Analysis of Blocking Networks with Temporal Dependence
Vittoria De Nitto Personè\textsuperscript{a}, Giuliano Casale\textsuperscript{b}, Evgenia Smirni\textsuperscript{c}

\textsuperscript{a}University of Rome Tor Vergata, Italy
\textsuperscript{b}Imperial College London, U.K.
\textsuperscript{c}College of William and Mary, Virginia, U.S.A.

Abstract—In this paper we extend the class of MAP queueing networks to include blocking models. We consider two different blocking mechanisms: Repetitive Service-Random Destination and Blocking After Service. We analyze the Markov process underlying the MAP queueing network and propose a methodology based on a partition of the state space into “marginal state spaces”. By using this partition, we prove a set of “partial” balance equations that relates blocking performance indexes. The proposed methodology can be a sound framework to define approximate solution methods for MAP queueing networks with blocking.

1. Introduction

Blocking queueing networks are powerful tools to deal with more representative models of real-life systems. Their solutions are difficult and, despite of their importance, there is a general lack of methodologies aimed to include blocking phenomenon into solution tools.

MAP queueing networks are shown to be the first analytical methodology to describe and predict accurately the performance of complex systems operating under temporal dependencies [CaMiSmi08, CaMiSmi10]. MAP queueing networks include servers whose service times are Markovian Arrival Processes (MAPs), a class of Markov-modulated point processes that can model general distributions as well as the main features of nonrenewal workloads, such as autocorrelation in service times and burstiness [Neu89].

In this paper, we extend the class of MAP queueing network to include blocking mechanism. We consider two different blocking models: the Repetitive Service-Random Destination (RS-RD) mechanism and the Blocking After Service mechanism (BAS). These two different models are extensively studied in the literature and are representative of telecommunication systems and computer systems with limited shared resources.

For these two mechanisms we analyze the Markov process underlying the MAP queueing network. Because of the state space explosion, the queueing network’s equilibrium behavior cannot be determined exactly, but we argue that it can still be bounded accurately by describing the system with “reduced” state spaces, which we call marginal state spaces. Marginal state spaces capture the behavior of the network conditioned on a given queue being active, blocked or idle.

In the following section, we present the considered blocking models and the related literature. In the section 3, we define the MAP blocking queueing networks and we prove the marginal balance for BAS blocking. In the section 4, we extend the methodology and the results to the RS-RD blocking.
Finally, the section 5 concludes the paper.

2. Blocking models
Queueing network models with finite capacity queues and blocking have been introduced and applied as more realistic models of systems with finite capacity resources and population constraints. In these models, when a queue reaches its maximum capacity, the flow of customers entering the service node is stopped, both from other service nodes and from external sources in case of open networks. This phenomenon is referred to as blocking. In particular, each blocking mechanism defines the service node blocking time, the behaviours of arriving customers to full capacity service center and the servers' activity in the network.

Queueing networks with blocking can be applied to telecommunication systems and computer systems with limited shared resources, such as interconnecting links or store and forward buffers, as well as in production systems with the finite storage buffers. Different blocking mechanisms representing distinct behaviours of real systems with limited resources have been defined and analyzed in the literature. The interested reader can refer to [BalDenOnv01, Onv90, Onv93, Per84, Per89, Per94] for an extensive bibliography.

More recent results can be found in several research application areas such as computer systems [DeKe00], communication systems and networks [AwYaWo06, DaHo08, LiToLe07], manufacturing systems [YaMiYaMa09], software architectures [BalDenIn03] and also in the emerging area of “health systems” [KoKuSm05].

Since the first results on product forms [Aky87, Aky88, Aky89, BalCl98, BalDen91, BouVan97, Cl98, Onv89, Ser99], and equivalence and monotonicity properties [AdVdw89, AmmGer89, BalDenIa87, BalDon89, DaLiTo94, DaTo91, Den94, OnPe89, ShaYa89, VanTi86], few effort has been devoted to develop general methodologies to deal with blocking queueing networks. More recently, some new solution approaches have been proposed [Be&alii07, OsBi09].

Bounding analysis is a one of the most attractive methodologies to estimate performance measures with a limited cost. Tandem networks with general service time distribution and Blocking Before Service mechanism or BAS blocking were considered in [Nak00]. The author defines bounds for the expected cycle time. The presented results are good for very limited size of the finite buffers. Open general topologies networks with multiserver exponential queues and RS-RD blocking were considered in [KuSriKu98]. The model includes multiclass population. The authors defined bounds to the queues throughput and the blocking probabilities.

---

1 Throughout this paper we use the terms queue, node and service center interchangeably.
Our methodology combines the blocking characteristic with the temporal dependencies characteristic in the service process.

As introduced above, we consider two different blocking models that are extensively used in the literature.

Let us consider a closed queueing network, with routing matrix $P = [p_{ij}]$, that is when $p_{ij} > 0$ queue $i$ is connected to queue $j$. We consider finite capacity queues, with $F_i$ the queue $i$ capacity. Note that $F_i$ includes the jobs in service. If $n_i$ denotes the instantaneous population at queue $i$, when $n_i = F_i$ queue $i$ is full and any incoming job cannot be accommodated until a departure from queue $i$ takes place. Each blocking model defines the behaviour of the “sending” queue and the behaviours of arriving customers to full capacity queue. The considered blocking mechanisms are defined in the following.

Repetitive Service-Random Destination (RS-RD)

A queue $i$, if not empty, processes a job regardless of the job population at its destination $j$ ($p_{ij} > 0$). When node $i$ completes, if node $j$ is full, the completed job is rerouted at node $i$ where, according to node $i$ scheduling, it will receive a new service. During the new service, the job will select a new destination independently from the previous one.

Note that according to RS-RD blocking a node will never be actually blocked, but it “wastes” its service, it could have to repeat it. In this case, we define blocking as the time the node is working for a full destination node.

RS-RD blocking is used to model mainly telecommunications system. Recently, this blocking mechanism was used to model congestion control in the internet [AwYaWo06].

Blocking After Service mechanism (BAS)

A queue $i$, if not empty, processes a job regardless of the job population at its destination $j$ ($p_{ij} > 0$). When node $i$ completes, if node $j$ is full, node $i$ stops its activity (it is blocked) and the completed job waits until a departure will occur from node $j$. At that moment two simultaneous transitions take place: the completed job from $i$ to $j$ and the job from $j$.

In a general network topology where more than one queue can complete job towards a full queue $j$, an “unblocking” policy has to be defined. Usually, the First Blocked First Unblocked (FBFU) policy is considered the fairest policy: first unblock the job was blocked first. In the following we assume FBFU policy.
BAS blocking has been used mainly to model production systems and disk I/O subsystems. Recently, this blocking was studied for two stages tandem queues with MAP and phase-type distributions [GoMar06].

