exact probabilities, e.g., for the upper bound

$$\begin{aligned}&\pi_{opt}^G(n_1 = 1, k_1 = 1, n_2 = 1, k_2 = 1) = 0.5000,\\ &\qquad\pi_{opt}^G(n_1 = 1, k_1 = 1, n_3 = 1, k_3 = 1) = 0.1905 \qquad (19)\end{aligned}$$

while for the lower bound

$$\begin{aligned}&\pi_{opt}^G(n_1 = 1, k_1 = 1, n_2 = 1, k_2 = 1) = 0.0,\\ &\qquad\pi_{opt}^G(n_1 = 1, k_1 = 1, n_3 = 1, k_3 = 1) = 0.0 \qquad (20)\end{aligned}$$

which are both impossible since the two marginal probabilities describe the same state $(n_1 = 1, n_2 = 1, n_3 = 1, k_1 = 1, k_2 = 1, k_3 = 1)$ in the original queueing network. In fact, in the original model

$$\begin{aligned}&\pi(n_1 = 1, k_1 = 1, n_2 = 1, k_2 = 1)\\ &\quad= \pi(n_1 = 1, k_1 = 1, n_3 = 1, k_3 = 1) = 0.1379, \qquad (21)\end{aligned}$$

We have verified that such an unconstrained mass can be allocated exactly by adding to $\mathcal{O}$ the following consistency constraint

$$\pi(n_j, k_j, n_i, k_i) = \pi(n_j, k_j, n_t = N - n_j - n_i, k_t),$$

for all choices of the stations $i \neq j \neq t$ and their states. This provides the optimal solution $\boldsymbol{\pi}_{opt}^G = \boldsymbol{\pi}$. This imposes that, in a model with $M = 3$ queues, there are at most two degrees of freedom in assigning the populations n_i and n_j at the queues, since the population at the last queue will be automatically set to $n_t = N - n_i - n_j$. This constraint is obvious but its integration in the QR marginal characterization requires in general a cubic number equations for a model with $M = 3$ which is not consistent with the approach that we have pursued; furthermore, for a model with $M \geq 4$ these constraints cannot be imposed using the QR marginal probabilities, since one would need to express the state of $M - 1$ queues simultaneously. This example highlights some consequences of the structural limitation of QR marginal probabilities; this limitation is that they cannot represent correctly the allocations of jobs (or the active phases) on more than two queues simultaneously.

We have then obtained the MEM and MMI solutions for the above model and found them as follows

	queue 1	queue 2	queue 3
U_k^{mem}	0.4887	0.5515	0.8464
U_k^{mmi}	0.4818	0.4316	0.9046
U_k	0.4828	0.4483	0.8966

which are much closer to the exact distribution that the QR bound solution. Further, we have now

$$\begin{aligned}&\pi_{mem}^G(n_1 = 1, k_1 = 1, n_2 = 1, k_2 = 1) = 0.2618,\\ &\qquad\pi_{mem}^G(n_1 = 1, k_1 = 1, n_3 = 1, k_3 = 1) = 0.0907 \qquad (22)\end{aligned}$$

and

$$\begin{aligned}&\pi_{mmi}^G(n_1 = 1, k_1 = 1, n_2 = 1, k_2 = 1) = 0.1180,\\ &\qquad\pi_{mmi}^G(n_1 = 1, k_1 = 1, n_3 = 1, k_3 = 1) = 0.1455 \qquad (23)\end{aligned}$$

which provide a substantial consistency improvement compared to the QR bounds, especially for the novel MMI method.

