Stochastic Complement Analysis of Multi-Server Threshold Queues with Hysteresis

John C.S. Lui
The Dept. of Computer Science & Engineering
The Chinese University of Hong Kong

Leana Golubchik
Dept. of Computer Science
University of Maryland at College Park

Abstract

We consider a K-server threshold-based queueing system with hysteresis, for which a set of forward thresholds $(F_1, F_2, \ldots, F_{K-1})$ and a set of reverse thresholds $(R_1, R_2, \ldots, R_{K-1})$ are defined. A simple version of this multi-server queueing system behaves as follows. When a customer arrives to an empty system, it is serviced by a single server. Whenever the number of customers exceeds a forward threshold F_i, a server is added to the system and server activation is instantaneous. Whenever the number of customer falls below a reverse threshold R_i, a server is removed from the system. We consider and solve several variation of this problem, namely: (1) homogeneous servers with Poisson arrivals, (2) homogeneous servers with bulk (Poisson) arrivals, and (3) heterogeneous servers with Poisson arrivals. We place no restrictions on the number of servers or the bulk sizes or the size of the waiting room. In [8], the authors solve a limited form of this problem using the Green's function method. More specifically, they give a closed-form solution for a K-server system, when the servers are homogeneous, and for a 2-server system, when the servers are heterogeneous; the authors experienced difficulties in extending the Green's function method beyond the case of 2 heterogeneous servers. Rather than using Green's function, we solve this problem using the concept of stochastic complementation, which is a more intuitive and more easily extensible method. For the case of a homogeneous multi-server system we are able to derive a closed-form solution for the steady state probability vector; for the remaining cases we give an algorithmic solution. Note, however, that we can use stochastic complementation to derive closed-form solutions for some limited forms of cases (2) and (3), such as heterogeneous servers with $K = 2$ and bulk arrivals with a limited bulk size. Finally, our technique works both for systems with finite and infinite waiting rooms.

1 Introduction

A K-server hysteresis threshold-based queueing system is considered in which the number of servers, employed for servicing customers, is governed by a *forward threshold* vector $\boldsymbol{F} = (F_1, F_2, \ldots, F_{K-1})$ and a *reverse threshold* vector $\boldsymbol{R} = (R_1, R_2, \ldots, R_{K-1})$. Without loss of generality, we assume that $F_1 < F_2 < \cdots < F_{K-1}$ and $R_1 < R_2 < \cdots < R_{K-1}$. The dynamics of this type of a multi-server queueing

system can be described as follows. When the system is empty, a single server is used to service an arriving customer. If a customer arriving to a system with i active servers finds that there are already F_i customers in the queueing system, then one additional server will be activated, i.e., this server will join the set active servers for servicing existing and incoming customers. A customer departure from a system with i active servers leaving R_i customers behind will force a removal of one server. In this paper, we assume that the activation and deactivation of server is an instantaneous operation.

There are many reasons for using the threshold-based approach to control the number of servers in the system. For instance, many systems incur significant server setup, usage and removal costs. As in most cases, what concerns the system designer is not only the system performance but also its cost/performance ratio. Therefore, what we would like is for the system to use an "appropriate" number of servers so as to satisfy some performance requirements, such as the mean system response time. One approach to improving the cost/performance ratio of a system is to react to changes in workload through the use of thresholds. For example, one can maintain the expected response time of a job in a system at an acceptable level, and at the same time maintain an acceptable operating cost by dynamically activating and deactivating servers as a function of the system load.

Note that in many situations, a simple threshold-based system may not be sufficient to guarantee that the system will operate in a "stable state". In fact, it is possible to cause the system to experience effects of oscillation. One reason for avoiding oscillations in a computer system is to reduce the above mentioned server setup and removal costs, i.e., oscillations coupled with non-negligible server setup and removal costs can result in a poor cost/performance ratio of a system. More specifically, it is desirable to add servers only when a system is moving towards a heavily loaded operation region, and it is desirable to remove servers only when a system is moving towards a lightly loaded operation region. Thus, to avoid oscillation, hysteresis is introduced into the system — this is the motivation for looking for general and efficient techniques for analyzing threshold-based queueing systems with hysteresis. Note that, the forward and reverse thresholds should be "sufficiently far apart" in order to insure that the system does not degenerate to a "simple" threshold-based system (i.e., one without hysteresis behavior).

Let us begin with a literature survey of several works on threshold-based queueing systems. In [13], a two-server heterogeneous system is presented, where a conjecture is made that for a $M/M/2$ queueing system with heterogeneous service rates, the policy that optimizes system performance, such as the mean response time, is of the threshold type. In [14], this conjecture is shown to be correct. An approximate solution for solving a degenerate form of this problem is presented in [6, 7], where an arriving customer is assigned to the fastest idle server. In this degenerate case, all thresholds are set to zero. An approximate solution for a multi-server queueing system that employs (non-zero) thresholds is presented in [20]; however, this queueing system lacks hysteresis. In [19], the waiting time distribution of a two-server threshold system without hysteresis is derived. In [8], the authors solve a limited form of the multi-server threshold queueing system with hysteresis, using the Green's function method [5, 9, 10]. More specifically, they give a closed-form solution for a K-server system, when the servers are homogeneous, and for a 2-server system, when the servers are heterogeneous; the authors experienced difficulties in extending the Green's function method beyond the case of 2 heterogeneous servers. In [3], authors consider a homogeneous server system where the server activation time is exponentially distributed. In general, no closed-form solution can be obtained but tight upper and lower bounds on some performance measures (i.e., expected response time and expected number of

customers) are derived.

In this paper, we consider and solve several variations of the multi-server threshold queueing system with hysteresis, namely: (1) homogeneous servers with Poisson arrivals, (2) homogeneous servers with bulk (Poisson) arrivals, and (3) heterogeneous servers with Poisson arrivals. We place no restrictions on the number of servers or the bulk sizes or the size of the waiting room. Rather than using the Green's function method, as in [8], we solve this problem using the concept of stochastic complementation [18], which is a more intuitive and a more easily extensible method. For case (1), we are able to derive a closed-form solution for the steady state probability vector; for the remaining cases, we give an algorithmic solution for computing the steady state probability vector. Of course, given the steady state probabilities, we can compute various performance measures of interest. Thus, the contributions of this work are as follows. We present a more intuitive and extensible method (than in the case of [8]) for obtaining a closed-form solution to the multi-server threshold queueing problem with hysteresis, when the servers are homogeneous and there is no restriction on the number of servers or the waiting room size. We also present algorithmic solutions for the bulk-arrivals and heterogeneous-servers variations of the problem (again, with no restrictions on the size of the bulk or the number of servers); to the best of our knowledge, these variations of the problem, with no restriction on the number of servers or the bulk size, have not been solved exactly in the past (except for the solution of the 2-heterogeneous-servers problem in [8]). The ease with which we are able to obtain solutions to these variations of the problem demonstrates the extensibility of our method. Note, that we can use stochastic complementation to derive closed-form solutions for some limited forms of heterogeneous-servers and bulk-arrivals variations of the problem, such as heterogeneous servers with $K = 2$ and bulk arrivals with a limited bulk size. Finally, our technique works both for systems with finite and infinite waiting rooms.

The remainder of the paper is organized as follows. In Section 2 we briefly review the concept of stochastic complementation and its implications, and in Section 3 we outline the basic solution approach. In Section 4 we formally define a model of a threshold-based queueing system with hysteresis and present several variations on this system; in Section 5 we present solutions to the different variations of the system using stochastic complementation and the basic approach outlined in Section 3. Numerical results obtained using our solution technique are given in Section 6. Our conclusions are given in Section 7.

2 Background on Stochastic Complementation

In this section, we briefly describe the concept of stochastic complementation [18], which we will use extensively to derive the solution of the threshold-based queueing systems with hysteresis. For the purpose of this presentation, we assume that we are given a discrete state space, discrete time, ergodic Markov chain with a transition probability matrix $\boldsymbol{P}$. Throughout the paper we will also consider continuous time Markov processes. Note, however, that there is a simple transformation between the two; that is, given a continuous time Markov process with a rate matrix $\boldsymbol{Q}$, we can transform it to a discrete time Markov chain via uniformization [4]:

$$\boldsymbol{P} = \boldsymbol{I} + \boldsymbol{Q}/\Lambda \tag{1}$$

where $\Lambda \geq \max_i\{|q_{ii}|\}$, q_{ii} is the i^{th} diagonal element of $\boldsymbol{Q}$, and $\boldsymbol{I}$ is an identity matrix. Note that the steady state probability vectors for $\boldsymbol{P}$ and $\boldsymbol{Q}$ are identical.

Given an irreducible discrete time Markov chain, $\mathcal{M}$, with state space S, let us partition this state space into two disjoint sets A and B. Then, the one-step transition probability matrix of $\mathcal{M}$ is:

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}_{A,A} & \boldsymbol{P}_{A,B} \\ \boldsymbol{P}_{B,A} & \boldsymbol{P}_{B,B} \end{bmatrix}$$

and $\boldsymbol{\pi} = [\boldsymbol{\pi}_A, \boldsymbol{\pi}_B]$ is the corresponding steady state probability vector of $\mathcal{M}$. In what follows, we define the notion of a stochastic complement and quote some useful results [18].

Definition 1 *The stochastic complement of $\boldsymbol{P}_{A,A}$, denoted by $\boldsymbol{C}_{A,A}$, is:*

$$\boldsymbol{C}_{A,A} = \boldsymbol{P}_{A,A} + \boldsymbol{P}_{A,B}[\boldsymbol{I} - \boldsymbol{P}_{B,B}]^{-1}\boldsymbol{P}_{B,A} \tag{2}$$

Theorem 1 *The stochastic complement is always a stochastic matrix and the associated Markov chain is always irreducible, if the original Markov chain is irreducible.*

Theorem 2 *Let $\boldsymbol{\pi}_{|A}$ be the stationary state probability vector for the stochastic complement $\boldsymbol{C}_{A,A}$, then*

$$\boldsymbol{\pi}_{|A} = 1/(\boldsymbol{\pi}_A \boldsymbol{e})\boldsymbol{\pi}_A \tag{3}$$

where $\boldsymbol{e}$ is the column vector with all entries equal to 1.

The implication of the above theorems is that the stationary state probabilities of the stochastic complement are the *conditional state probabilities* of the associated states of the original Markov chain.

Let diag($\boldsymbol{v}$) be a diagonal matrix where the i^{th} diagonal element is the i^{th} element of the vector $\boldsymbol{v}$. We can re-write Equation (2) as:

$$\boldsymbol{C}_{A,A} \;=\; \boldsymbol{P}_{A,A} + \text{diag}(\boldsymbol{P}_{A,B}\boldsymbol{e})\boldsymbol{Z} \tag{4}$$

where $\boldsymbol{Z} = \boldsymbol{P}^*_{A,B}\,[\boldsymbol{I} - \boldsymbol{P}_{B,B}]^{-1}\,\boldsymbol{P}_{B,A}$ and $\boldsymbol{P}^*_{A,B}$ is simply $\boldsymbol{P}_{A,B}$ but with all the rows normalized to sum to 1. The square matrix $\boldsymbol{Z}$ is also an irreducible stochastic matrix, provided that the original Markov chain is irreducible. Let r_i be the i^{th} diagonal element of diag($\boldsymbol{P}_{A,B}\boldsymbol{e}$). The probabilistic interpretation of r_i is that it is the total probability of making a transition from state $s_i \in A$ to any state in B. Also, let $\boldsymbol{z}_i$ be the i^{th} row of $\boldsymbol{Z}$; then we can re-write Equation (4) as:

$$\boldsymbol{C}_{A,A} \;=\; \boldsymbol{P}_{A,A} + \begin{bmatrix} r_1\boldsymbol{z}_1 \\ r_2\boldsymbol{z}_2 \\ \vdots \\ r_n\boldsymbol{z}_n \end{bmatrix} \tag{5}$$

Remarks: the probabilistic interpretation of Equation (5) is as follows. If in the original Markov chain there is a transition from state $s_i \in A$ to any state in B, then in the stochastic complement this transition becomes a transition to some state(s) in A instead. In other word, the derived Markov chain "skips over" the period of time spent in B. The transition from $s_i \in A$ to B becomes a transition to $s_j \in A$ with probability z_{ij}. The stochastic complement of $P_{A,A}$ is therefore equal to $P_{A,A}$ plus any transition probabilities, which used to go from A to B, "folded" back to A and redistributed according to the stochastic matrix $\boldsymbol{Z}$. This interpretation implies that the i^{th} row of matrix $\boldsymbol{Z}$ determines how r_i should be redistributed back to A. In general, it is not an easy task to compute $\boldsymbol{Z}$, but for some special cases where sufficient "structure" exists in the original Markov chain, $\boldsymbol{Z}$ can be obtained with little or no computation.