3. MAP Blocking Queueing Networks

We introduce the class of MAP queueing networks supporting nonrenewal service which is studied in the rest of the paper. A summary of the main notation is given in Table I. In this section, we present the case of BAS blocking mechanism. The RS-RD blocking mechanism case is presented in Section 4.

3.1 Model Definition

We consider a closed network with finite single-server queues, which serve jobs according to a MAP service time process and under work-conserving FCFS scheduling. Each queue has finite capacity $F_i$ and the same blocking mechanism. 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 queue $i$, a job joins queue $j$ with fixed probability $p_{ij}$.

<table>
<thead>
<tr>
<th>TABLE I</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUMMARY OF MAIN NOTATION</td>
</tr>
<tr>
<td>$h, k, u$ phase indexes</td>
</tr>
<tr>
<td>$i, j, m$ queue indexes</td>
</tr>
<tr>
<td>$k$ phase vector, i.e., active phases</td>
</tr>
<tr>
<td>$k_i$ active phase at queue $i$ in $\bar{k}$</td>
</tr>
<tr>
<td>$K_i$ number of phases in queue $i$’s MAP</td>
</tr>
<tr>
<td>$K_{max}$ maximum $K_i$, $1 \leq i \leq M$</td>
</tr>
<tr>
<td>$M$ number of queues in the network</td>
</tr>
<tr>
<td>$\mu_i$ mean service rate of queue $i$</td>
</tr>
<tr>
<td>$\mu_i^{k,h}$ completion rate of queue $i$, phase change $k \rightarrow h$</td>
</tr>
<tr>
<td>$\nu_i^{k,h}$ background trans. rate of queue $i$, phase $k \rightarrow h$</td>
</tr>
<tr>
<td>$N$ number of jobs in the network</td>
</tr>
<tr>
<td>$n_i$ number of jobs at queue $i$</td>
</tr>
<tr>
<td>$p_{ij}$ routing prob. from queue $i$ to queue $j$</td>
</tr>
<tr>
<td>$\pi(n_i, h, n_j, k)$ prob. of $n_i$ jobs in queue $i$ in phase $h$ and $n_j$ jobs in queue $j$ in phase $k$</td>
</tr>
<tr>
<td>$q_i^{k,h}$ rate for a queue $i$ job completion towards queue $j$, phase change $k \rightarrow h$</td>
</tr>
<tr>
<td>$Q_i$ mean queue-length at queue $i$</td>
</tr>
<tr>
<td>$Q_i^k$ mean queue-length at queue $i$ in phase $k$</td>
</tr>
<tr>
<td>$U_i$ mean utilization of queue $i$</td>
</tr>
<tr>
<td>$U_i^k$ mean utilization of queue $i$ in phase $k$</td>
</tr>
<tr>
<td>$U_{ef}$ mean effective utilization of queue $i$</td>
</tr>
</tbody>
</table>
Uefi\textsuperscript{k} mean effective utilization of queue i in phase k

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 [ZhaCaSmi08]. If $K_i = 1$, then the MAP reduces to an exponential distribution, otherwise it generates service time samples that are phase-type (PH) distributed [Neu89]. 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) [FiMe93]. It should be nevertheless remarked that MAP fitting can be still a challenging problem if the data has an irregular temporal dependence structure, see [HorTe02] for a review. We point to [ZhaCaSmi08] 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 leaves 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). It is worth noting that if a queue is blocked it completely stops its activity. As a consequence, a phase transition cannot occur during the blocking time. Note that this holds for BAS blocking, but not for RS-RD where a queue is never effectively blocked.

In this representation of queue i’s MAP, $\mu_i^{k,h}$ is the element in row $k$ and column $h$ of the $D_1$ matrix; $v_i^{k,h}$ is in row $k$ and column $h$ of $D_0$. We point the reader to [HorTe02] and references therein for background on MAPs and MAP fitting.

### 3.2. Underlying Markov Process for BAS blocking

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 $s = (s_1, s_2, \ldots, s_M)$, where $s_i$ is the state of queue $i$ with $s_i = (n_i, b_i, m_i, k)$ defined as follows: $n_i$ is the queue population (including the job in service); $b_i$ is the state of node $i$ (1=blocked, 0=active); $m_i$ is the list of queues blocked on queue $i$; $k \in K_i$ is the phase of queue $i$. As stated in section 2., the First Blocked First Unblocked order is assumed. Head($m_i$) is the queue that will be unblocked by a departure from queue $i$ in state $s$.

In the following, for the sake of simplicity, we assume to omit $b_i=0$ (queue $i$ is blocked) and $m_i$ when this is empty (there are no queues blocked on the considered queue). Finally, let $E_{\text{BAS}}$ be the state space of the queueing network when all queues behave according to BAS blocking.
According to this space, the Markov process transitions have rate \( q_{i,j}^{k,h} \) from state \( s=(s_1, s_2, ..., s_M) \) to \( s'=(s'_1, s'_2, ..., s'_M) \); the rate is given by

\[
q_{i,j}^{k,h} = \begin{cases} 
    p_{ij} \mu_{i}^{k,h}, & i \neq j, \\
    v_{i}^{k,h} + p_{ij} \mu_{i}^{k,h}, & i = j \text{ and } k \neq h.
\end{cases}
\]

The state \( s' \) is defined in Table II according to the blocking and unblocking rules and the simultaneous transitions characterizing BAS blocking. In particular, the first row corresponds to the normal transition between \( i \) and \( j \) when queue \( j \) is not full and queue \( i \) does not have blocked queues on it. The second row corresponds to the case when queue \( i \) is full with queue \( m \) blocked on it and queue \( j \) is not full (for the aim of simplicity, we omit the case of multiple simultaneous transitions. The interested reader can refer [BalDenOnv01]). Finally the third row corresponds to the case of blocking, since queue \( j \) is full. Note that the remaining components of \( s' \) are unchanged in respect of \( s \). All these transitions have rate \( q_{i,j}^{k,h} \).