V. Numerical Validation

We illustrate the accuracy of the BAS and RS-RD bounds on a set of case studies having different level of complexities, number of queues, and network topology. Throughout the experiments, we use a combination of exponential service processes and nonrenewal autocorrelated MAPs. We use the GLPK linear programming solver to compute bounds and the MINOS solver for nonlinear programs required to evaluate the MEM and MMI approximations. For simplicity of comparison, we always use a short-range dependent MAP process with two-phases having representation [15]

$$\begin{aligned}\mathbf{D}_0 &= \begin{bmatrix} -1.016212022108574 & 0 \\ 0 & -0.015702871508448 \end{bmatrix}\\ \mathbf{D}_1 &= \begin{bmatrix} 1.016186165025678 & 0.000025857082896 \\ 0.001569887597955 & 0.014132983910493 \end{bmatrix}\end{aligned} \qquad (24)$$

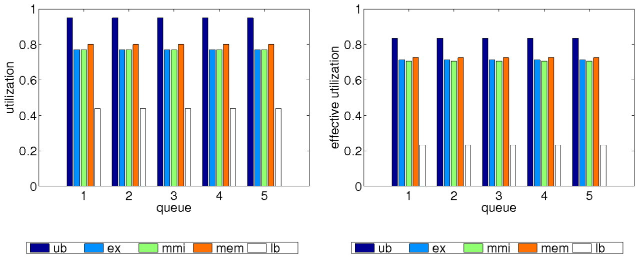
This yields a process with moments $E[X] = 1$, $E[X^2] = 4$, $E[X^3] = 400$, and positive autocorrelation function $\rho_k = \frac{1}{3}(\frac{9}{10})^k$ such that $\rho_1 = 0.300$, $\rho_2 = 0.270$, $\rho_3 = 0.243, \ldots$. On a laptop computer, the hardest case study execution times were less than 5 seconds for the QR bounds, about 300 seconds for the nonlinear programs used for MIM/MEM. Note that we used a single CPU core, nonlinear solvers running on multi-core machines are usually 8-10 times faster, thus the nonlinear solution can be significantly accelerated.

A. Case Study 1

Let us first consider a model composed of $M = 5$ queues with $N = 10$ jobs, capacity $F_i = 5$ for each queue $i = 1, \ldots, M$, and service processes all equal to the short-range dependent MAP given in (24). Hence, all stations can

TABLE II
CONSTRAINTS IMPOSING BASIC PROPERTIES OF $\pi(n_i, k_i, n_j, k_j, \boldsymbol{m})$ INCLUDED IN THE OPTIMIZATION PROGRAMS