The following theorem illustrates a special structure which we will use in analyzing the threshold-based queueing system with hysteresis.

Theorem 3 *Given an irreducible Markov process with state space S, let us partition the state space into two disjoint sets A and B. The transition rate matrix $\boldsymbol{Q}$ of this Markov process is:*

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Q}_{A,A} & \boldsymbol{Q}_{A,B} \\ \boldsymbol{Q}_{B,A} & \boldsymbol{Q}_{B,B} \end{bmatrix}$$

where $\boldsymbol{Q}_{i,j}$ is the transition rate sub-matrix corresponding to transitions from partition i to partition j. If $\boldsymbol{Q}_{B,A}$ has all zero entries except for some non-zero entries in the i-th column, then the conditional steady state probability vector (corresponding to the states in A), given that the system is in partition A, is denoted by $\boldsymbol{\pi}_{|A}$ and is the solution to the following system of linear equations:

$$\begin{aligned} \boldsymbol{\pi}_{|A} \left[\boldsymbol{Q}_{A,A} + \boldsymbol{Q}_{A,B}\ \boldsymbol{e}\ \boldsymbol{e}_i^T \right] &= \mathbf{o} \\ \boldsymbol{\pi}_{|A}\ \boldsymbol{e} &= 1 \end{aligned}$$

where $\boldsymbol{e}_i^T$ is a row vector with a 0 in each component, except a 1 in the the i-th component.

Proof: This is intuitively clear based on the stochastic complementation arguments. For detailed derivation, please refer to [2, 16, 17]. ∎

3 Basic Approach

Before we proceed with a more detailed definition of our model and the presentation of the details of the analysis, let us briefly describe the general approach we intend to use to solve the queueing problem described in Section 1. We will model this queueing system as a Markov chain, $\mathcal{M}$ (see Section 4 for a detailed definition), where: (1) the main goal is to compute the steady state probabilities of the Markov chain and use these to compute various performance metrics of interest (see Section 5) and (2) the main difficulty is that the Markov chain is infinite and thus "difficult" to solve using a "direct" approach[1].

[1] We could consider finite versions of the model; however, the Markov chain would still be very large and the computational complexity of a "direct" solution for a reasonable size system still unacceptable.

As is often done in these cases, we need to look for special structure that might exist in the Markov chain; specifically, we intend to take advantage of the stochastic complementation technique briefly described in Section 2. The basic approach to computing the steady state probabilities of the Markov process and the corresponding performance measures is as follows. We will first partition the state space of the original Markov chain $\mathcal{M}$ into disjoint sets. Using the concept of stochastic complementation (see Section 2), for each set, we will compute the conditional steady state probability vector, given that the original Markov chain $\mathcal{M}$ is in that set. By applying the state aggregation technique [1], we will aggregate each set into a single state and then compute the steady state probabilities for the aggregated process, i.e., the probabilities of the system being in any given set. Lastly, we will apply the disaggregation technique [1] to compute the individual (unconditional) steady state probabilities of the original Markov process $\mathcal{M}$. These can in turn be used to compute various related performance measures, as already mentioned.

4 System Model

In this section we present the model of a multi-server threshold queueing system with hysteresis which can be defined as follows. There are K servers in the system, where K is unrestricted, each with an exponential service rate μ_i. Customer arrivals are governed by a Poisson process with rate λ. Addition and removal of servers in this queueing system is governed by the forward and the reverse threshold vectors $\boldsymbol{F} = (F_1, F_2, \ldots, F_{K-1})$ and $\boldsymbol{R} = (R_1, R_2, \ldots, R_{K-1})$, where $R_i < F_i$ for all i. Note that, there are multiple ways to create a total order between the F_i's and the R_i's; for clarity and ease of presentation, in the remainder of this paper (unless otherwise stated) we assume that $R_{i+1} < F_i \ \forall i$. However, our solution technique can be easily extended to all other cases as well.

There are several variations of this queueing system that can be considered. In this paper, we consider three such variations, namely: (1) homogeneous-server system, (2) bulk-arrival system, and (3) heterogeneous-server system. Each of the variations of the system can be modeled by a Markov process, of a similar structure. In the following sections we formally describe the Markov processes corresponding to each of the variations; the solution of each of these Markov processes is given in Section 5.

4.1 Homogeneous Servers

Given a K-server *homogeneous* threshold-based queueing system with hysteresis, i.e., $\mu_i = \mu$ for all i, we can construct a corresponding Markov process $\mathcal{M}$ with the following state space $\mathcal{S}$:

$$\mathcal{S} = \{(N, N_s) \mid N \geq 0, N_s \in \{0, 1, 2, \ldots, K\}\}$$

where N is the number of customers in the queueing system and N_s is the number of busy servers. Figure 1 illustrates the state transition diagram for such a system where $K = 3$. Formally, the

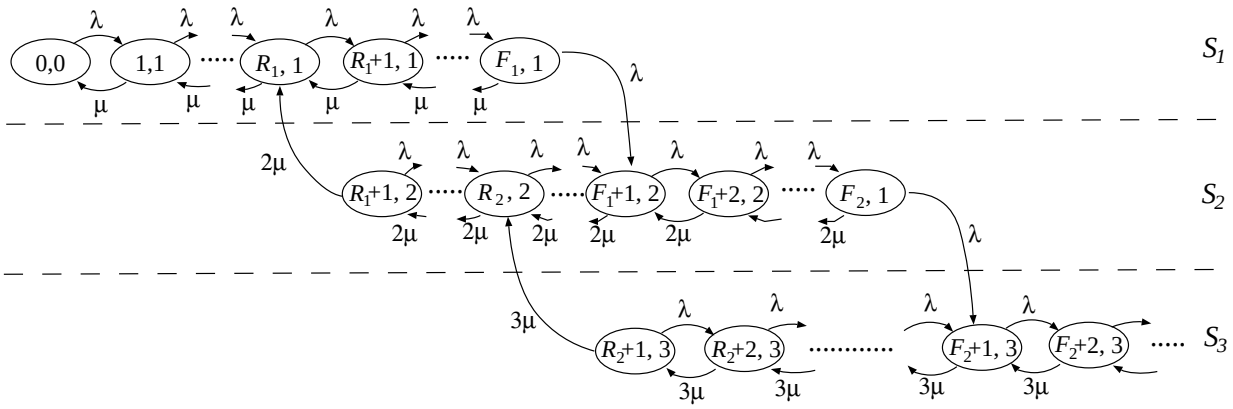

Figure 1: State transition diagram for a three identical server system.

transition structure of $\mathcal{M}$ can be specified as follows:

$$
\begin{array}{lcll}
(0,0) & \longrightarrow & (1,1) & \lambda \\
(i,j) & \longrightarrow & (i+1,j) & \lambda \mathbf{1}\{\ (j \in \{1,\ldots,K\}) \wedge ((i \notin \boldsymbol{F}) \vee ((i = F_z \in \boldsymbol{F}) \wedge (j \neq z)))\ \} \\
(i,j) & \longrightarrow & (i+1,j+1) & \lambda \mathbf{1}\{\ (j \in \{1,\ldots,K-1\}) \wedge (i = F_z \in \boldsymbol{F}) \wedge (j = z)\ \} \\
(i,j) & \longrightarrow & (i-1,j) & j\mu \mathbf{1}\{\ (i \geq 1) \wedge ((i,j) \neq (1,1)) \wedge (j \in \{1,\ldots,K\}) \wedge ((i-1 \notin \boldsymbol{R}) \\
 & & & \quad \vee ((i-1 = R_z \in \boldsymbol{R}) \wedge (j \neq z+1)))\ \} \\
(i,j) & \longrightarrow & (i-1,j-1) & j\mu \mathbf{1}\{\ (j \in \{2,\ldots,K\}) \wedge (i-1 = R_z \in \boldsymbol{R}) \wedge (j = z+1)\ \} \\
(1,1) & \longrightarrow & (0,0) & \mu
\end{array}
\tag{6}
$$

where $\mathbf{1}\{x\}$ is an indicator function that $\mathbf{1}\{x\} = 1$ if condition x is true and 0 condition x is false.

4.2 Bulk Arrivals

In another variation of the threshold-based queueing system with hysteresis each arrival event corresponds to an arrival of multiple customers. This type of a *bulk* arrival process is a generalization of the Poisson arrival process with a single customer, as used in Section 4.1; note that, we do not restrict the bulk arrival size, and (as in the case of Section 4.1) we do not restrict the number of servers in the system. More specifically, the difference from the model considered in Section 4.1 is that each arrival event corresponds to a bulk arrival of size g_i, where:

$$g_i \quad = \quad \text{Prob[arrival of } i \text{ customers]} \qquad i \geq 1 \tag{7}$$

We can construct a corresponding Markov process $\mathcal{M}_b$ with the state space $\mathcal{S}_b$:

$$\mathcal{S}_b = \{(N, N_s) \mid N \geq 0, N_s \in \{0,1,2,\ldots,K\}\}$$

where N is the number of customers in the system and N_s is the number of busy servers. Figure 2 illustrates the state transition diagram for such as system where $K = 3$. Formally, the transition structure of $\mathcal{M}_b$ is defined as follows. Transitions that are due to arrivals have the following structure:

$$
\begin{array}{lcll}
(0,0) & \longrightarrow & (k, \xi(0,1,k)) & \lambda g_k \\
(i,j) & \longrightarrow & (i+k, \xi(i,j,k)) & \lambda g_k \mathbf{1}\{\ j \in \{1,2,\ldots,K\}\ \}
\end{array}
\tag{8}
$$

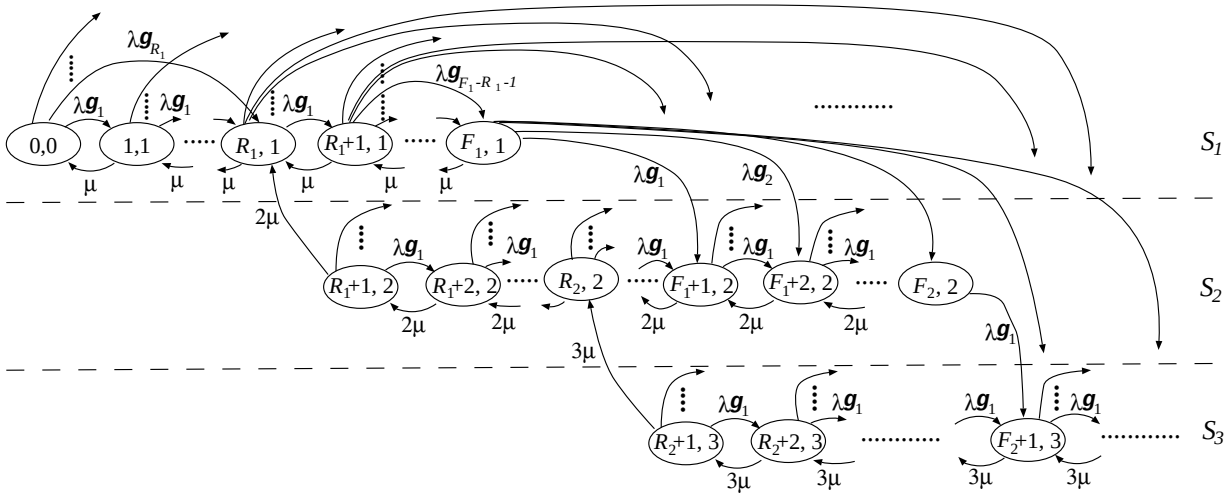

Figure 2: State transition diagram for a three homogeneous servers system with bulk arrivals.

where the mapping function ξ is defined as:

$$\xi(i,j,k) = \begin{cases} j & \text{if } (i+k) \leq F_j \\ \max\{j^* | ((i+k) \leq F_{j^*}) \wedge j^* \geq j \wedge (j^* \in \{j,\ldots,K\})\} & \text{otherwise} \end{cases}$$

Transitions that are due to departures have the following structure:

$$\begin{array}{lcll} (i,j) & \longrightarrow & (i-1,j) & j\mu \mathbf{1}\{\ (i \geq 1) \wedge ((i,j) \neq (1,1)) \wedge (j \in \{1,2,\ldots,K\}) \wedge ((i-1 \notin \boldsymbol{R}) \\ & & & \quad \vee ((i-1 = R_z \in \boldsymbol{R}) \wedge (j \neq z+1)))\ \} \\ (i,j) & \longrightarrow & (i-1,j-1) & j\mu \mathbf{1}\{\ (j \in \{2,\ldots,K\}) \wedge (i-1 = R_z \in \boldsymbol{R}) \wedge (j = z+1)\ \} \\ (1,1) & \longrightarrow & (0,0) & \mu \end{array} \tag{9}$$