**TABLE II**

<table>
<thead>
<tr>
<th>( s'=(s'_1, s'_2, ..., s'_M) )</th>
<th>Conditions on ( s=(s_1, s_2, ..., s_M) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s'_1 = (n_i-1, b'_i, m'_i, h) \land s'_2 = (n_j+1, b'_j, m'_j, k'_j) )</td>
<td>( b_i = 0, n_i &lt; F_j, m_i = \emptyset )</td>
</tr>
<tr>
<td>( s'_2 = (n_i, b'_i, m'_i, k'_i) \land s'_3 = (n_j+1, b'_j, m'_j, k'_j) \land s'_4 = (n_m-1, b'_m, m'_m, k'_m) )</td>
<td>( b_i = 0, n_i = F_i, n_j &lt; F_j, m_i = \emptyset, \text{ head}(m_i)=m, m_m = \emptyset )</td>
</tr>
<tr>
<td>( s'_3 = (n_i, b'_i, m'_i, h) \land s'_4 = (F_j, b'_j, m'_j, k'_j) )</td>
<td>( b_i = 0, n_j = F_j )</td>
</tr>
</tbody>
</table>

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 \( p_{ij} \mu_{i}^{k,h} \). 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. 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. In particular, all states with \( n_i > F_i \) are cutted. On the other hand, all the different components \( b_i, m_i \) increase the state space cardinality of the cutted subspace. To the best of our knowledge, a formula does not exist to compute the state space cardinality for blocking networks.
The state subspace of queue $i$ in phase $k$ can be partitioned in three subspaces: an idle subspace in which queue $i$ is idle $I_i^k$, a blocked subspace in which queue $i$ is blocked $B_i^k$, an active subspace in which queue $i$ is neither empty nor blocked $A_i^k$, that is queue $i$ is active.

**Definition 1 (Marginal State Spaces)**

The idle condition subspace $I_i^k = \{s : s_i(0,0,\emptyset, k)\}$ is the set of states of the MAP network where queue $i$ is empty ($n_i=0$), active ($b_i=0$), with no node blocked on it ($m_i=\emptyset$) and in phase $k \in K_i$.

The active condition subspace $A_i^k = \{s : s_i(n_i,0,m_i,k)\}$ is the set of states of the MAP network where queue $i$ is not empty ($n_i>0$), active ($b_i=0$) and in phase $k \in K_i$. Note that if queue $i$ is full ($n_i=F_i$), it is also possible that some node is blocked on it ($m_i \neq \emptyset$).

The blocked condition subspace $B_i^k = \{s : s_i(n_i,1,m_i,k)\}$ is the set of states of the MAP network where queue $i$ is not empty ($n_i>0$), blocked ($b_i=1$) and in phase $k \in K_i$. Note that if queue $i$ is full ($n_i=F_i$), it is also possible that some node is blocked on it ($m_i \neq \emptyset$).

As a summarizing example, the MAP network in Figure 1 with routing probabilities $p_{11}, p_{12}, p_{13} = 1−p_{11}−p_{12}$ at the first queue and $p_{21} = 1, p_{31} = 1$, at the remaining queues has underlying Markov process as shown in Figure 2. In the last figure, each queue $i$ has finite capacity $F_i = 2$ and the total number of jobs circulating in the network is $N=3$. Two queues are exponential with rates $\mu_1=\mu_1^{k,1}$ and $\mu_2=\mu_2^{k,1}$, 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.

![Fig. 1. Example network composed by two exponential queues and a MAP queue.](image-url)
process transition will be into the state $(2, 1, 0, 1)$ in the idle-subspace if the completed job is destined to queue 2, while it will be into the state $(2, 0, 1)$ in the active-subspace if the completed job is destined to queue 3.

In the following figures 3.a and 3.b the active subspaces are shown for queues 1 and 2 respectively. The blocked and empty subspaces can be easily derived.

In the following section, we present a “partial” balance analysis. Indeed, we prove partial balance equations that relate the performance indexes of the queues.
3.3. Marginal balance and performance indexes for BAS blocking

In this Section, we define a sort of “partial” balance, although the class of models considered in this paper in non-product-form. The first step is the definition of a marginal steady state probability that relates two queues of the network.

Let us define the marginal probability function that relates two queues in the network:

\[ \pi(n_i,k,n_j,u) = \sum_{s' \in E_{BAS}} \pi(s') \]

(2)

where \( E_{BAS} \) is the state space, that is \( E_{BAS} = \{ (s_1, s_2, ..., s_M) \mid s_i = (n_i, b_i, m_i, k_i), 0 \leq n_i \leq F_i, \sum_{i=1}^M n_i = N \} \).

The function represents the joint steady state probability to have queue \( i \) in phase \( k \) with \( n_i \) jobs and queue \( j \) in phase \( u \) with \( n_j \) jobs.

We also define a subspace \( Y_i^k \) marginal probability function \( \pi_i^k(n_i,k,n_j,u) \) as the marginal probability function defined in (2), but restricted to the states in the subspace \( Y_i^k \), where \( Y_i^k \subseteq \{ I_i^k, B_i^k, A_i^k \} \), that is

\[ \pi_i^k(n_i,k,n_j,u) = \sum_{s' \in Y_i^k} \pi(s') \]

(3)

Let us define the performance indexes in terms of the marginal probability function.

The mean queue length of queue \( i \) is \( Q_i = \sum_{k=1}^{K_i} Q_i^k \) with

\[ Q_i^k = \sum_{n_j=0}^{\min\{N,F_j\}} \sum_{n_j=0}^{\min\{N,F_j\}} \sum_{h=1}^{K_j} n_j \pi(n_i,k,n_j,h) \]

(4)

with respect to any queue \( j, 1 \leq j \leq M \).

Let us define the classical utilization of queue \( i \) as

\[ U_i = \sum_{k=1}^{K_i} U_i^k \]

(5)

where we denote by \( U_i^k \) the classical utilization of queue \( i \) in phase \( k \) and it can be computed as follows:
March, 22nd 2010

\[ U_i^k = \sum_{n_i=1}^{\min\{N,F_i\}} \sum_{n_j=0}^{\min\{N,F_j\}} \sum_{h=1}^{K_i} \pi(n_i,k,n_j,h) \]  

(6)

with respect to any queue j, 1 \leq j \leq M.

According to the blocking theory, the classical utilization is a measure of the occupancy degree of the queue, but this does not correspond to the queue productivity, since a not-idle queue could even be blocked if its destination is full.

For the aim to measure the effective productivity of a queue, we have to define an “effective utilization“. The effective utilization of queue i is defined as follows:

\[ Ue_i = \sum_{k=1}^{K_i} Uef_i^k \]  

(7)

where we denote by \( Uef_i^k \) the effective utilization of queue i in phase k and it can be computed as follows:

\[ Uef_i^k = \sum_{s' \in A_i^k : n'_i = n_i, k' = k} \pi(n_i,k,n_i,k) \]  

(8)

As introduced above, the effective utilization takes into account the “useful” 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.

Let us define the mean queue length of i \( C_j^k(i) \) in the active and blocked subspace, \( A_j^k \cup B_j^k \), of queue j in phase k:

\[ C_j^k(i) = \sum_{n_j=1}^{\min\{N,F_j\}} \sum_{n_i=1}^{\min\{N,F_i\}} \sum_{h=1}^{K_i} n_i \pi(n_i,h,n_j,k) \]  

(9)

As a consequence, \( C_j^k(j) = Q_j^k \). The following theorem relates classical utilization with mean queue length.