$\pi(n_i, k_i, n_j, k_j, \mathbf{m}) \geq 0,$	$\forall_{i=1}^{M} \forall_{n_i=0}^{F_i} \forall_{k_i=1}^{K_i} \forall_{j=1}^{M} \forall_{n_j=0}^{F_i} \forall_{k_j=1}^{K_j} \forall_{\boldsymbol{m}}$
$\sum_{n_j=0}^{F_j} \sum_{k_j=1}^{K_j} \sum_{\boldsymbol{m}} \pi(n_j, k_j, n_j, k_j, \mathbf{m}) = 1,$	$\forall_{j=1}^{M}$
$\pi(n_j, k, n_j, h, \boldsymbol{m}) = 0,$	$\forall_{j=1}^{M} \forall_{k=1}^{K_j} \forall_{h=1, h \neq k}^{K_j} \forall_{n_j=0}^{F_j} \forall_{\boldsymbol{m}}$
$\pi(n_j, k, n_j', h, \boldsymbol{m}) = 0,$	$\forall_{j=1}^{M} \forall_{k=1}^{K_j} \forall_{n_j=0}^{F_j} \forall_{h=1}^{K_j} \forall_{n_j'=0, n_j' \neq n_j}^{F_j} \forall_{\boldsymbol{m}}$
$\pi(n_j, k, n_i, h, \boldsymbol{m}) = 0$	$\forall_{j=1}^{M} \forall_{k=1}^{K_j} \forall_{n_j=0}^{F_j} \forall_{i=1, i \neq j}^{M} \forall_{h=1}^{K_i} \forall_{n_i=N-n_j+1}^{F_i} \forall_{\boldsymbol{m}}$
$\sum_{n_f=0}^{F_f-1} \sum_{h=1}^{K_f} \pi(n_j, k, n_f, h, \boldsymbol{m}) = 0,$	$\forall_{j=1, f \neq j}^{M} \forall_{k=1}^{K_j} \forall_{n_j=0}^{F_j} \forall_{\boldsymbol{m}: \boldsymbol{m} \notin \emptyset}$
$\pi(n_j = 0, k, n_i, h, \boldsymbol{m}) = 0,$	$\forall_{j=1}^{M} \forall_{k=1}^{K_j} \forall_{i=1}^{M} \forall_{h=1}^{K_i} \forall_{n_i=0}^{F_i} \forall_{\boldsymbol{m}: j \in \boldsymbol{m}}$
$\pi(n_j, k, n_i, h, \boldsymbol{m}) = 0,$	$\forall_{j=1}^{M} \forall_{k=1}^{K_j} \forall_{n_j=F_j+1}^{N} \forall_{i=1}^{M} \forall_{h=1}^{K_i} \forall_{n_i=0}^{F_i} \forall_{\boldsymbol{m}}$
$\pi(n_j, k, n_i, h, \boldsymbol{m}) = 0$	$\forall_{j=1}^{M} \forall_{k=1}^{K_j} \forall_{n_j=1}^{F_j} \sum_{\substack{i=1 \\ i \neq j \neq f}}^{M} \forall_{h=1}^{K_i} \forall_{n_i=N-nj-F_f+1}^{F_i} \forall_{\boldsymbol{m}: Head(\boldsymbol{m})=j}$
$\pi(n_j, k, n_i, h, \boldsymbol{m}) = \pi(n_i, h, n_j, k, \boldsymbol{m}),$	$\forall_{j=1}^{M} \forall_{n_j=0}^{F_j} \forall_{k=1}^{K_j} \forall_{i=1}^{M} \forall_{n_i=0}^{F_i} \forall_{h=1}^{K_i} \forall_{\boldsymbol{m}}$
$\pi(n_j, k, n_j, k, \boldsymbol{m}) = \sum_{n_i=0}^{N-n_j} \sum_{h=1}^{K_i} \pi(n_j, k, n_i, h, \boldsymbol{m}),$	$\forall_{j=1, j \neq i}^{M} \forall_{k=1}^{K_j} \forall_{n_j=0}^{F_j} \forall_{i=1}^{M} \forall_{\boldsymbol{m}}$
$\pi(n_j, k, n_j, k, \boldsymbol{m}) = 0,$	$\forall_{j=1}^{M} \forall_{k=1}^{K_j} \forall_{n_j=0}^{F_j} : N - n_j > \sum_{\substack{y=1 \\ y \neq j}}^{M} F_y$
$\pi(n_j, k, n_i, h, \boldsymbol{m}) = 0,$	$\forall_{i}^{M} \forall_{j=1, j \neq i}^{M} \forall_{k=1}^{K_j} \forall_{n_j=0}^{F_j} \forall_{h=1}^{K_i} \forall_{n_i=0}^{F_i} : N - n_j - n_i > \sum_{\substack{y=1 \\ y \neq j \neq i}}^{M} F_y$



(a) Utilization

(b) Effective Utilization

Fig. 3. **Case 1 - MAP network with RS-RD blocking**

be blocked. The routing matrix is

$$P = \begin{bmatrix} 0 & 0.5000 & 0 & 0 & 0.5000 \\ 0.5000 & 0 & 0.5000 & 0 & 0 \\ 0 & 0.5000 & 0 & 0.5000 & 0 \\ 0 & 0 & 0.5000 & 0 & 0.5000 \\ 0.5000 & 0 & 0 & 0.5000 & 0 \end{bmatrix}$$

This is a case where we compare approximations and bounds under multiple RS-RD blocking. We see in Figure 3 that the upper and lower bounds (“ub” and “lb”, respectively) are not able to generate a tight envelope around the exact utilization and exact effective utilizations (“ex”). However, both MEM and MMI return almost perfect results within less than 2% utilization. Similarly to the toy example, MMI appears slightly more effective than MEM for capturing the probability distribution. Notice also that the MEM solution is slightly affected by numerical perturbations due to the fully symmetric routing of this network.