4.3 Heterogeneous Servers

Finally, a third variation on the problem, is the case with *heterogeneous* servers. More specifically, the customer arrival process is still Poisson with rate λ and again restricted to a single customer per arrival, but the K servers are heterogeneous, each with an exponential service rate of $\mu_i, 1 \leq i \leq K$. We make no restrictions on the number of servers or the relative values of service rates μ_i and μ_j, where $i \neq j$ and $1 \leq i \leq K, 1 \leq j \leq K$. Since in the case of heterogeneous servers, the servers are distinct, we must also make the following modifications to the rules which govern addition and removal of servers, based on the values of the threshold vectors $\boldsymbol{F}$ and $\boldsymbol{R}$:

- when an arrival occurs to a system with F_i customers, server $i+1$, with a service rate of μ_{i+1} is added to the system (as opposed to an "arbitrary" server), i.e., servers are added to the system in *ascending* order, where server i is added before server j if $i < j$
- when a departure, corresponding to service completion *at server* $i+1$, occurs, leaving behind a system with j customers where $j \leq R_i$, server $i+1$ is removed from the system

For this threshold-based queueing system with hysteresis and heterogeneous servers we can construct a Markov process $\mathcal{M}_h$ with the following state space $\mathcal{S}_h$:

$$\mathcal{S}_h = \{(N, \boldsymbol{N}_s) \mid N \geq 0, \boldsymbol{N}_s \in \{0,1\}^K\}$$

where N is the number of customers in the queueing system and $\boldsymbol{N}_s$ is a string of K bits indicating busy and idle servers, i.e., $\boldsymbol{N}_s = N_s^1 N_s^2 \cdots N_s^K$, where

$$N_s^k = \begin{cases} 1 & \text{if server } k \text{ is busy} \\ 0 & \text{if server } k \text{ is idle} \end{cases}$$

Figure 3 illustrates a Markov process corresponding to such a system where $K = 3$.

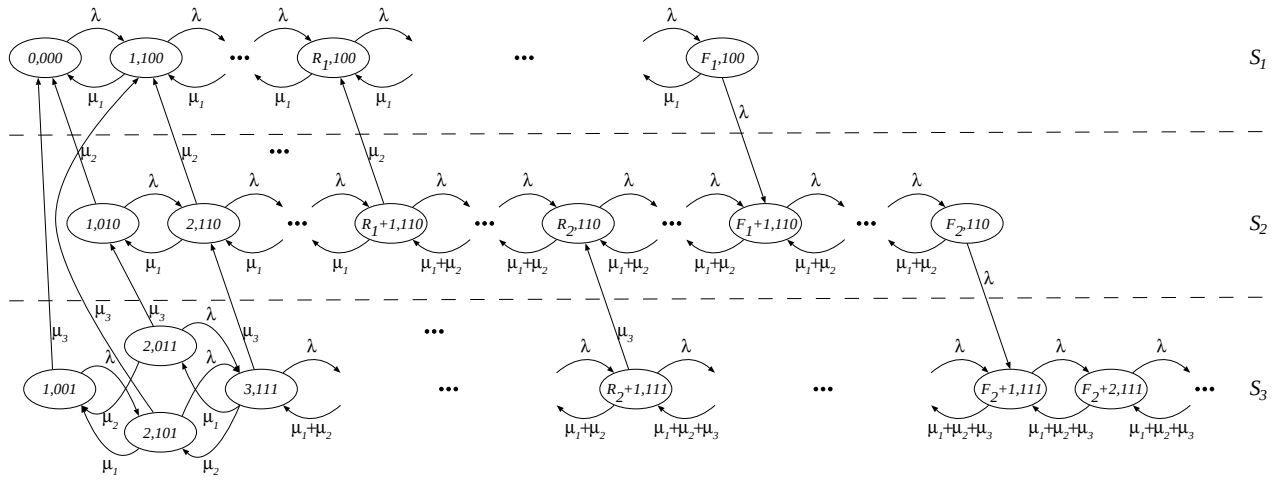

Figure 3: State transition diagram for a heterogeneous servers system with $K = 3$.

Before formally defining the transition structure of $\mathcal{M}_h$, let us define the following notation. Let $\boldsymbol{j}^k$, $\boldsymbol{j} \in \{0,1\}^K$, represent a string of K bits with the k^{th} bit equal to 1, i.e., $\boldsymbol{j}^k = \{0,1\}^{(k-1)}\{1\}\{0,1\}^{(K-k)}$. Let $\boldsymbol{j}(k)$, $\boldsymbol{j} \in \{0,1\}^K$, represent a string of K bits with the first k bits equal to 1, i.e., $\boldsymbol{j}(k) = \{1\}^k\{0,1\}^{(K-k)}$. Let $G_+^n(\boldsymbol{j})$, $1 \leq n \leq K$, be a function which, given $\boldsymbol{j}$, returns a new string $\boldsymbol{j}'$ which has all bits identical to those of $\boldsymbol{j}$, except for the n^{th} bit, which is equal to 1. Let $G_-^n(\boldsymbol{j})$, $1 \leq n \leq K$, be a function which, given $\boldsymbol{j}$, returns a new string $\boldsymbol{j}'$ which has all bits identical to those of $\boldsymbol{j}$, except for the n^{th} bit, which is equal to 0. Then, formally, the transition structure of $\mathcal{M}_h$ can be specified as

follows:

$$
\begin{array}{lcll}
(0,\{0\}^K) & \longrightarrow & (1,1\{0\}^{(K-1)}) & \lambda \\
(i,\boldsymbol{j}(k)) & \longrightarrow & (i+1,\boldsymbol{j}(k)) & \lambda\mathbf{1}\{\ (i\geq k)\wedge(\boldsymbol{j}\in\{1\}^k\{0\}^{(K-k)})\wedge((i\notin\boldsymbol{F}) \\
 & & & \quad\vee((i=F_z\in\boldsymbol{F})\wedge(k\neq z)))\ \} \\
(i,\boldsymbol{j}(k)) & \longrightarrow & (i+1,\boldsymbol{j}(k+1)) & \lambda\mathbf{1}\{\ (i\geq k)\wedge(\boldsymbol{j}\in\{1\}^k\{0\}^{(K-k)}) \\
 & & & \quad\wedge(k<K)\wedge(i=F_z\in\boldsymbol{F})\wedge(k=z)\ \} \\
(i,\boldsymbol{j}(k)) & \longrightarrow & (i-1,\boldsymbol{j}(k)) & \sum_{n=1}^{k}\mu_n\mathbf{1}\{\ (i\geq k+1)\wedge((i,\boldsymbol{j}(k))\neq(1,\{1\}\{0\}^{K-1})) \\
 & & & \quad\wedge(\boldsymbol{j}\in\{1\}^k\{0\}^{(K-k)}) \\
 & & & \quad\wedge((i-1\notin\boldsymbol{R})\vee((i-1=R_z\in\boldsymbol{R})\wedge(k\neq z+1)))\ \} \\
(i,\boldsymbol{j}(k)) & \longrightarrow & (i-1,\boldsymbol{j}(k-1)) & \mu_k\mathbf{1}\{\ (i\geq k+1)\wedge(\boldsymbol{j}\in\{1\}^k\{0\}^{(K-k)}) \\
 & & & \quad\wedge(i-1=R_z\in\boldsymbol{R})\wedge(k=z+1)\ \} \\
(i,\boldsymbol{j}^k) & \longrightarrow & (i-1,G_-^n(\boldsymbol{j}^k)) & \mu_n\mathbf{1}\{\ (n<k)\wedge(1<i\leq k)\wedge(\boldsymbol{j}\in\{0,1\}^K)\ \} \\
(i,\boldsymbol{j}^k) & \longrightarrow & (i+1,G_+^n(\boldsymbol{j}^k)) & \lambda\mathbf{1}\{\ (n<k)\wedge(1\leq i<k) \\
 & & & \quad\wedge((n=1)\vee(\boldsymbol{j}\in\{1\}^{(n-1)}\{0\}\{0,1\}^{(K-n)}))\ \} \\
(1,\{1\}\{0\}^{K-1}) & \longrightarrow & (0,\{0\}^K) & \mu_1
\end{array}
\tag{10}
$$

Note that in [8], the authors describe a solution for a system with $K = 2$ heterogeneous servers; however, they experience difficulties in extending the Green's function method to the general case of $K > 2$. In Section 5, we present a solution for the general case heterogeneous servers system using the approach of stochastic complementation, as in the other two cases.

5 Analysis

In this section we present the details of the basic analysis approach outlined in Section 3. We first illustrate this technique using the simpler case, of homogeneous servers, and then show how it can be extended (fairly simply) to the other two cases, namely, the bulk arrivals and the heterogeneous servers cases.

5.1 Homogeneous Servers

The goal of this section is to compute the steady state probabilities $\boldsymbol{\pi}(\boldsymbol{n})$ for all $\boldsymbol{n} \in \mathcal{S}$, where $\mathcal{S}$ is the state space of the Markov process $\mathcal{M}$ (see Section 4.1). As outlined in Section 3, the first step is to partition the state space. Specifically, given the original Markov process $\mathcal{M}$, let us partition the state space $\mathcal{S}$ into K disjoint sets $\mathcal{S}_l$, where:

$$\mathcal{S}_l = \{(i,j) \mid (i,j) \in \mathcal{S} \text{ and } j = l\} \qquad l = 1, 2, \ldots, K$$

We can view partition $\mathcal{S}_l$ as representing all states corresponding to exactly l busy servers[2]. For $2 \leq l \leq K-1$, we can order the states in $\mathcal{S}_l$ as follows:

$$\{(R_{l-1}+1, l), \ldots, (R_l, l), \ldots, (F_{l-1}+1, l), \ldots, (F_l, l)\}$$

[2]To simplify notation, we assume that state $(0,0)$ is also in $\mathcal{S}_1$.

Let us define another Markov processes $\mathcal{M}_l$, for $l \in \{2, \ldots, K-1\}$, such that the state space of $\mathcal{M}_l$ corresponds to the states in $\mathcal{S}_l$. The transition structure of $\mathcal{M}_l$ is similar to the transition structure of $\mathcal{M}$ for the states in $\mathcal{S}_l$, except for the following modifications: (rule 1) a transition from $(R_{l-1}+1, l)$ to $(R_{l-1}, l-1)$ in the original process $\mathcal{M}$ is replaced by a transition from $(R_{l-1}+1, l)$ to $(F_{l-1}+1, l)$ in $\mathcal{M}_l$ and (rule 2) a transition from (F_l, l) to $(F_l+1, l+1)$ in the original process $\mathcal{M}$, is replaced by a transition from (F_l, l) to (R_l, l) in $\mathcal{M}_l$. Figure 4 illustrates the state transition diagram for $\mathcal{M}_l$, for $l \in \{2, \ldots, K-1\}$. Similarly, for $l = 1$, we can order the states in $\mathcal{S}_1$ as follows:

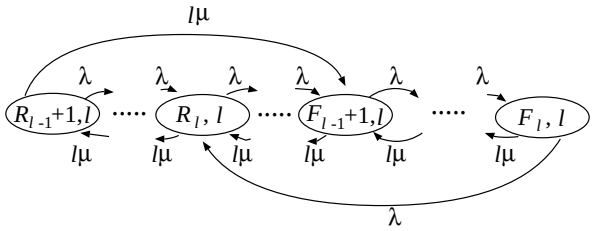

Figure 4: State transition diagram for $\mathcal{M}_l$.

$$\{(0,0), \ldots, (R_1, 1), \ldots, (F_1, 1)\}$$

and then define the Markov process $\mathcal{M}_1$ such that the state space of $\mathcal{M}_1$ corresponds to the states in $\mathcal{S}_1$. The transition structure of $\mathcal{M}_1$ is similar to that of $\mathcal{M}$ for the states in $\mathcal{S}_1$, except that a transition from $(F_1, 1)$ to $(F_1+1, 2)$ in $\mathcal{M}$ is replaced by a transition from $(F_1, 1)$ to $(R_1, 1)$ in $\mathcal{M}_1$. That is, for the $l = 1$ case, (rule 1) above simply does not apply. Finally, for $l = K$, we can order the states in $\mathcal{S}_K$ as follows:

$$\{(R_{K-1}+1, K), \ldots, (F_{l-1}+1, K), \ldots\}$$

and then define the Markov process $\mathcal{M}_K$ such that the state space of $\mathcal{M}_K$ corresponds to the states in $\mathcal{S}_K$. The transition structure of $\mathcal{M}_K$ is similar to that of $\mathcal{M}$ for the states in $\mathcal{S}_K$, except that a transition from $(R_{K-1}+1, K)$ to $(R_{K-1}, K-1)$ in $\mathcal{M}$ is replaced by a transition from $(R_{K-1}+1, K)$ to $(F_{K-1}+1, K)$ in $\mathcal{M}_K$. That is, for the $l = K$ case, (rule 2) above simply does not apply.