**Theorem 1**

In the state subspace where queue j is not-idle and in phase k, the \( C_j^k(i) \) sum to \( NU_j^k \), i.e.

\[ \sum_{i=1}^{M} C_j^k(i) = NU_j^k \]  

(10)

\[ 1 \leq k \leq K_j \]
Proof

By using (6) and the population constraint, we have

\[ NU^k_j = N \sum_{n_j=1}^{\min\{N,F_j\}} \sum_{n_i=0}^{\min\{N,F_i\}} \sum_{h=1}^{K_i} \pi(n_j,k,n_i,h) = \sum_{n_j=1}^{\min\{N,F_j\}} \sum_{n_i=0}^{\min\{N,F_i\}} \sum_{h=1}^{K_i} \pi(n_j,k,n_i,h) = \]

\[ = \sum_{n_j=1}^{\min\{N,F_j\}} \sum_{n_i=0}^{\min\{N,F_i\}} \sum_{h=1}^{K_i} n_z \pi(n_j,k,n_i,h) \]

and by the definition (6) one can choose any \( i \), with \( 1 \leq i \leq M \). So by choosing \( i=z \), the following holds:

\[ NU^k_j = \sum_{n_j=1}^{\min\{N,F_j\}} \sum_{n_z=0}^{\min\{N,F_z\}} \sum_{h=1}^{K_z} n_z \pi(n_j,k,n_z,h) = \sum_{z=1}^{M} C^k_j(z) \]

Note that in the theorem 1 the classical utilization is considered, since we are interested in the computation of mean queue length. So we have to include the “blocking period”.

In the following theorem we derive a balance between the effective utilization of queue \( i \) in all its phases. Indeed, for the BAS blocking the transitions are associated only to the non-blocking states.

As a consequence, since the Theorem 2 represents a balance between transitions, it can only consider the effective utilization.

**Theorem 2**

The utilizations of queue \( i \) in its \( K_i \) phases are in equilibrium, i. e.,

\[ \sum_{j=1}^{M} q_{i,j}^{k,h} \left( n_i \right) U_{ef,i}^k = \sum_{j=1}^{M} q_{i,j}^{h,k} \left( n_i \right) U_{ef,i}^h \]

(11)

**Proof**

Let \( \delta_m \) be a binary variable that is one if and only if queue \( m \) is not-idle in state \( (\bar{n},\bar{k}) \), i.e. \( n_m \geq 1 \).

Let us consider the global balance equation for state \( (\bar{n},\bar{k}) \):

\[ \sum_{m=1}^{M} \sum_{j=1}^{M} O_{m,j}(\bar{n},\bar{k}) = \sum_{m=1}^{M} \sum_{j=1}^{M} I_{m,j}(\bar{n},\bar{k}) \]

where for \( j=m \)

\[ O_{m,m}(\bar{n},\bar{k}) = \delta_m \sum_{h=1}^{K_m} q_{m,m}^{k,h} \pi(\bar{n},\bar{k}) \]
\[ I_{m,m}(\bar{n}, \bar{k}) = \delta_m \sum_{h=1}^{K_m} q_{m,m}^{h,k} \pi(\bar{n}, \bar{k} + (h-k)\bar{e}_m) \]

and for \( j \neq m \)

\[ O_{m,j}(\bar{n}, \bar{k}) = \delta_m \sum_{h=1}^{K_m} q_{m,j}^{h,k} \pi(\bar{n}, \bar{k}) \]

\[ I_{m,j}(\bar{n}, \bar{k}) = \delta_j \sum_{h=1}^{K_m} q_{m,j}^{h,k} \pi(\bar{n} - \bar{e}_j + \bar{e}_m, \bar{k} + (h-k)\bar{e}_m) \]

By considering all global balance equations of states in which queue \( i \) is in phase \( k \), we evaluate the following identity relation:

\[ \sum_{(\bar{n}, \bar{k}):k_i = k} \sum_{m=1}^{M} \sum_{j=1}^{M} (O_{m,j}(\bar{n}, \bar{k}) - I_{m,j}(\bar{n}, \bar{k})) = 0. \]

For \( m = i \) and \( j = i \)

\[ \sum_{(\bar{n}, \bar{k}):k_i = k} \sum_{h=1}^{K_i} \left( q_{i,i}^{k,h} \pi(\bar{n}, \bar{k}) - q_{i,i}^{h,k} \pi(\bar{n}, \bar{k} + (h-k)\bar{e}_i) \right) = \]

\[ = \min\{N, F_j\} \sum_{n_i=1}^{n} \left( \sum_{h=1}^{K_i} \left( q_{i,i}^{k,h} \pi(n_i, k, n_i, k) - q_{i,i}^{h,k} \pi(n_i, h, n_i, h) \right) \right) = \]

\[ = \sum_{h=1}^{K_i} \left( q_{i,i}^{k,h} U_i^k - q_{i,i}^{h,k} U_i^h \right) \]

for \( m = i \) and \( j \neq i \)

\[ \sum_{(\bar{n}, \bar{k}):k_i = k} \sum_{j=1}^{M} \sum_{h=1}^{K_i} \left( \delta_j q_{i,j}^{k,h} \pi(\bar{n}, \bar{k}) - \delta_j q_{i,j}^{h,k} \pi(\bar{n} - \bar{e}_j + \bar{e}_i, \bar{k} + (h-k)\bar{e}_i) \right) = \]

\[ = \sum_{h=1}^{K_i} \left( q_{i,j}^{k,h} U e f_i^k - q_{i,j}^{h,k} U e f_i^h \right) \]

for \( m \neq i \) and \( 1 \leq j \leq M \)

\[ \sum_{(\bar{n}, \bar{k}):k_i = k} \sum_{m=1}^{M} \sum_{j=1}^{M} \sum_{h=1}^{K_m} \left( \delta_m q_{m,j}^{u,h} \pi(\bar{n}, \bar{k}) - \delta_j q_{m,j}^{h,u} \pi(\bar{n} - \bar{e}_j + \bar{e}_m, \bar{k} + (h-k)\bar{e}_m) \right) = 0 \]
\[ \sum_{\tilde{n}, \tilde{k} : k_i = k} \sum_{m=0}^{M} \left( \sum_{j=1}^{M} \sum_{h \neq m}^{K_j} \frac{d_{m,j}^e f^u - d_{m,j}^h f^h}{d_{m,m}^u - q_{m,m}^h f^h} \right) + \sum_{h \neq u}^{K_j} \frac{d_{m,m}^h f^u - q_{m,m}^h f^h}{d_{m,m}^u - q_{m,m}^h f^h} = 0 \]

In the following theorem, we prove that a “marginal balance” holds between the marginal probabilities similarly to the global balance between the steady state probabilities.