B. Case Study 2

This model differs from Case Study 1 in that we consider BAS blocking and the topology is now full mesh with routing matrix

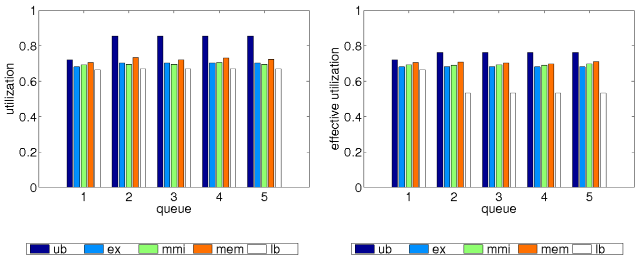
$$P = \begin{bmatrix} 0 & 0.2500 & 0.2500 & 0.2500 & 0.2500 \\ 0.2500 & 0 & 0.2500 & 0.2500 & 0.2500 \\ 0.2500 & 0.2500 & 0 & 0.2500 & 0.2500 \\ 0.2500 & 0.2500 & 0.2500 & 0 & 0.2500 \\ 0.2500 & 0.2500 & 0.2500 & 0.2500 & 0 \end{bmatrix}$$

Furthermore, station capacities are now $F_1 = 5$, $F_2 = F_3 = F_4 = F_5 = N$ so that only station 1 has finite capacity. The population is $N = 10$ jobs. Service processes are again identical short-range dependent MAPs. The results in Figure 4 indicate that the bounds are very effective in capturing the performance of the finite capacity queue 1, while more uncertainty is left on queues $2-5$ where the gap between upper and lower bounds is approximately up to 20%. In spite of such uncertainty, MEM and MMI again find very accurate results, again within a few percent of the exact results, with MMI again being slightly better than MEM. This is a relevant result, since despite its apparent simplicity, the number of possible combinations of $\mathbf{m}$ vectors is 64 for each state in which queue 1 is full, which is significant. Hence, this experiment suggests that the MEM and MMI approximation are effective also on cases where the portion of the state-space due to the BAS precedence constraints is non-negligible.

C. Case Study 3

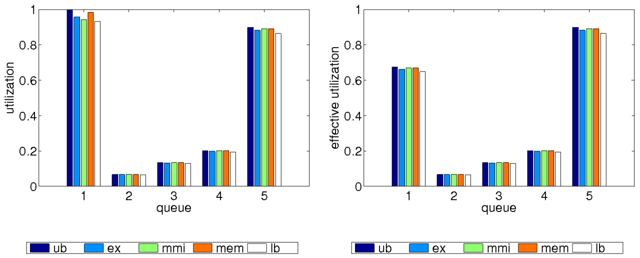
We now consider a classic central-server-type topology, where queue 1 feeds parallel stations. We assume $M = 5$, $N = 10$, and routing matrix

$$P = \begin{bmatrix} 0 & 0.1000 & 0.2000 & 0.3000 & 0.4000 \\ 1.0000 & 0 & 0 & 0 & 0 \\ 1.0000 & 0 & 0 & 0 & 0 \\ 1.0000 & 0 & 0 & 0 & 0 \\ 1.0000 & 0 & 0 & 0 & 0 \end{bmatrix}$$



(a) Utilization (b) Effective Utilization

Fig. 4. **Case 2 - A model with BAS blocking and mesh topology**



(a) Utilization (b) Effective Utilization

Fig. 5. **Case 3 - A MAP network with central-server-type topology**

Station 1 has MAP service (24), all other stations are exponential. All mean service times are equal to 1, expect for queue 5 where it is 3.333. Station 5 is also the only finite capacity queue with capacity $F_5 = 3$. Figure 5 reports experimental results. We see again that the proposed approximations are very effective, however this illustrates a case where also the bounds are very tight, and one may for instance take their middle point as a first approximation of the exact value of the utilizations. Quite interestingly, this shows a case where station 1 has a dramatic difference between utilization and effective utilization, due to the blocking on queue 5. This is perfectly captured by the proposed techniques.

VI. Conclusion

In this paper, we have proposed a major extension of MAP queueing network models to support BAS and RS-RD blocking mechanisms. Based on the recently proposed Quadratic Reduction (QR) technique [5], we have described the state space with a set of marginal probabilities having cardinality that is much smaller than for the original state space. Then, we have derived new exact characterization results that describe the relations between such marginal probabilities in the context of BAS and RS-RD blocking. Using a numerical optimization approach, we have derived boundable approximations on performance metrics based on the maximum entropy and minimum mutual information principles. Experimental results indicate that such approximations are highly accurate. Possible extensions of this work include considering additional blocking mechanisms and the generalization of the proposed techniques to queueing networks that limit the maximum number of jobs within a subnetwork, which are important to model admission control mechanisms.