Theorem 4 *The steady state probabilities solution of the Markov process $\mathcal{M}_l$ is the conditional steady state probabilities solution for the states in $\mathcal{S}_l$ of the original Markov process $\mathcal{M}$, given that the system is in partition $\mathcal{S}_l$.*

Proof: For the Markov processes $\mathcal{M}_1$ and $\mathcal{M}_K$, this follows from a simple application of Theorem 3. For the Markov processes $\mathcal{M}_l$, where $2 \leq l \leq K-1$, let us define the following:

$$\mathcal{S}_l^- = \bigcup_{i=1}^{l-1} \mathcal{S}_i \;\; ; \;\; \mathcal{S}_l \;\; ; \;\; \mathcal{S}_l^+ = \bigcup_{i=l+1}^{K} \mathcal{S}_i$$

Since there is a single return from $\mathcal{S}_l^+$ to $\{\mathcal{S}_l^- \cup \mathcal{S}_l\}$, using Theorem 3, we can obtain the conditional steady state probabilities for the states in $\{\mathcal{S}_l^- \cup \mathcal{S}_l\}$, given that the process is in $\{\mathcal{S}_l^- \cup \mathcal{S}_l\}$. Since there is a single entry from $\mathcal{S}_l^-$ to $\mathcal{S}_l$, using Theorem 3, we can obtain the conditional steady state

probabilities for states in $\mathcal{S}_l$, given that the process is in $\mathcal{S}_l$. ∎

In the following section, we show how to compute the steady state probability vector for each of the Markov processes $\mathcal{M}_l$, where $l \in \{1, 2, \ldots, K\}$.

5.1.1 Analysis of $\mathcal{M}_l$

As outlined in Section 3, the next step is to derive the steady state probability vector for the states in $\mathcal{M}_l$ where $l \in \{1, \ldots, K\}$, namely $\boldsymbol{\pi}_{\mathcal{M}_l}(\boldsymbol{n})$. Since all states in $\mathcal{M}_l$ represent l busy servers, for ease of presentation, we can ignore that portion of the state description, i.e., we can identify states in $\mathcal{M}_l$ based on the number of customers. Figure 4 illustrates the state transition diagram of $\mathcal{M}_l$, where $l \in \{2, \ldots, K-1\}$; let us begin with the analysis of these Markov processes. Based on the flow balance equations for all states i, where $R_{l-1} + 1 \leq i \leq R_l$, we can define the coefficient terms C_i^l such that:

$$\begin{aligned} \pi_l(i) &= \pi_l(R_{l-1}+1)C_i^l \qquad \text{where} \\ C_i^l &= \sum_{j=0}^{i-R_{l-1}-1} \left(\frac{\lambda}{l\mu}\right)^j = \frac{l}{l-\rho}\left[1 - \left(\frac{\rho}{l}\right)^{i-R_{l-1}}\right] \qquad i = R_{l-1}+1, \ldots, R_l \end{aligned} \tag{11}$$

and $\rho = \lambda/\mu$. (Note that above we assume that $\frac{\lambda}{l\mu} \neq 1$; a similar derivation can be given for $\frac{\lambda}{l\mu} = 1$, which we omit for clarity of presentation.)

If we consider the flow balance equations for all states i, where $R_l + 1 \leq i \leq F_{l-1} + 1$, we can express their state probabilities $\pi_l(i)$ in term of $\pi_l(R_{l-1}+1)$ and $\pi_l(F_l)$ as:

$$\pi_l(i) = \pi_l(R_{l-1}+1)\left[\sum_{j=0}^{i-R_{l-1}-1}\left(\frac{\lambda}{l\mu}\right)^j\right] - \pi_l(F_l)\left[\sum_{j=1}^{i-R_l}\left(\frac{\lambda}{l\mu}\right)^j\right] \tag{12}$$

Similarly, the flow balance equations for all states i, where $F_{l-1} + 2 \leq i \leq F_l - 1$ are:

$$\pi_l(i) = \pi_l(R_{l-1}+1)\left[\sum_{j=i-F_{l-1}-1}^{i-R_{l-1}-1}\left(\frac{\lambda}{l\mu}\right)^j\right] - \pi_l(F_l)\left[\sum_{j=1}^{i-R_l}\left(\frac{\lambda}{l\mu}\right)^j\right] \tag{13}$$

Lastly, the flow balance equation for $i = F_l$ is:

$$\pi_l(F_l - 1)\lambda = \pi_l(F_l)(\lambda + l\mu) \tag{14}$$

Now, observe that based on Equations (11), (12), (13) and (14), we can express $\pi_l(F_l)$ in terms of $\pi_l(R_{l-1}+1)$. After simplifying the necessary expressions, we have:

$$\begin{aligned} \pi_l(F_l) &= \pi_l(R_{l-1}+1)C_{F_l}^l \qquad \text{where} \\ C_{F_l}^l &= \left[1 + \frac{l\mu}{\lambda} + \sum_{j=1}^{F_l-1-R_l}\left(\frac{\lambda}{l\mu}\right)^j\right]^{-1}\left[\sum_{j=F_l-F_{l-1}}^{F_l-1-(R_{l-1}+1)}\left(\frac{\lambda}{l\mu}\right)^j\right] \\ &= \frac{\rho}{l + \rho\,(\rho/l)^{F_l-R_l}}\left[\left(\frac{\rho}{l}\right)^{F_l-F_{l-1}} - \left(\frac{\rho}{l}\right)^{F_l-R_{l-1}+1}\right] \end{aligned} \tag{15}$$

Now that $\pi_l(F_l)$ depends only on $\pi_l(R_{l-1}+1)$, we can substitute the expression for $\pi_l(F_l)$ back into Equations (12) and (13) and find the corresponding coefficients C_i^l for $R_l+1 \le i \le F_l-1$; then,

$$\begin{aligned}
\pi_l(i) &= \pi_l(R_{l-1}+1)C_i^l \qquad \text{where} \\
C_i^l &= \begin{cases} \sum_{j=0}^{i-R_{l-1}-1}\left(\frac{\lambda}{l\mu}\right)^j - C_{F_l}^l \sum_{j=1}^{i-R_l}\left(\frac{\lambda}{l\mu}\right)^j & \text{for } R_l+1 \le i \le F_{l-1}+1 \\ \sum_{j=i-F_{l-1}-1}^{i-R_{l-1}-1}\left(\frac{\lambda}{l\mu}\right)^j - C_{F_l}^l \sum_{j=1}^{i-R_l}\left(\frac{\lambda}{l\mu}\right)^j & \text{for } F_{l-1}+2 \le i \le F_l-1 \end{cases}
\end{aligned}$$

After further simplifications, we have:

$$C_i^l = \begin{cases} \frac{l}{l-\rho}\left[1-\left(\frac{\rho}{l}\right)^{i-R_{l-1}} - \frac{\rho\, C_{F_l}^l}{l}\left[1-\left(\frac{\rho}{l}\right)^{i-R_l}\right]\right] & \text{for } R_l+1 \le i \le F_{l-1}+1 \\ \frac{l}{l-\rho}\left[\left(\frac{\rho}{l}\right)^{i-F_l-1} - \left(\frac{\rho}{l}\right)^{i-R_{l-1}} - \frac{\rho C_{F_l}^l}{l}\left[1-\left(\frac{\rho}{l}\right)^{i-R_l+1}\right]\right] & \text{for } F_{l-1}+2 \le i \le F_l-1 \end{cases} \tag{16}$$

With all coefficient C_i^l defined in Equations (11), (15), and (16), we can determine $\pi_l(R_{l-1}+1)$ through normalization, that is, the sum of all the steady state probabilities in $\mathcal{M}_l$ has to be equal to 1:

$$\pi_l(R_{l-1}+1) = \left[\sum_{i=R_{l-1}+1}^{F_l} C_i^l\right]^{-1} \tag{17}$$

For the Markov process $\mathcal{M}_1$, we can use a similar approach to derive the steady state probabilities. They are:

$$\pi_1(0) = \left[\frac{1-\rho^{R_1+1}}{1-\rho} + \frac{\rho^{F_1+1}}{\rho^{F_1-R_1+1}-1}\left(F_1 - R_1 - \frac{\rho^{F_1-R_1}-1}{[\rho-1]\,[\rho^{F_1-R_1}]}\right)\right]^{-1} \tag{18}$$

$$\pi_1(j) = \pi_1(0)\rho^j \qquad j = 1,2,\ldots,R_1 \tag{19}$$

$$\pi_1(j) = \pi_1(0)\rho^j\left(\frac{\rho^{F_1-j+1}-1}{\rho^{F_1-R_1+1}-1}\right) \qquad j = R_1+1,\ldots,F_1 \tag{20}$$

where $\rho = \lambda/\mu$. Finally, the steady state probabilities for the Markov process $\mathcal{M}_K$ are:

$$\pi_K(R_{K-1}+1) = \frac{K-\rho}{K(F_{K-1}-R_{K-1}+1)} \tag{21}$$

$$\pi_K(j) = \frac{1}{F_{K-1}-R_{K-1}+1}\left[1-\left(\frac{\rho}{K}\right)^{j-R_{K-1}}\right] \qquad j=R_{K-1}+2,\ldots,F_{K-1}+1 \tag{22}$$

$$\pi_K(j) = \frac{1}{F_{K-1}-R_{K-1}+1}\left[\left(\frac{\rho}{K}\right)^{j-F_{K-1}-1} - \left(\frac{\rho}{K}\right)^{j-R_{K-1}}\right] \qquad j>F_{K-1}+1 \tag{23}$$

5.1.2 Analysis of the Aggregated Process

Once we have obtained an expression for the steady state probability vector of each $\mathcal{M}_l$, which is also the conditional state probability vector of $\mathcal{M}$, given that the system is in $\mathcal{S}_l$, the only remaining step (as outlined in Section 3) is to find the aggregate state probability of the system being in $\mathcal{S}_l$.

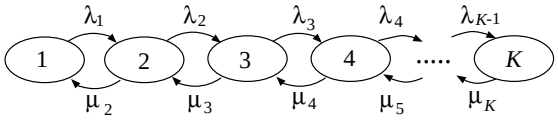

Figure 5: State transition diagram for aggregated process.

Therefore, for each l, $1 \leq l \leq K$ let us aggregate all the states in $\mathcal{S}_l$ into a single state. The transition state diagram of the resulting aggregated process is illustrated in Figure 5. The transition rates of the aggregated process can be computed as follows:

$$\lambda_i = \lambda \pi_i(F_i) \qquad i = 1, 2, \ldots, K-1 \tag{24}$$

$$\mu_i = i\mu \pi_i(R_{i-1}+1) \qquad i = 2, 3, \ldots, K \tag{25}$$

where $\pi_i(F_i)$ and $\pi_i(R_{i-1}+1)$ are the conditional state probabilities obtained in Section 5.1. The steady state probabilities of this aggregated process are as follows [12]:

$$\pi(1) = \left[1 + \sum_{k=2}^{K} \prod_{j=1}^{k-1} \left(\frac{\lambda_j}{\mu_{j+1}} \right) \right]^{-1} \tag{26}$$

$$\pi(i) = \left[1 + \sum_{k=2}^{K} \prod_{j=1}^{k-1} \left(\frac{\lambda_j}{\mu_{j+1}} \right) \right]^{-1} \prod_{j=1}^{i-1} \left(\frac{\lambda_j}{\mu_{j+1}} \right) \qquad i = 2, 3, \ldots, K \tag{27}$$

5.1.3 Performance Measures

At this point we have all the necessary information to compute the steady probabilities for $\mathcal{M}$. That is, once we determine, for each l: 1) the conditional state probabilities of all states in $\mathcal{S}_l$, given that the system is in $\mathcal{S}_l$ and 2) the steady state probability of being in state l of the aggregated process, then the steady state probability of each individual state (i, j) in $\mathcal{M}$ can be expressed as:

$$\pi(i,j) = \pi_j(i)\pi(j) \qquad \text{where } (i,j) \in \mathcal{S}_j \tag{28}$$

Then (as outlined in Section 3) we can compute various performance measures; more specifically, we can compute many performance measures which can be expressed in the form of a Markov reward function, $\mathcal{R}$, where $\mathcal{R} = \sum_{i,j} \pi(i,j) R(i,j)$ and $R(i,j)$ is the reward for state (i,j). Two useful performance measures for our system are the expected number of customers and the expected response time. Below, we illustrate how easy it is to obtain such performance measures, once we have the steady state probabilities; for instance, the expected number of customers can be expressed as a Markov reward functions, where $R(i,j) = i$.