**Theorem 3**

The arrival rate at queue \( i \) when its queue-length is \( n_i \) jobs, \( 0 < n_i \leq \min \{ N, F_i j-1 \} \), is balanced by the rate of departures when the queue-length is \( n_i + 1 \), i.e.

\[ \sum_{j=1}^{M} \sum_{k_i}^{K_i} \sum_{m \in A_j^k} \sum_{n_i, n_j, k}^{n_i, n_j, k} q_{j,i}^{k,h} \pi_j^k(n_i, u, n_j, k) \]

\[ = \sum_{j=1}^{M} \sum_{n_j = 0}^{n_j} \sum_{n_j = 1}^{n_j} \sum_{n_i, n_j, k}^{n_i, n_j, k} q_{j,i}^{k,h} \pi_j^k(n_i, u, n_j, k) + \]

\[ + \sum_{j=1}^{M} \sum_{n_j = 1}^{n_j} \sum_{n_j = 1}^{n_j} \sum_{n_i, n_j, k}^{n_i, n_j, k} q_{j,i}^{k,h} \pi_j^k(n_i, u, n_j, k) \]

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

\[ \sum_{j=1}^{M} \sum_{k_i}^{K_i} \sum_{m \in A_j^k} \sum_{n_i, n_j, k}^{n_i, n_j, k} q_{j,i}^{k,h} \pi_j^k(n_i, u, n_j, k) \]

\[ = \sum_{j=1}^{M} \sum_{n_j = 0}^{n_j} \sum_{n_j = 1}^{n_j} \sum_{n_i, n_j, k}^{n_i, n_j, k} q_{j,i}^{k,h} \pi_j^k(n_i, u, n_j, k) + \]

\[ + \sum_{j=1}^{M} \sum_{n_j = 1}^{n_j} \sum_{n_j = 1}^{n_j} \sum_{n_i, n_j, k}^{n_i, n_j, k} q_{j,i}^{k,h} \pi_j^k(n_i, u, n_j, k) \]

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

**Proof**

The left side of the equation considers all the departures from queue \( j \) towards queue \( i : j \) has to be in the active subspace and \( i \) cannot be full. All these departures yield the population in \( i \) to become \( n_i + 1 \), except for the case in which \( i \) is unblocked by the departure from \( j \) (head(\( m_j \)) = \( i \)), that is \( i \) was waiting for free space in \( j \), and by effect of the simultaneous transitions the population in \( i \) would remain \( n_i \).
The right side of the equation considers all the departures from queue \(i\), when the population in \(i\) is \(n_i+1\). All these departures yield the population in \(i\) to become \(n_i\). These departures are:

a. the transitions from \(i\) towards \(j\): \(i\) has to be active and \(j\) non-full. Moreover, queue \(i\) doesn’t have blocked queues on it (\(m_{i} \neq \emptyset\)), otherwise its population would still remain \(n_i+1\);

b. the transitions between any pair of nodes \(j\) and \(l\), with \(j, l \neq i\), when the queue \(j\) is full and the queue \(i\) is the first blocked node in the \(j\) list \(m_j\), that is \(head(m_j) = i\). These transitions trigger a simultaneous one from queue \(i\), thus decreasing its population to \(n_i\).

Let \(S(k, n) = \{s = (s_1, s_2, ..., s_M) \mid s_i = (n_i', b_i', m_i', k_i') : n_i' \leq n_i, k_i' = k\}\), since the theorem requires \(n_i \leq min\{N, F_{i}\}-1\) there always exists the related set \(\bar{S}(k, n) = \{s = (s_1, s_2, ..., s_M) \mid s_i = (n_i', b_i', m_i', k_i') : n_i' \geq n_i + 1, k_i' = k\}\).

The equilibrium probability flux exchanged by \(\bigcup_{k=1}^{K_i} S(k, n)\) and \(\bigcup_{k=1}^{K_i} \bar{S}(k, n)\) must be in balance because their union is the entire state space. We seek for a representation of the exchanged probability flux using the marginal probabilities. The flux \(F\) from \(\bigcup_{k=1}^{K_i} \bar{S}(k, n)\) to \(\bigcup_{k=1}^{K_i} S(k, n)\) needs to decrease the queue-length of queue \(i\) to \(n_i\). By considering that batch completions are not allowed, these transitions are the two cases a. and b. described above. Therefore \(F\) is the following flux of job completions:

\[
F = \sum_{j=1}^{M} \sum_{k=1}^{K_j} \sum_{s' \in \Lambda_j^{k} : n_i' = n_i + 1, m_i' = \emptyset, n_i' < F_j} \sum_{h=1}^{K_j} q_{j,h}^{k} \pi(s') + \\
+ \sum_{j=1}^{M} \sum_{l=1}^{M} \sum_{k=1}^{K_j} \sum_{s' \in \Lambda_j^{k} : n_i' = n_i + 1, head(m_j) \neq i, n_i' < F_l} \sum_{l=1}^{K_l} q_{j,l}^{k} \pi(s')
\]

which excludes the self-routed jobs (case \(j = i\)) that do not decrease \(n_i + 1\) to \(n_i\). Note that the job transition towards queue \(x\) is possible only if \(x\) is not full \((n_x < F_x)\), \(x \neq j, l\). The opposite flux \(G\) needs to increase the queue-length of queue \(i\) to \(n_i+1\). The transitions towards states where \(i\) has \(n_i+1\) jobs are allowed from states where the following conditions hold: the sending queue \(j\) is not blocked, \(head(m_j) \neq i\) such that a simultaneous transition doesn’t occur (otherwise the \(i\) population doesn’t change), queue \(i\) is not full \((n_i < F_i)\). Therefore \(G\) is the following flux of job completions:

\[
G = \sum_{j=1}^{M} \sum_{k=1}^{K_j} \sum_{h=1}^{K_j} \sum_{s' \in \Lambda_j^{k} : n_i' = n_i, m_j' = \emptyset, \text{head}(m_j) \neq i} \pi(s')
\]

and describes all possible transitions that bring a job from queue \(j \neq i\).
However by the given definition (3)

\[
\sum_{j\neq i}^{M} \sum_{k=1}^{K_j} \sum_{s'\in A_j^k : n'_i = n_i+1, \mathbf{m}_j = \emptyset, n'_j < F_j} \sum_{h=1}^{K_i} q_{i,j}^{k,h} \pi(s') + \\
+ \sum_{j\neq i}^{M} \sum_{l=1}^{M} \sum_{k=1}^{K_j} \sum_{s'\in A_l^k : n'_i = n_i+1, \text{head}(\mathbf{m}_j) = i, n'_j < F_j} \sum_{h=1}^{K_i} q_{i,j}^{k,h} \pi(s') = \\
= \sum_{j\neq i}^{M} \sum_{n_j = 0}^{\min\{N,F_j\}} \sum_{l=1}^{K_j} \sum_{k=1}^{K_i} \sum_{s'\in A_l^k : n'_i = n_i+1, \mathbf{m}_l = \emptyset, n'_j = n_j, k' j = u} \sum_{h=1}^{K_i} q_{i,j}^{k,h} x_i^k(n_i+1,k,n_j,u) + \\
+ \sum_{j\neq i}^{M} \sum_{l=1}^{M} \sum_{n_j = 0}^{\min\{N,F_j\}} \sum_{l=1}^{K_j} \sum_{k=1}^{K_i} \sum_{s'\in A_l^k : n'_i = n_i+1, \text{head}(\mathbf{m}_j) = i} \sum_{h=1}^{K_i} q_{i,j}^{k,h} x_j^k(n_j,k,n_i,u)
\]