Acnowledgement

The work of Giuliano Casale is supported by the Imperial College Junior Research Fellowship. This research is partially supported by NSF grants CNS-0720699, CCF-0811417 and CCF-0937925. The authors thank Valentina Lenti for extensive experimentation in her MSc thesis that helped in understanding better the characteristics of MAP queueing networks with blocking.

References

[1] I. Awan, A. Yar, M.E. Woodward. Analysis of Queueing Networks with Blocking under Active Queue Management Scheme. In *Proc. of IPDPS*, 61-68, 2006.

[2] S. Balsamo, V. De Nitto Personé, P. Inverardi. A review on Queueing Network Models with finite capacity queues for Software Architectures performance prediction. *Perform. Eval.*, 51(2-4), 269-288, 2003.

[3] S. Balsamo, V. De Nitto Personé, R. Onvural. Analysis of Queueing Networks with Blocking. Kluwer Academic, 2001.

[4] G. Casale, E. Smirni. MAP-AMVA: Approximate Mean Value Analysis of Bursty Systems . In Proc. of *IEEE/IFIP DSN*, 409–418, IEEE Press, Jul 2009.

[5] G. Casale, N. Mi, E. Smirni. Model-Driven System Capacity Planning Under Workload Burstiness. *IEEE Trans. on Computers*, 59(1):66-80, Jan 2010.

[6] G. Casale, E.Z. Zhang, E. Smirni. Trace Data Characterization and Fitting for Markov Modeling. *Elsevier Performance Evaluation*, vol. 67, 61-79, February 2010.

[7] H. Daduna, M. Holst. Customer Oriented Performance Measures for Packet Transmission in a Ring Network with Blocking. In Proc. of *14th GI/ITG Conf. On Measurement, Modeling and Evaluation of Computer and Comm. Systems*, 2008.

[8] V. De Nitto, G. Casale. E. Smirni. Analysis of Blocking Networks with Temporal Dependence. Tech. Rep. RR-10.83, Dept. Comp. Science, University of Rome Tor Vergata, March 2010.

[9] D. De Almeida, P. Kellert. Markovian and analytical models for multiple bus multiprocessor systems with memory blockings. Journal of Systems Architecture, 46, 455-477, 2000.

[10] A.E. Ferdinand. A Statistical Mechanical Approach to Systems Analysis. *IBM J. Res. Dev.*, 14(5):539-547, Sep 1970.

[11] D. D. Kouvatsos. Maximum Entropy Analysis of Queueing Network Models. Performance/SIGMETRICS Tutorials, Springer LNCS Vol. 729, 245-290, 1993.

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

[13] N. Mi, G. Casale, L. Cherkasova, E. Smirni Burstiness in Multi-Tier Applications: Symptoms, Causes, and New Models. In Proc. of *Middleware 2008*, LNCS 5346, 265-286, Springer, Dec 2008.

[14] K. Nakade. New bounds for expected cycle times in tandem queues with blocking, *Europ. J. Oper. Res.*, 125(1):84-92, 2000.

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

[16] R.O. Onvural. Survey of Closed Queueing Networks with Blocking. *ACM Computing Surveys*, 22:(2) 83-121, 1990.

[17] R.O. Onvural. Special Issue on Queueing Networks with Finite Capacity. *Perform. Eval.*, 17 (3), 1993.

[18] H.G. Perros. Queueing networks with blocking. Oxford University Press, 1994.

[19] X. Zhang, A. Riska, E. Riedel. Characterization of the E-commerce Storage Subsystem Workload. In *Proc. of QEST*, 297-306, 2008.

[20] T. Yamadaa, N. Mizuharab, H. Yamamotoc, M. Matsuib. A performance evaluation of disassembly systems with reverse blocking. Computers & Industrial Engineering Intelligent Manufacturing and Logistics, 56:(3), 1113-1125, 2009.