Let N and T denote the expected number of customers and the expected response time, respectively, of the original threshold-based queueing system with hysteresis, corresponding to the Markov process $\mathcal{M}$. Then N can be expressed as:

$$N = \sum_{i=1}^{F_1} i\pi_1(i)\pi(1) + \sum_{j=2}^{K-1} \sum_{i=R_{j-1}+1}^{F_j} i\pi_j(i)\pi(j) + \sum_{i=R_{K-1}+1}^{\infty} i\pi_K(i)\pi(K) \tag{29}$$

Using Little's result [15], we can express T as:

$$T = \frac{1}{\lambda}\left[\sum_{i=1}^{F_1} i\pi_1(i)\pi(1) + \sum_{j=2}^{K-1}\sum_{i=R_{j-1}+1}^{F_j} i\pi_j(i)\pi(j) + \sum_{i=R_{K-1}+1}^{\infty} i\pi_K(i)\pi(K)\right] \quad (30)$$

Remarks on Complexity of Solution: it would be useful at this point to briefly discuss the complexity of computing N, where we consider the number of multiplications required by a computation as a measure of time complexity. The major contributors to the time complexity of computing N are: (a) computation of the aggregate state probabilities and (b) evaluation of the summations in Equation (29). The complexity of computing the aggregate state probabilities is $O(K)$. The complexity of evaluating the finite summations in Equation (29), each corresponding to a partition $\mathcal{S}_l$, is $O(F_l - R_{l-1})$ for each $2 \leq l \leq K-1$ and $O(F_1)$ for $l = 1$. What remains is the complexity of evaluating the infinite summation, which may not be apparent directly from Equation (29). Using Equation (23), we can evaluate the tail of the infinite summation in Equation (29) to be (assuming that $\frac{\rho}{K} \neq 1$):

$$\pi(K)\left[\frac{(\frac{\rho}{K})^{-1} - (\frac{\rho}{K})^{(F_{K-1}-R_{K-1})}}{F_{K-1} - R_{K-1} + 1}\right]\left[\frac{F_{K-1}+1}{1-\frac{\rho}{K}} + \frac{1}{(1-\frac{\rho}{K})^2}\right]$$

which requires $O(F_{K-1} - R_{K-1})$ multiplications to compute. The remainder of the infinite summation, which can be computed using Equations (22) and (23), requires $O(F_{K-1} - R_{K-1})$ multiplications. Thus the total time complexity of evaluating N is

$$O(\max(K, (F_1 + 2(F_{K-1} - R_{K-1}) + \sum_{l=2}^{K-1}(F_l - R_{l-1}))))$$

and the corresponding space complexity is $O(1)$. Note that, one advantage of the homogeneous case solution is that the different partitions can be solved *in parallel*, i.e., the construction of stochastic complements for all partitions and their solution can proceed in parallel.

5.2 Bulk Arrivals

Although, in general, there may not exist a closed-form solution for the steady state probabilities of a threshold-based queueing system with hysteresis and *bulk* arrivals, we can still devise an efficient algorithm for computing the steady state probability vector $\boldsymbol{\pi}(\boldsymbol{n})$ (where $\boldsymbol{n} \in \mathcal{S}_b$ in the original Markov process $\mathcal{M}_b$) as well as the expected number of customers, N_b, and the expected response time of a customer, T_b. This can be accomplished using the approach outlined in Section 3, similarly to the procedure used in Section 5.1.

As in the case of homogeneous servers, we first partition the state space $\mathcal{S}_b$ of the Markov process $\mathcal{M}_b$ into K disjoint sets, S_l, where:

$$\mathcal{S}_l = \{(i,j) \mid (i,j) \in \mathcal{S}_b \text{ and } j = l\} \qquad l = 1, 2, \ldots, K$$

and, as before, $\mathcal{S}_l$ represents all the states with exactly l busy servers[3]. Also, we define:

$$\mathcal{S}_i^* = \{\mathcal{S}_{i+1} \cup \mathcal{S}_{i+2} \cup \cdots \cup \mathcal{S}_K\} \qquad \text{for } i = 1, 2, \ldots, K-1$$

[3]Recall that, to simplify notation, we can assume that state $(0,0)$ is in $\mathcal{S}_1$.

Using Theorem 3, we can easily compute the conditional steady state probabilities for states in $\mathcal{S}_1$, given that $\mathcal{M}_b$ is in $\mathcal{S}_1$. This is accomplished by constructing a Markov process $\mathcal{M}_1$ which has the state space $\mathcal{S}_1$, with all transitions being the same as those (corresponding to states in $\mathcal{S}_1$) in the original Markov process $\mathcal{M}_b$, except that any transition from a state in $\mathcal{S}_1$ to a state in $\mathcal{S}_j$ (where $1 < j \leq K$) becomes a transition to $(R_1, 1) \in \mathcal{S}_1$. In general, there does not exist a closed-form solution for $\mathcal{M}_1$; however, since the state space of $\mathcal{M}_1$ is usually small and finite, we can easily obtain the steady state probability vector using any chosen solution technique, as described in [21]. Let us denote the steady state probability vector of $\mathcal{M}_1$ by $\boldsymbol{\pi}_{\mathcal{M}_1}$.

Other than for $\mathcal{S}_1$, it appears to be difficult to apply Theorem 3 directly to other sets $\mathcal{S}_l$, $2 \leq l \leq K$, in $\mathcal{M}_b$ since there are *multiple ways* of entering $\mathcal{S}_j$ from $\mathcal{S}_i$ for $j > i$. To solve this problem, let us take advantage of stochastic complementation once again. Since we are able to compute the steady state probability vector $\boldsymbol{\pi}_{\mathcal{M}_1}$, which is also the conditional steady state probability vector (for the states in $\mathcal{S}_1$) of $\mathcal{M}_b$, given than the system is in $\mathcal{S}_1$, we can easily construct the stochastic complement for states in $\mathcal{S}_1^* = \{\mathcal{S}_2 \cup \mathcal{S}_3 \cdots \cup \mathcal{S}_K\}$. A probabilistic interpretation of this approach is that we are *redistributing* the transition rates[4] from states in $\mathcal{S}_1^*$ to $\mathcal{S}_1$ (which exist in the original Markov process) back to states in $\mathcal{S}_1^*$. This redistribution should be proportional to the relative visit ratios at which $\mathcal{S}_1^*$ is entered from $\mathcal{S}_1$. These relative visit ratios are known since we have an efficient procedure for computing $\boldsymbol{\pi}_{\mathcal{M}_1}$. Thus, the relative rates back to $\mathcal{S}_1^*$ are:

$$\boldsymbol{f}_1 \quad = \quad \left[\boldsymbol{\pi}_{\mathcal{M}_1} \boldsymbol{Q}_{\mathcal{S}_1,\mathcal{S}_1^*} \boldsymbol{e}\right]^{-1} \boldsymbol{\pi}_{\mathcal{M}_1} \boldsymbol{Q}_{\mathcal{S}_1,\mathcal{S}_1^*} \tag{31}$$

where $\boldsymbol{Q}_{\mathcal{S}_1,\mathcal{S}_1^*}$ is the transition rate matrix from $\mathcal{S}_1$ to $\mathcal{S}_1^*$ and $\boldsymbol{f}_1$ denotes the row vector of relative visit ratios to $\mathcal{S}_1^*$. It is not difficult to observe that $\boldsymbol{f}_1 \boldsymbol{e} = 1$. The validity of this claim is reflected in the following theorem.

Theorem 5 *The steady state probability vector for the Markov process $\mathcal{M}_{\mathcal{S}_1^*}$ is the conditional steady state probability vector for the states in $\mathcal{S}_1^*$ of the original Markov process $\mathcal{M}_b$, given that the system is in partition $\mathcal{S}_1^*$.*

Proof: Let us rearrange the states of the original process such that the transitional rate matrix of $\mathcal{M}_b$ is:

$$\begin{bmatrix} \boldsymbol{Q}_{\mathcal{S}_1^*,\mathcal{S}_1^*} & \boldsymbol{Q}_{\mathcal{S}_1^*,\mathcal{S}_1} \\ \boldsymbol{Q}_{\mathcal{S}_1,\mathcal{S}_1^*} & \boldsymbol{Q}_{\mathcal{S}_1,\mathcal{S}_1} \end{bmatrix}$$

where $\boldsymbol{Q}_{i,j}$ is the transition rate sub-matrix corresponding to transitions from partition i to partition j. Note that the sub-matrix $\boldsymbol{Q}_{\mathcal{S}_1^*,\mathcal{S}_1}$ has only a single non-zero row which has only a single non-zero entry. This row, call it row i of $\boldsymbol{Q}_{\mathcal{S}_1^*,\mathcal{S}_1}$, corresponds to state $(R_1 + 1, 2)$ in $\mathcal{S}_1^*$ and the non-zero entry has the value of 2μ. Referring to the form of a stochastic complement given in Equation (5), $r_i = 2\mu$ and $r_j = 0$ for $j \neq i$. Thus, we only need to construct $\boldsymbol{z}_i$, a vector which determines how r_i is redistributed between the states in $\mathcal{S}_1^*$. This vector $\boldsymbol{z}_i$ is determined by the conditional steady state probabilities of states in $\mathcal{S}_1$, i.e., $\boldsymbol{\pi}_{\mathcal{M}_1}$, and the transitional matrix $\boldsymbol{Q}_{\mathcal{S}_1,\mathcal{S}_1^*}$. Therefore, the redistribution is governed precisely by the vector $\boldsymbol{f}_1$, as specified in Equation (31). ∎

[4] In this case, the transition rate in question is 2μ.

Thus, we have constructed a stochastic complement for the states in $\mathcal{S}_1^*$. The Markov process $\mathcal{M}_{\mathcal{S}_1^*}$, corresponding to the example of Figure 2, is illustrated in Figure 6. Note that this newly

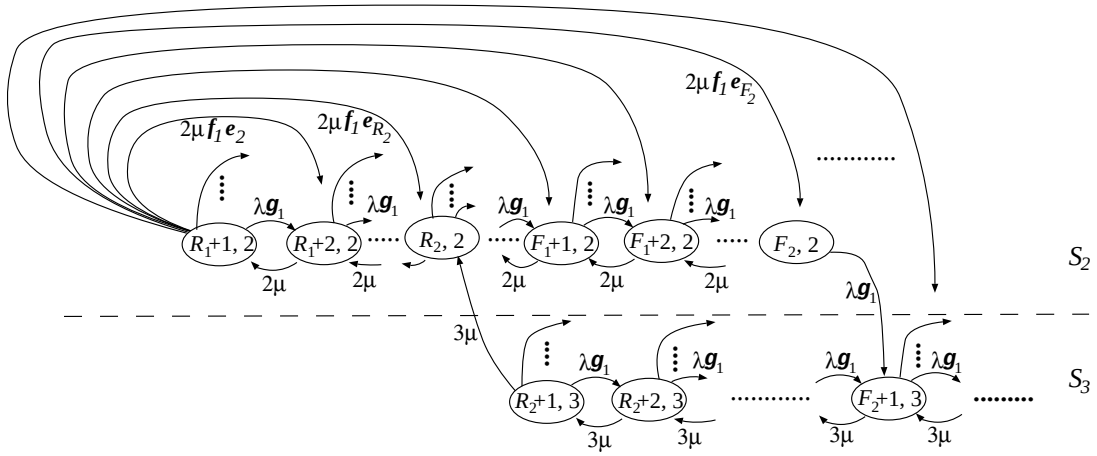

Figure 6: State transition diagram after elimination of $\mathcal{S}_1$.

derived Markov process has a structure similar to that of the original Markov process $\mathcal{M}_b$. Namely, there is a single entry from $\mathcal{S}_3$ to $\mathcal{S}_2$. Therefore, using a similar argument (to the one used in deriving $\boldsymbol{\pi}_{\mathcal{M}_1}$), we can easily compute the conditional steady state probabilities for the states in $\mathcal{S}_2$, given that $\mathcal{M}_b$ is in $\mathcal{S}_2$. The conditional steady state probabilities for $\mathcal{S}_2$ in this newly derived process $\mathcal{M}_{\mathcal{S}_1^*}$ are clearly identical to the conditional steady state probabilities of the original Markov process $\mathcal{M}_b$, given that the system is in $\mathcal{S}_2$. At this point, we are in a position to construct $\mathcal{M}_{\mathcal{S}_2^*}$, using an argument similar to that of Theorem 5. Continuing in this manner, we can recursively solve for all the conditional steady state probabilities for states in $\mathcal{S}_1$ through $\mathcal{S}_{K-1}$; we denote these by $\boldsymbol{\pi}_{\mathcal{M}_j}$, for $j = 1, \ldots, K-1$.