and

\[
\sum_{j\neq i}^{M} \sum_{k=1}^{K_j} \sum_{h=1}^{\mathbf{h}_j^k} \sum_{s'\in A_j^k : n'_i = n_i, \mathbf{m}_j = \emptyset, \text{head}(\mathbf{m}_j) = i} \pi(s') = \\
= \sum_{j\neq i}^{M} \sum_{k=1}^{K_j} \sum_{h=1}^{\mathbf{h}_j^k} \sum_{s'\in A_j^k : n'_i = n_i, \mathbf{m}_j = \emptyset, \text{head}(\mathbf{m}_j) = i} \sum_{h=1}^{K_i} q_{i,j}^{k,h} x_j^k(n_i, u, n_j, k)
\]

and by substituting them

\[
F = \sum_{j\neq i}^{M} \sum_{n_j = 0}^{\min\{N,F_j\}} \sum_{l=1}^{K_j} \sum_{k=1}^{K_i} \sum_{s'\in A_l^k : n'_i = n_i+1, \mathbf{m}_l = \emptyset, n'_j = n_j, k' j = u} \sum_{h=1}^{K_i} q_{i,j}^{k,h} x_i^k(n_i+1,k,n_j,u) + \\
+ \sum_{j\neq i}^{M} \sum_{n_j = 0}^{\min\{N,F_j\}} \sum_{l=1}^{M} \sum_{n_j = 0}^{\min\{N,F_j\}} \sum_{l=1}^{K_j} \sum_{k=1}^{K_i} \sum_{s'\in A_l^k : n'_i = n_i+1, \text{head}(\mathbf{m}_j) = i} \sum_{h=1}^{K_i} q_{i,j}^{k,h} x_j^k(n_j,k,n_i,u)
\]

\[
G = \sum_{j\neq i}^{M} \sum_{k=1}^{K_j} \sum_{h=1}^{\mathbf{h}_j^k} \sum_{s'\in A_j^k : \mathbf{m}_j = \emptyset, \text{head}(\mathbf{m}_j) = i} \sum_{h=1}^{K_i} q_{i,j}^{k,h} x_j^k(n_i, u, n_j, k)
\]

and by imposing the equilibrium balance \( F = G \) for \( n_i \geq 1 \) we find immediately (12).

Note that (12) would hold also for \( n_i = 0 \); nevertheless, in this case we can give the more detailed condition (13) by recalling that if \( n_i = 0 \) 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 \( S(k, n_i=0) \equiv \{ s = (s_1, s_2, ..., s_M) | s_1 = (n'_i, h'_i, \mathbf{m}_i, k'_i) : n'_i \leq 0, k'_i = u \} \), and \( \cup_{k=1}^{K_i} \mathbb{S}(k, n_i=1) \).
The following theorem can be seen as an extension of the Theorem 1. Indeed Theorem 4 defines a relation between the sum of mean queue-lengths of all queues and the classical utilization of any queue \( i \) when a queue \( j \) is in phase \( k \), for any \( j \) and \( k \).

**Theorem 4**

The sum of mean queue-lengths of all queues \( t, t=1, \ldots, M \), when queue \( j \) is in phase \( k \), satisfies

\[
\sum_{t=1}^{M} \left( \min\{N,F_i\} \sum_{n_j=0}^{K_i} \sum_{h=1}^{K_j} n_t \pi(n_t,h,n_j,k) \right) \geq N \sum_{h=1}^{K_i} \sum_{n_j=0}^{K_j} \sum_{n_i=1}^{K_i} n_i \pi(n_i,h,n_j,k) \tag{16}
\]

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

**Proof**

Letting \( \sum_{s_j=(n_j,b_j,m_j,k_j):k_j=k} = \sum_{s \in A_j^k \cup B_j^k \cup I_j^k} s_j=(n_j,b_j,m_j,k_j):k_j=k \), we have

\[
N \sum_{s_j=(n_j,b_j,m_j,k_j):k_j=k} \pi(s) = \sum_{t=1}^{M} \sum_{s_j=(n_j,b_j,m_j,k_j):k_j=k} n_t \pi(s) = \sum_{t=1}^{M} \left( \min\{N,F_i\} \sum_{n_j=0}^{K_j} \sum_{h=1}^{K_i} n_t \pi(n_t,h,n_j,k) \right)
\]

On the other hand

\[
N \sum_{s_j=(n_j,b_j,m_j,k_j):k_j=k} \pi(s) \geq N \sum_{n_i=1}^{K_i} \sum_{n_j=0}^{K_j} \pi(n_i,h,n_j,k)
\]

since in the first member there are also states with \( n_i=0 \).

Finally in the following theorem, all the performance indexes of queue \( i \) in all its phase are related.

**Theorem 5**

The performance indexes of queue \( i \) in phase \( k \) and in phase \( h \) are related by the following equation
\[ \sum_{h=1}^{K_i} \sum_{j=1}^{M} \sum_{n_j=0}^{\min\{N,F_j\}} \sum_{u=1}^{K_j} q_{i,j}^{k,h} n \pi(n_i,k,n_j,u) + \]
\[ + \sum_{j=1}^{M} \sum_{h=1}^{K_i} \sum_{n_j=0}^{\min\{N,F_j\}} \sum_{u=1}^{K_j} \sum_{s' \in A^h_{j,n'_i=n_i,k'_i=k,n'_j=n_j,k'_j=u}} q_{i,j}^{h,k} \pi(n_i,h,n_j,u) = (17) \]
\[ = \sum_{j=1}^{M} \sum_{h=1}^{K_i} \sum_{u=1}^{K_j} \sum_{s' \in A^h_{j,n'_i=n_i,k'_i=k,n'_j=n_j,k'_j=h}} \sum_{m_j=\text{top}(m_j)=i}^{\min\{N,F_j\}} q_{i,j}^{h,k} \pi(n_i,k,n_j,h) + \]
\[ + \sum_{h=1}^{K_i} \sum_{j=1}^{M} \sum_{n_j=0}^{\min\{N,F_j\}} \sum_{u=1}^{K_j} \sum_{s' \in A^h_{j,n'_i=n_i,k'_i=h,n'_j=n_j,k'_j=u}} q_{i,j}^{h,k} n \pi(n_i,h,n_j,u) \]

**Proof**

Let us consider the weighted sum of all global balance equations of states in which queue \( i \) is in phase \( k \):

\[ \sum_{A^k_i} \cup_{B^k_i} \cup_{U^k_i} n_i \sum_{m=1}^{M} (O_{m,j}(s) - I_{m,j}(s)) = 0 \]

The proof follows a similar high-level structure than the proof of theorem 2.