To solve for $\boldsymbol{\pi}_{\mathcal{M}_K}$, which is a vector of conditional steady state probabilities for states in $\mathcal{S}_K$, given that $\mathcal{M}_b$ is in $\mathcal{S}_K$, we cannot simply apply the above stated approach. The reason being that the state space $\mathcal{S}_K$ is infinite, since we have not restricted the queueing capacity of our system. However, we can express $\boldsymbol{\pi}_{\mathcal{M}_K}$ via a $Z-$transform. Let us define $\boldsymbol{f}_{K-1}$ to be the relative ratios back to $\mathcal{S}_K$ for the derived Markov process $\mathcal{M}_{\mathcal{S}_{K-1}^*}$; these relative visit ratios are:

$$\boldsymbol{f}_{K-1} \;=\; \left[\boldsymbol{\pi}_{\mathcal{M}_{K-1}} \boldsymbol{Q}_{\mathcal{S}_{K-1},\mathcal{S}_K} \boldsymbol{e}\right]^{-1} \boldsymbol{\pi}_{\mathcal{M}_{K-1}} \boldsymbol{Q}_{\mathcal{S}_{K-1},\mathcal{S}_K} \;=\; [f_0, f_1, f_2, \ldots] \tag{32}$$

The derived Markov process $\mathcal{M}_{\mathcal{S}_{K-1}^*}$, corresponding to the example of Figure 2, is depicted in Figure 7. To simplify our notation, let us express $\boldsymbol{\pi}_{\mathcal{M}_K}$ as follows:

$$\boldsymbol{\pi}_{\mathcal{M}_K} = [p_0, p_1, p_2, \ldots]$$

Then, we can express the flow balance equations of the Markov process $\mathcal{M}_{\mathcal{S}_{K-1}^*}$ as:

$$\begin{aligned} [\lambda + K\mu(1 - f_0)]\, p_0 &= K\mu p_1 \\ [\lambda + K\mu]\, p_k &= K\mu f_k p_0 + K\mu p_{k+1} + \sum_{i=0}^{k-1} p_i \lambda g_{k-i} \qquad \text{for } k > 0 \end{aligned}$$

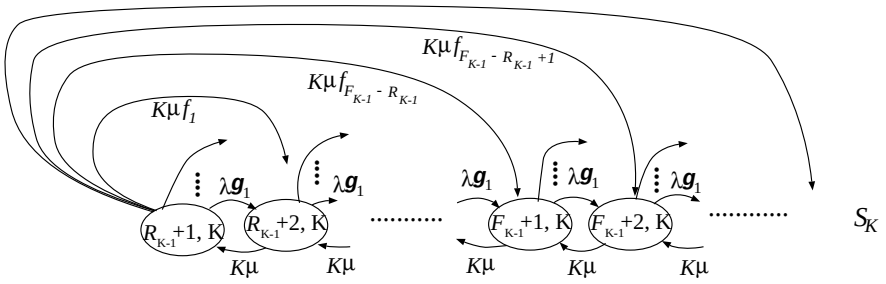

Figure 7: State transition diagram after elimination of $\mathcal{S}_1$ to $\mathcal{S}_{K-1}$.

Let

$$P(z) = \sum_{k=0}^{\infty} p_k z^k \ ; \ F(z) = \sum_{k=0}^{\infty} f_k z^k \ ; \ \overline{f} = \left.\frac{d\,F(s)}{dz}\right|_{z=1} \ ; \ G(z) = \sum_{k=1}^{\infty} g_k z^k \ ; \ \overline{g} = \left.\frac{d\,G(z)}{dz}\right|_{z=1}$$

And, we can express the Z-transform of $\boldsymbol{\pi}_{\mathcal{M}_K} = [p_0, p_1, \ldots]$ as:

$$P(z) \quad = \quad \frac{p_0 K\mu\,[1 - zF(z)]}{\lambda z\,[1 - G(z)] - K\mu\,(1 - z)} \tag{33}$$

where p_0 can be computed by evaluating $P(z)|_{z=1} = 1$; thus,

$$p_0 \quad = \quad \frac{K\mu - \lambda\overline{g}}{K\mu(\overline{f} + 1)} \tag{34}$$

Once we find the $Z-$transform, we can easily evaluate the expected number of customers in $\mathcal{M}_{\mathcal{S}^*_{K-1}}$ as:

$$\left.\frac{d\,P(z)}{dz}\right|_{z=1} + R_{K-1} + 1$$

5.2.1 Analysis of the Aggregated Process

All that remains at this point is to determine the probabilities of being in each set $\mathcal{S}_i$. As in Section 4.1, we can aggregate each set $\mathcal{S}_i$ in the original Markov process, $\mathcal{M}_b$, into a single state, i, for $i = 1, \ldots, K$. Since we have obtained the conditional steady state probabilities for all states in $\mathcal{S}_1$ through $\mathcal{S}_{K-1}$ as well as the steady state probability for state $(R_{K-1} + 1, K)$ in $\mathcal{S}_K$ (we refer to it as p_0 above), we can easily compute the transition rates of the aggregated process. We denote a transition, in the aggregated process, from state i to state j by $r_{i,j}$ and describe each transition as follows:

$$r_{i,j} \quad = \quad \boldsymbol{\pi}_{\mathcal{M}_i} \boldsymbol{Q}_{\mathcal{S}_i,\mathcal{S}_j} \boldsymbol{e} \qquad \text{for } 1 \le i < j \le K \tag{35}$$

$$r_{i,i-1} \quad = \quad i\mu \boldsymbol{\pi}_{\mathcal{M}_i} \boldsymbol{e}_1 \qquad \text{for } 2 \le i \le K \tag{36}$$

Since the state space of the aggregated process is finite, we can use the flow balance equations to compute $\boldsymbol{\pi} = [\pi(1), \pi(2), \ldots, \pi(K)]$, the steady state probability vector of the aggregated process, as follows. We define

$$r^*_{i,j} = \sum_{k=j}^{K} r_{i,k} \qquad 1 \le i < j \le K$$

and express $\pi(j)$ in terms of $\pi(1)C_j$, where:

$$C_j = \left(\frac{1}{r_{j,j-1}}\right)\left[\sum_{k=1}^{j-1} C_k r^*_{k,j}\right] \qquad 2 \le j \le K \tag{37}$$

with the initial value of $C_1 = 1$. With all the coefficients C_j defined, we have:

$$\pi(1) = \left(\sum_{k=1}^{K} C_k\right)^{-1} \qquad \text{and} \tag{38}$$

$$\pi(j) = \left(\sum_{k=1}^{K} C_k\right)^{-1} C_j \qquad 2 \le j \le K \tag{39}$$

5.2.2 Performance Measures

Let $\boldsymbol{V}_{\mathcal{M}_i}$ denote a column vector such that the j^{th} component of the vector represents the number of customers in the queueing system when the system is in the j^{th} state in $\mathcal{S}_i$. Then the average number of customers in the original Markov process $\mathcal{M}_b$, denoted by N_b, can be expressed as:

$$N_b = \sum_{i=1}^{K-1} \pi(i)\boldsymbol{\pi}_{\mathcal{M}_i}\boldsymbol{V}_{\mathcal{M}_i} + \pi(K)\left(\left.\frac{d\,P(z)}{dz}\right|_{z=1} + R_{K-1} + 1\right) \tag{40}$$

Using Little's result, we can obtain the average customer response time, denoted by T_b, as:

$$T_b = \frac{1}{\lambda\overline{g}}\left[\sum_{i=1}^{K-1} \pi(i)\boldsymbol{\pi}_{\mathcal{M}_i}\boldsymbol{V}_{\mathcal{M}_i} + \pi(K)\left(\left.\frac{d\,P(z)}{dz}\right|_{z=1} + R_{K-1} + 1\right)\right] \tag{41}$$

Remarks on Complexity of Solution: As in Section 5.1, it is useful at this point to consider the complexity of our technique where the number of multiplications required by the computation of N_b is used as the measure of time complexity. As before, the major contributors to the time complexity of computing N_b are: (a) computation of the aggregate state probabilities and (b) evaluation of the summation in Equation (40). The complexity of computing the aggregate state probabilities is $O(K^2)$. The time complexity of evaluating the finite summation in Equation (40) is due to the method chosen to compute the steady state probabilities, for instance, using the power method [21] gives the complexity[5] of $O((F_1^3+\sum_{l=2}^{K-1}(F_l-R_{l-1})^3))$. Of course, the corresponding space complexity, for storing a transition matrix for partition $\mathcal{S}_l$, is $O((\max(F_1^2, \max_{2\le l\le K-1}(F_l - R_{l-1})))^2)$. What remains is the complexity of evaluating the infinite part of Equation (40); this is a function of the bulk arrival sizes distribution, i.e., it depends on the actual Z-transform of Equation (33). Since we do not assume a specific distribution for bulk sizes in the derivation of our solution, we do not pursue this matter any further. Note that, one drawback of the bulk arrivals case solution is that the different partitions can not be solved in parallel, as in the homogeneous servers case, i.e., we need to "fold the partitions" one partition at a time, and thus the computation must necessarily proceed in a sequential manner.

[5]Empirical evidence indicates that other iterative as well as direct methods are more efficient than the power method, which we use here for simplicity of presentation; however, since that is not the focus of the paper, we will not discuss it here any further.

5.3 Heterogeneous Servers

As in the case of homogeneous servers, we first partition the state space, $\mathcal{S}_h$, of the original Markov process $\mathcal{M}_h$ into disjoint sets (refer to Figure 3), where the states in each set $\mathcal{S}_l$, $1 \leq l \leq K$, correspond to the states of the original Markov process where server l is busy[6] (of course, other servers may be busy in set $\mathcal{S}_l$, that is any server $k < l$ maybe be busy as well), i.e.,

$$\mathcal{S}_l = \{(i, \boldsymbol{j}) \mid (i, \boldsymbol{j}) \in \mathcal{S}_h \ \wedge \ \boldsymbol{j} \in \{0,1\}^{(l-1)}\{1\}\{0\}^{(K-l)}\}$$

Also let us define

$$\mathcal{S}_l^- = \bigcup_{i=1}^{l-1} \mathcal{S}_i \quad \text{and} \quad \mathcal{S}_l^+ = \bigcup_{i=l+1}^{K} \mathcal{S}_i$$

The heterogeneous case is somewhat more complicated than the homogeneous servers case, however, we can still use the method of stochastic complementation as follows. The last partition, $\mathcal{S}_K$, has only a single entry state from a state in $\mathcal{S}_{K-1}$, namely the state with $F_{K-1}+1$ customers. This means that however we leave $\mathcal{S}_K$, and whatever partition we go to, we will always come back to $\mathcal{S}_K$, from a state in $\mathcal{S}_{K-1}$, through the state with $F_{K-1}+1$ customers. Therefore, in creating a stochastic complement for the states in $\mathcal{S}_K$, all rates out of the states in $\mathcal{S}_K$ (regardless of what state they are from and where they lead) can be "folded back" to the state with $F_{K-1}+1$ customers. Thus, we can compute the conditional steady state probabilities for states in $\mathcal{S}_K$, given that $\mathcal{M}_h$ is in $\mathcal{S}_K$. This is accomplished by constructing a Markov process, $\mathcal{M}_K$, which has state space $\mathcal{S}_K$ with all transitions being the same as those (corresponding to states in $\mathcal{S}_K$) in the original Markov process $\mathcal{M}_h$, except that any transition from $\mathcal{S}_K$ to $\mathcal{S}_j$ (where $1 \leq j < K$) in $\mathcal{M}_h$ becomes a transition to state $(F_{K-1}+1, \{1\}^K)$ in $\mathcal{M}_K$. This is precisely an application of Theorem 3. The resulting process, $\mathcal{M}_K$, corresponding to the 3-server example of Figure 3, is illustrated in Figure 8. The solution for the steady state probabilities for all

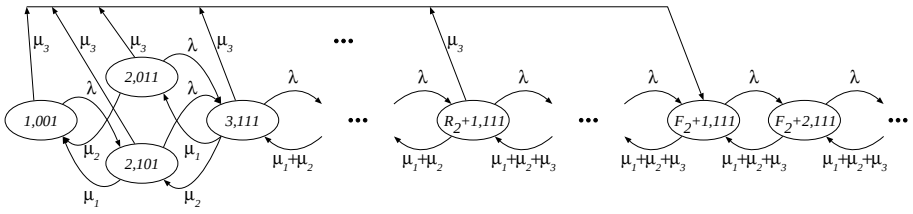

Figure 8: State transition diagram for $\mathcal{M}_3$.

states in $\mathcal{M}_K$, denoted by $\boldsymbol{\pi}_{\mathcal{M}_K}$, is given in Section 5.3.3[7]. Note that, these are exactly the conditional steady state probabilities of states in $\mathcal{S}_K$ of the original process $\mathcal{M}_h$, given that $\mathcal{M}_h$ is in $\mathcal{S}_K$.