**4. MAP Queueing Networks with RS-RD blocking**

In this section we present MAP queueing networks with RS-RD blocking. In this case, since the given mechanism definition, a simple state notation is enough to characterize unambiguously the network behaviour. A feasible network state in the queueing network underlying Markov process is a tuple \((\vec{n}, \vec{k})\), where \( \vec{n} = (n_1, n_2, \ldots, n_M)\), \( 0 \leq n_i \leq F_i, \sum_{i=1}^{M} n_i = N \), 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. Note that this state description coincides with the one used for the non-blocking case. According to this space, the Markov process transitions have rate \( q_{i,j}^{k,h} \) from state \((\vec{n}, \vec{k})\) to \((\vec{n}-\vec{e}_t+\vec{e}_j, \vec{k}')\), \( k_i = k, k'_i = h \), where \( \vec{e}_j \) is a vector of zeros with a one in the \( t \)-th position; the rate is given by (1). Finally, let \( E_{RS-RD} \) be the state space of the queueing network when all queues behave according to RS-RD blocking.

As stated above for the BAS case, the state space of a blocking network is a subset of the state space of the same network but with infinite capacity queues. In particular, for RS-RD, the state space can be simply obtained by cutting all states with \( n_i > F_i \). To the best of our knowledge, a
formula does not exist to compute the state space cardinality for blocking networks. The interested reader can refer to [BalDenOnv01] for a recursive expression to compute the state space cardinality for a queueing network in which all queues have the same capacity and RS-RD blocking.

By considering the network of Figure 1, the underlying Markov process is shown in Figure 4.

The notation, e.g., (102, 1) indicates that the exponential queue 2 is idle, the exponential queue 1 has 1 job in service, and the MAP queue has two jobs and is in phase 1. According to RS-RD mechanism, in this state queue 1 is “blocked” with probability $0 < p_{13} < 1$ (it is working for a full queue). In (210, 2), the phase 2 is the phase left active by the last served job and queue 2 is “blocked” with probability $p_{21} = 1$.

In the following section, we present the “partial” balance analysis for the case of RS-RD blocking.

4.1. Marginal balance and performance indexes for RS-RD blocking

Let us define the marginal probability function that relates two queues of the network:

$$\pi(n_i,k,n_j,u) = \sum_{(\bar{n}', \bar{k}') \in E_{RS-RD}} \pi(\bar{n}', \bar{k}')$$

(18)

where $E_{RS-RD}$ is the state space, that is $E_{RS-RD} = \{(\bar{n}, \bar{k}) \mid \bar{n} = (n_1, n_2, \ldots, n_M), 0 \leq n_i \leq F_i, \sum_{i=1}^M n_i = N\}$.

Similarly to equation (2), this represents the joint steady state probability to have queue $i$ in phase $k$ with $n_i$ jobs and queue $j$ in phase $u$ with $n_j$ jobs.

For the sake of brevity, we do not repeat the definitions that are identical to those defined for the BAS blocking. As a consequence, the mean queue length $Q_i$ and the classical utilization $U_i$ for
queue $i$ with RS-RD blocking, are defined by using the marginal probability function (18), instead of (2), in the equations (4), (5) and (6).

To measure the effective utilization for RS-RD blocking, we have to exclude the useless work destined to full queues. Indeed, according to the blocking definition, this work will have to be repeated, therefore it is considered useless work. The effective utilization $U_{ef}^{k}$ of queue $i$ in phase $k$ can be computed as follows:

$$U_{ef}^{k} = \min\{N,F_i\} \left( \pi(n_i,k,n_i,k) - \sum_{j=1}^{M} \sum_{h=1}^{K} p_{ij} \sum_{h=1}^{K} p_{ij} \pi(n_i,k,F_j,h) \right)$$

(19)

The effective utilization $U_{ef}^{k}$ of queue $i$, can be simply obtained by substituting (19) into equation (7).

Note that the mean queue length $C_{j}^{k}(i)$, defined in (9) for BAS corresponds to the mean queue length of $i$ when queue $j$ is not idle and it is in phase $k$ for RS-RD. As a consequence, the Theorem 1 holds with the same proof.

In the following theorem we derive a balance between the utilization of queue $i$ in all its phases, both classical and effective. Note that this is different from the Theorem 2 for the BAS case. Indeed, for the RS-RD mechanism, queue blocking doesn’t really occur. As a consequence, since this Theorem represents a balance between transitions, it includes also the classical utilization.

**Theorem 2’**

The utilizations of queue $i$ in its $K_i$ phases are in equilibrium, i.e.,

$$\sum_{j=1}^{M} \sum_{h=1}^{K} q_{i,j}^{h} U_{ef}^{k} + \sum_{j=1}^{M} \sum_{h=1}^{K} q_{i,j}^{h} U_{i}^{h} = \sum_{j=1}^{M} \sum_{h=1}^{K} q_{i,j}^{h} U_{ef}^{h} + \sum_{j=1}^{M} \sum_{h=1}^{K} q_{i,j}^{h} U_{i}^{h}$$

(20)

**Proof**

Let $\delta_m$ be a binary variable that is one if and only if queue $m$ is not-idle in state $(\bar{n}, \bar{k})$, i.e. $n_m \geq 1$.

Let us consider the global balance equation for state $(\bar{n}, \bar{k})$:

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

where for $j=m$

$$O_{m,m}(\bar{n}, \bar{k}) = \delta_m \sum_{h=1}^{K_m} q_{m,m}^{h}(\bar{n}, \bar{k})$$
\[ I_{m,m}(\bar{n}, \bar{k}) = \delta_m \sum_{h=1}^{K_m} q_{m,m,k}^{h,k} \pi(\bar{n}, \bar{k} + (h - k)\bar{c}_m) \]

and for \( j \neq m \)

\[ O_{m,j}(\bar{n}, \bar{k}) = \delta_m \sum_{h=1}^{K_m} q_{m,j,k}^{h,k} \pi(\bar{n}, \bar{k}) \]

\[ I_{m,j}(\bar{n}, \bar{k}) = \delta_j \sum_{h=1}^{K_m} q_{m,j,k}^{h,k} \pi(\bar{n} - \bar{c}_j + \bar{c}_m, \bar{k} + (h - k)\bar{c}_m) \]