Let us further examine the transition structure of Figure 3. Since there is only a single entry from $\mathcal{S}_{K-1}^-$ to $\{\mathcal{S}_{K-1} \bigcup \mathcal{S}_{K-1}^+\}$, namely through the state with $F_{K-2}+1$ customers, we can compute the stochastic complement for $\{\mathcal{S}_{K-1} \bigcup \mathcal{S}_{K-1}^+\}$, using Theorem 3. The transition diagram corresponding to the stochastic complement of $\{\mathcal{S}_{K-1} \bigcup \mathcal{S}_{K-1}^+\}$, for the example system of Figure 3, is illustrated in Figure 9. Note that, in Figure 9, there is a single exit from $\mathcal{S}_{K-1}$ to $\mathcal{S}_{K-1}^+$, namely from the state with

[6] Once again, to simplify the notation, we assume that state $(0, \{0\}^K)$ is in $\mathcal{S}_1$.

[7] We postpone the details of the steady state probabilities solution until Section 5.3.3 so as not to distract the reader from the basic solution approach.

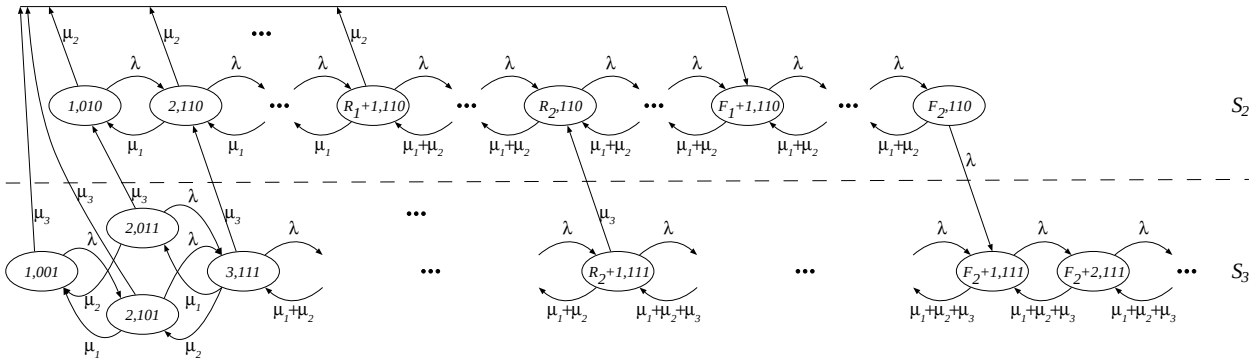

Figure 9: State transition diagram for $\{\mathcal{S}_{K-1} \bigcup \mathcal{S}_{K-1}^{+}\}$, where $K = 3$.

F_{K-1} customers, but multiple returns from $\mathcal{S}_{K-1}^{+}$ to $\mathcal{S}_{K-1}$. Since $\mathcal{S}_K = \mathcal{S}_{K-1}^{+}$, and since we are able to compute $\boldsymbol{\pi}_{\mathcal{M}_K}$, the conditional steady state probability vector for the states in $\mathcal{S}_K$, given that $\mathcal{M}_h$ is in $\mathcal{S}_K$, we can use this information to complete the construction of the stochastic complement for the states in $\mathcal{S}_{K-1}$. (Note that, since $\mathcal{S}_K = \mathcal{S}_{K-1}^{+}$, from now on we will refer to $\mathcal{S}_K$ only.) Similarly to the case of bulk arrivals, the probabilistic interpretation of this approach is that we are *redistributing* the transition rate of λ (from the state with F_{K-1} customers in $\mathcal{S}_{K-1}$ to the state with $F_{K-1}+1$ customers in $\mathcal{S}_K$) back to the states in $\mathcal{S}_{K-1}$. This redistribution should be proportional to the relative visit ratios at which $\mathcal{S}_{K-1}$ is entered from $\mathcal{S}_K$, either directly or by first going to some other partition $\mathcal{S}_j$, $j < K-1$ and then returning to $\mathcal{S}_{K-1}$ through $\mathcal{S}_{K-2}$. We can compute a vector, $\boldsymbol{f}$, corresponding to these visit ratios, using our solution for $\boldsymbol{\pi}_{\mathcal{M}_K}$; the elements of $\boldsymbol{f}$ are as follows:

$$f(i,\boldsymbol{j}) \;=\; \begin{cases} 0 & \text{for } R_{K-1} < i \le F_{K-1} \\ & \text{and } (i,\boldsymbol{j}) \neq (F_{K-2}+1, \boldsymbol{j} = \{1\}^{(K-1)}\{0\}) \\ \dfrac{\boldsymbol{\pi}_{\mathcal{M}_K}(i+1, G_{+}^{K}(\boldsymbol{j}))\mu_K}{\sum_{(1\le k\le R_{K-1}+1)\wedge(\boldsymbol{n}\in\{0,1\}^{(K-1)}\{1\})} \boldsymbol{\pi}_{\mathcal{M}_K}(k,\boldsymbol{n})\mu_K} & \text{for } 1 \le i \le R_{K-1} \\ \dfrac{\sum_{(1\le k\le K-1)\wedge(\boldsymbol{n}\in\{0,1\}^{(K-2)}\{01\})} \boldsymbol{\pi}_{\mathcal{M}_K}(k,\boldsymbol{n})\mu_K}{\sum_{(1\le k\le R_{K-1}+1)\wedge(\boldsymbol{n}\in\{0,1\}^{(K-1)}\{1\})} \boldsymbol{\pi}_{\mathcal{M}_K}(k,\boldsymbol{n})\mu_K} & \text{for } i = F_{K-2}+1,\ \boldsymbol{j} = \{1\}^{(K-1)}\{0\} \end{cases} \tag{42}$$

where $\boldsymbol{\pi}_{\mathcal{M}_K}(i,\boldsymbol{j})$ is the conditional steady state probability of being in state $(i,\boldsymbol{j})$ in $\mathcal{S}_K$, given that the original Markov process $\mathcal{M}_h$ is in $\mathcal{S}_K$ and

$$\sum_{(1\le i\le F_{K-1})\wedge(\boldsymbol{j}\in\{0,1\}^{(K-2)}\{10\})} f(i,\boldsymbol{j}) \;=\; 1$$

The transition diagram for $\mathcal{M}_{K-1}$ is illustrated in Figure 10, where $l = K-1$.

Thus, we can compute the conditional steady state probabilities for states in $\mathcal{S}_{K-1}$, given that $\mathcal{M}_h$ is in $\mathcal{S}_{K-1}$ by constructing a Markov process, $\mathcal{M}_{K-1}$ which has state space $\mathcal{S}_{K-1}$ with all transitions being the same as those in the original Markov process $\mathcal{M}_h$, except that any transition from $\mathcal{S}_{K-1}$ to $\mathcal{S}_{K-1}^{-}$ in $\mathcal{M}_h$ becomes a transition to state $(F_{K-2}+1, \{1\}^{(K-1)}\{0\})$ in $\mathcal{M}_{K-1}$. Furthermore, the single transition from state $(F_{K-1}, \{1\}^{(K-1)}\{0\})$ to state $(F_{K-1}+1, \{1\}^{K})$ in $\mathcal{M}_h$ is redistributed back to the states in $\mathcal{S}_{K-1}$ according to the visit ratios given in Equation (42). This is reflected in the following theorem (which is the "heterogeneous counterpart" to Theorems 4 and 5).

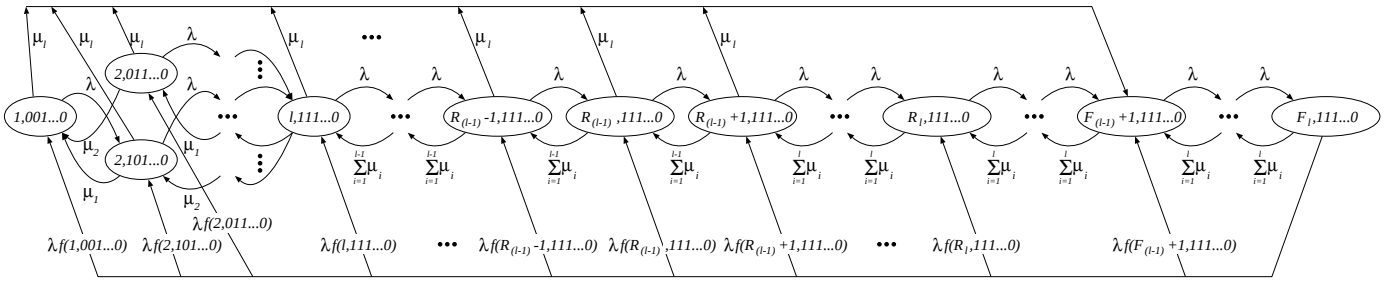

Figure 10: State transition diagram for subset S_l, where $1 \leq l < K$.

Theorem 6 *The steady state probabilities solution of the Markov process $\mathcal{M}_{K-1}$ is the conditional steady state probabilities solution for the states in $\mathcal{S}_{K-1}$ of the original Markov process $\mathcal{M}_h$, given that $\mathcal{M}_h$ is in partition $\mathcal{S}_{K-1}$.*

Proof: Since there is a single return from $\mathcal{S}^-_{K-1}$ to $\{\mathcal{S}_{K-1} \cup \mathcal{S}^+_{K-1}\}$, using Theorem 3, we can obtain conditional steady state probabilities for the states in $\{\mathcal{S}_{K-1} \cup \mathcal{S}^+_{K-1}\}$, given that the original process $\mathcal{M}_h$ is in $\{\mathcal{S}_{K-1} \cup \mathcal{S}^+_{K-1}\}$. Thus we can construct a stochastic complement for the states in $\{\mathcal{S}_{K-1} \cup \mathcal{S}^+_{K-1}\}$; the transition rate matrix of the corresponding Markov process is:

$$\begin{bmatrix} \boldsymbol{Q}_{\mathcal{S}_{K-1},\mathcal{S}_{K-1}} & \boldsymbol{Q}_{\mathcal{S}_{K-1},\mathcal{S}^+_{K-1}} \\ \boldsymbol{Q}_{\mathcal{S}^+_{K-1},\mathcal{S}_{K-1}} & \boldsymbol{Q}_{\mathcal{S}^+_{K-1},\mathcal{S}^+_{K-1}} \end{bmatrix}$$

where $\boldsymbol{Q}_{i,j}$ is the transition rate sub-matrix corresponding to transitions from partition i to partition j. Note that, $\boldsymbol{Q}_{\mathcal{S}_{K-1},\mathcal{S}^+_{K-1}}$ has only a single non-zero row which has a single non-zero entry (corresponding to state $(F_{K-1}, \{1\}^{(K-1)}\{0\})$). $\boldsymbol{Q}_{\mathcal{S}^+_{K-1},\mathcal{S}_{K-1}}$, however, has multiple non-zero columns. Each non-zero column that corresponds to a state $(i, \{0,1\}^{(K-2)}\{10\})$, where $(1 \leq i \leq R_{K-1})$, has only a single non-zero element. The non-zero column that corresponds to the state $(F_{K-2}+1, \{1\}^{(K-1)}\{0\})$ has multiple non-zero entries. Referring to Theorem 3, in order to compute the stochastic complement of Equation (4), we must compute $\text{diag}(\boldsymbol{Q}_{\mathcal{S}_{K-1},\mathcal{S}^+_{K-1}}\boldsymbol{e})\boldsymbol{Z}$. Computing $\text{diag}(\boldsymbol{Q}_{\mathcal{S}_{K-1},\mathcal{S}^+_{K-1}}\boldsymbol{e})$ is simple, since $\boldsymbol{Q}_{\mathcal{S}_{K-1},\mathcal{S}^+_{K-1}}$ has only a single non-zero element. Thus, referring to Equation (5), $r_i = \lambda$ for i corresponding to state $(F_{K-1}, \{1\}^{(K-1)}\{0\})$ and $r_j = 0$, for all $j \neq i$.[8] This means that, referring to Equation (5) again, we need only to construct $\boldsymbol{z}_i$, i.e., a vector which determines how r_i is redistributed between the states in $\mathcal{S}_{K-1}$. This redistribution is governed by the original process $\mathcal{M}_h$, namely by the relative frequencies with which the states in $\mathcal{S}_{K-1}$ are visited, when the original process makes a transition out of $\mathcal{S}_K$; these frequencies are determined by the conditional steady state probabilities of the states in $\mathcal{S}_K$, i.e., $\boldsymbol{\pi}_{\mathcal{M}_K}$, and the transitions which take the original process, $\mathcal{M}_h$, from states in $\mathcal{S}_K$ to states in $\mathcal{S}_{K-1}$, either directly or by first going to some other partition $\mathcal{S}_j$, $j < K-1$, and then coming back to $\mathcal{S}_{K-1}$ through $\mathcal{S}_{K-2}$. Thus, the redistribution of r_i is governed precisely by the vector $\boldsymbol{f}$ computed in Equation (42), i.e., $\boldsymbol{z}_i = \boldsymbol{f}$. ∎

If we continue in this manner, we can solve for all conditional steady state probabilities for states in $\mathcal{S}_K$ through $\mathcal{S}_1$.