By considering all global balance equations of states in which queue \( i \) is in phase \( k \), we evaluate the following identity relation:

\[ \sum_{(\bar{n}, \bar{k}): k_i = k} M_{j=1}^{M} \left( O_{m,j}(\bar{n}, \bar{k}) - I_{m,j}(\bar{n}, \bar{k}) \right) = 0. \]

For \( m = i \) and \( j = i \)

\[ \sum_{(\bar{n}, \bar{k}): k_i = k} \delta_i \sum_{h=1}^{K_i} \left( q_{i,i}^{k,h} \pi(\bar{n}, \bar{k}) - q_{i,i}^{h,k} \pi(\bar{n}, \bar{k} + (h - k)\bar{c}_i) \right) = \]

\[ = \min \{ N_i, F_i \} \sum_{n_i=1}^{N_i} \left( \sum_{h=1}^{K_i} \left( q_{i,i}^{k,h} \pi(n_i, k, n_i, k) - q_{i,i}^{h,k} \pi(n_i, h, n_i, h) \right) \right) = \]

\[ = \sum_{h=1}^{K_i} \left( q_{i,i}^{k,h} U_{i}^k - q_{i,i}^{h,k} U_{i}^h \right) \]

for \( m = i \) and \( j \neq i \)

\[ \sum_{(\bar{n}, \bar{k}): k_i = k} \sum_{j=1}^{M} \delta_j \sum_{h=1}^{K_j} \left( \delta_j q_{i,j}^{k,h} \pi(\bar{n}, \bar{k}) - \delta_j q_{i,j}^{h,k} \pi(\bar{n} - \bar{c}_j + \bar{c}_i, \bar{k} + (h - k)\bar{c}_i) \right) = \]

\[ = \sum_{h=1}^{K_i} \left( q_{i,j}^{k,h} U_{ef_i}^k - q_{i,j}^{h,k} U_{ef_i}^h \right) \]

for \( m \neq i \) and \( 1 \leq j \leq M \)

\[ \sum_{(\bar{n}, \bar{k}): k_i = k} \sum_{m=1}^{M} \sum_{j=1}^{M} \delta_m \sum_{h=1}^{K_m} \left( \delta_m q_{m,j}^{u,h} \pi(\bar{n}, \bar{k}) - \delta_m q_{m,j}^{h,u} \pi(\bar{n} - \bar{c}_j + \bar{c}_m, \bar{k} + (h - k)\bar{c}_m) \right) = 0 \]
\[
\sum_{(n,k):k_i=k} \sum_{m=1}^{M} \left( \sum_{j=1}^{M} \sum_{h=1}^{K_i} q_{m,j}^{u,h} u m^h - q_{m,m}^{h,u} u m^h \right) + \sum_{h=1}^{K_i} \left( d_{m,m}^{h,u} u m^h - q_{m,m}^{h,u} u m^h \right) = 0
\]

In the following theorem, we prove that a “marginal balance” holds between the marginal probabilities similarly to the global balance between the steady state probabilities.

**Theorem 3’**

The arrival rate at queue \(i\) when its queue-length is \(n_i\) jobs, \(0 < n_i \leq \min\{N, F_i\}-1\), is balanced by the rate of departures when the queue-length is \(n_i + 1\), i.e.

\[
\sum_{j=1}^{M} \sum_{n_j=1}^{\min\{N, F_j\}} \sum_{k=1}^{K_j} \sum_{u=1}^{K_i} \sum_{h=1}^{K_j} q_{j,i}^{k,h} \pi(n_i,u,n_j,k) = \\
\sum_{j=1}^{M} \sum_{n_j=1}^{\min\{N, F_j\}-1} \sum_{k=1}^{K_j} \sum_{h=1}^{K_j} q_{i,j}^{k,h} \pi(n_i+1,k,n_j,u) \tag{21}
\]

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

\[
\sum_{j=1}^{M} \sum_{n_j=1}^{\min\{N, F_j\}} \sum_{k=1}^{K_j} \sum_{h=1}^{K_j} q_{j,i}^{k,h} \pi(n_i=0,u,n_j,k) = \\
\sum_{j=1}^{M} \sum_{n_j=1}^{\min\{N, F_j\}-1} \sum_{k=1}^{K_j} \sum_{h=1}^{K_j} q_{i,j}^{k,h} \pi(n_i=1,k,n_j,h) \tag{22}
\]

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

**Proof**

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

The equilibrium probability flux exchanged by \(\bigcup_{k=1}^{K_i} S(k, n)\) and \(\bigcup_{k=1}^{K_i} \bar{S}(k, n)\) must be in balance because their union is the entire state space. We seek for a representation of the exchanged probability flux using the marginal probabilities. The flux \(F\) from \(\bigcup_{k=1}^{K_i} \bar{S}(k, n)\) to \(\bigcup_{k=1}^{K_i} S(k, n)\) needs to decrease the queue-length of queue \(i\) to \(n_i\). Only states where \(i\) has \(n_i + 1\) jobs can have transitions to states where \(i\) has \(n_i\) jobs (batch completions are not allowed); therefore \(F\) is the following flux of job completions
The performance indexes of queue $j$ are related. Finally, as for the BAS case, in the following theorem all the performance indexes of queue $i$ in all its phases are related.

**Theorem 5’**

The performance indexes of queue $i$ in phase $k$ and in phase $h$ are related by the following equation.
\[ \sum_{h=1}^{M} \sum_{j=1}^{K_i} \sum_{n_j=0}^{\min\{N,F_j\}} q_{i,j}^{h,n_j} \pi(n_i,k,n_j,u) + \sum_{n_i=1}^{\min\{N,F_i\}} q_{i,i}^{h,k} \pi(n_i,h,n_j,u) = 0 \]  

\[ \sum_{n_i=1}^{\min\{N,F_i\}} q_{i,i}^{h,k} \pi(n_i,h,n_j,u) = 0 \]  

\[ \sum_{n_i=1}^{\min\{N,F_i\}} q_{i,i}^{h,k} \pi(n_i,h,n_j,u) = 0 \]  

**Proof**

Let us consider the weighted sum of all global balance equations of states in which queue \( i \) is in phase \( k \):

\[ \sum_{i} \sum_{k_i=k} n_i \sum_{m=1}^{M} \sum_{j=1}^{M} \left( O_{m,j}(\pi,k) - I_{m,j}(\pi,k) \right) = 0 \]

The proof follows a similar high-level structure than the proof of theorem 2'.

5. Conclusions

In this paper, we consider the class of closed queueing networks with blocking and MAP service times. We consider two of the most used blocking mechanisms: BAS and RS-RD blocking. We analyze the underlying Markov processes and we prove a set of marginal balance equations that relates the blocking performance indexes. We argue that the marginal balance can be a sound framework for approximate solution techniques. A bounding analysis will be presented in a forthcoming paper.
References