[8] In the discussion of Section 2, r_i refers to a transition *probability* and here we are discussing transition *rates*; however, as mentioned in that section, we can easily convert between the two.

5.3.1 Analysis of the Aggregated Process

As in Sections 4.1 and 4.2, we can now create an aggregated version of $\mathcal{M}_h$ which is illustrated in Figure 11 and all that remains to compute are the probabilities of being in each set $\mathcal{S}_l$. The aggregated process is fairly simple to solve and has the following transition rates (refer to Figure 11)

$$\lambda_i = \lambda \boldsymbol{\pi}_{\mathcal{M}_i}(F_i, \boldsymbol{n}) \qquad \text{for } (\boldsymbol{n} \in \{1\}^i\{0\}^{(K-i)}) \text{ and } (i = 1, 2, \ldots, K-1) \tag{43}$$

$$\mu_{ij} = \mu_i \sum_{\substack{(1 \le k \le i) \\ \wedge(\boldsymbol{n} \in H)}} \boldsymbol{\pi}_{\mathcal{M}_i}(k, \boldsymbol{n}) \qquad \text{for } 1 \le j < i \le K \tag{44}$$

where $H = \{\boldsymbol{c} \,|\, \boldsymbol{c} = G^i_+(\boldsymbol{d}) \wedge \boldsymbol{d} \in \{0,1\}^{(j-1)}\{1\}\{0\}^{(K-j)}\}$. Note that the structure of the aggregated

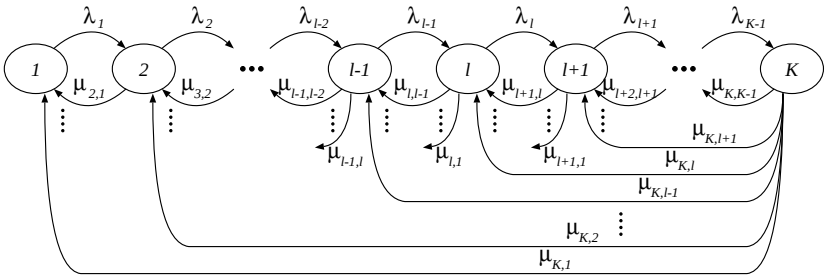

Figure 11: State transition diagram for aggregated system.

processes for the heterogeneous servers case can be made identical to that of the aggregated process for the bulk arrivals case, if we rename the states properly, i.e., in the bulk arrivals case, there is a single transition due to a departure from state i to state $i-1$ and multiple transition due to arrivals, from state i to state j, for all $j > i$; in the heterogeneous servers case, there is a single transition due to an arrival from state i to state $i+1$ and multiple transitions due to departures, from state i to state j, for all $j < i$. Thus, we can use similar equations to obtain the aggregate steady state probabilities for the heterogeneous servers case as we used for the bulk arrivals case (refer to Section 5.2). In the interests of brevity, we do not repeat these equations here.

Given the steady state probabilities, $\boldsymbol{\pi} = (\pi(1), \pi(2), \ldots, \pi(K))$, for the aggregated process, we can use them to adjust the conditional steady state probabilities computed in Section 5.3.3 to obtain the final solution, i.e., to compute the steady state probabilities for the original Markov process $\mathcal{M}_h$ as follows:

$$\boldsymbol{\pi}_{\mathcal{M}_h}(i, \boldsymbol{j}) = \pi(l) * \boldsymbol{\pi}_{\mathcal{M}_l}(i, \boldsymbol{j})$$

where $\boldsymbol{\pi}_{\mathcal{M}_l}(i, \boldsymbol{j})$ is the conditional steady state probability of being in state $(i, \boldsymbol{j}) \in \mathcal{S}_l$, where $\boldsymbol{j} \in \{0,1\}^{(l-1)}\{1\}\{0\}^{(K-l)}$, and $\pi(l)$ is the probability of being in partition $\mathcal{S}_l$.

5.3.2 Performance Measures

Finally, we can compute some performance measures. Let N_h be the average number of customers and T_h be the average customer waiting time in the original process $\mathcal{M}_h$. Also, let $G(\boldsymbol{n})$ be a function

which returns the number of 1's in the string $\boldsymbol{n}$. Then,

$$\begin{aligned} N_h &= \sum_{i=1}^{K-1} \pi(i) \left[\sum_{k=1}^{i-1} G(\boldsymbol{n}) \pi_i(k, \boldsymbol{n}) + \sum_{k=i}^{F_i} k \pi_i(k, \boldsymbol{n}) \right] \\ &+ \pi(K) \left[\sum_{k=1}^{K-1} G(\boldsymbol{n}) \pi_K(k, \boldsymbol{n}) + \sum_{k=K}^{\infty} k \pi_K(k, \boldsymbol{n}) \right] \end{aligned} \tag{45}$$

where $\boldsymbol{\pi}_i(k, \boldsymbol{n})$ is the conditional steady state probability of being in state $(k, \boldsymbol{n})$, given that the original process is in $\mathcal{S}_i$, and $\boldsymbol{n} \in \{0,1\}^{(i-1)}\{1\}\{0\}^{(K-i)}$. Using Little's result, we can obtain T_h, that is:

$$\begin{aligned} T_h &= \frac{1}{\lambda} \left(\sum_{i=1}^{K-1} \pi(i) \left[\sum_{k=1}^{i-1} G(\boldsymbol{n}) \pi_i(k, \boldsymbol{n}) + \sum_{k=i}^{F_i} k \pi_i(k, \boldsymbol{n}) \right] \right. \\ &+ \left. \pi(K) \left[\sum_{k=1}^{K-1} G(\boldsymbol{n}) \pi_K(k, \boldsymbol{n}) + \sum_{k=K}^{\infty} k \pi_K(k, \boldsymbol{n}) \right] \right) \end{aligned} \tag{46}$$

Remarks on Complexity of Solution: As in Sections 5.1 and 5.2, it is useful at this point to consider the complexity of our technique where the number of multiplications required by the computation of N_h is used as the measure of time complexity. As before, the major contributors to the time complexity of computing N_h are: (a) computation of the aggregate state probabilities and (b) evaluation of the summations in Equation (45). The complexity of computing the aggregate state probabilities is $O(K^2)$. The complexity of evaluating the finite summations in Equation (45), each corresponding to a partition $\mathcal{S}_l$, is $O(F_l)$ for each $1 \leq l \leq K-1$. What remains is the complexity of evaluating the infinite summation, which may not be apparent directly from Equation (45). Using the equations for conditional steady probabilities of $\mathcal{S}_K$ (refer to Section 5.3.3), we can evaluate the infinite summation in Equation (45) — in fact, to simplify this problem, we need only to evaluate the following part of the infinite summation

$$\pi(K) \sum_{k=F_{K-1}+2}^{\infty} k \pi_K(k, \boldsymbol{n})$$

which can be done by using Equation (47) alone; this equation can be simplified to be (assuming that $\frac{\lambda}{\sum_{j=1}^{K} \mu_j} \neq 1$):

$$\pi(K) \pi_K(F_{K-1}+1, \boldsymbol{n}) \frac{\lambda}{\sum_{j=1}^{K} \mu_j} \left[\frac{F_{K-1}+1}{1 - \frac{\lambda}{\sum_{j=1}^{K} \mu_j}} + \frac{1}{\left(1 - \frac{\lambda}{\sum_{j=1}^{K} \mu_j}\right)^2} \right]$$

where evaluation of this equation requires $O(1)$ multiplications, and thus evaluation of the entire infinite summation in Equation (45) requires $O(K^3 + F_{K-1})$ multiplications. Thus, the overall time complexity of evaluating N_h is $O(K^3 + F_{K-1} + \sum_{l=1}^{K-1} F_l^3)$, where the overall space complexity is $O((\max_{1 \leq l \leq K-1}(F_l))^2)$. Note that, one drawback of the heterogeneous case solution (just as in the case of bulk arrivals) is that the different partitions can not be solved in parallel, as in the homogeneous case, i.e., we need to "fold the partitions up" one partition at a time, and thus the computation must necessarily proceed in a sequential manner.

5.3.3 Analysis of $\mathcal{M}_l$

We begin by solving for the conditional steady state probabilities of the states in $\mathcal{S}_K$, given that the original process $\mathcal{M}_h$ is in $\mathcal{S}_K$. In the following derivation, referring to Figure 8, we give a set of flow balance equations for the states in $\mathcal{S}_K$, where the notation $\pi_l(i, \boldsymbol{j})$ refers to a state with i customers in partition $\mathcal{S}_l$, i.e., a partition where the l^{th} server is busy and $\boldsymbol{j} = \{0,1\}^{(l-1)}\{1\}\{0\}^{(K-l)}$. Furthermore, in the remainder of this section

$$\sum_{k=1}^{i} \pi_l(k, \boldsymbol{n}) = \sum_{\substack{(1 \le k \le l-1) \wedge \\ (\boldsymbol{n} \in \{0,1\}^{(l-1)}\{1\}\{0\}^{(K-l)})}} \pi_l(k, \boldsymbol{n}) \quad + \sum_{\substack{(l \le k \le i) \wedge \\ (\boldsymbol{n} \in \{1\}^{(l)}\{0\}^{(K-l)})}} \pi_l(k, \boldsymbol{n})$$

where for ease of presentation, we use the simpler notation of $\sum_{k=1}^{i} \pi_l(k, \boldsymbol{n})$.

For all states where the number of customers is i, where $i \ge F_{K-1} + 2$:

$$\pi_K(i-1, \boldsymbol{j})\lambda = \pi_K(i, \boldsymbol{j}) \sum_{j=1}^{K} \mu_j$$

$$\pi_K(i, \boldsymbol{j}) = \pi_K(F_{K-1} + 1, \boldsymbol{j}) \left[\frac{\lambda}{\sum_{j=1}^{K} \mu_j}\right]^{(i - F_{K-1} - 1)} \quad \text{for } (i \ge F_{K-1} + 2) \wedge (\boldsymbol{j} \in \{1\}^K) \tag{47}$$

For all states where the number of customers is i, where $R_{K-1} + 1 < i < F_{K-1} + 2$:

$$\pi_K(i-1, \boldsymbol{j})\lambda + \mu_K \sum_{k=1}^{R_{K-1}+1} \pi_K(k, \boldsymbol{n}) = \pi_K(i, \boldsymbol{j}) \sum_{j=1}^{K} \mu_j$$

$$\pi_K(i, \boldsymbol{j}) = \pi_K(R_{K-1} + 1, \boldsymbol{j}) \left(\frac{\lambda}{\sum_{j=1}^{K} \mu_j}\right)^{(i - R_{K-1} - 1)} + \left(\frac{\mu_K \sum_{k=1}^{R_{K-1}+1} \pi_K(k, \boldsymbol{n})}{\sum_{j=1}^{K} \mu_j}\right) \sum_{n=0}^{i-R_{K-1}-2} \left(\frac{\lambda}{\sum_{j=1}^{K} \mu_j}\right)^n$$

If we let $C_1 = \frac{\lambda}{\sum_{j=1}^{K} \mu_j}$ and assume that $C_1 \neq 1$ (a similar derivation can be given for $C_1 = 1$), then we can simplify the above equation to

$$\begin{aligned} \pi_K(i, \boldsymbol{j}) &= \pi_K(R_{K-1} + 1, \boldsymbol{j})(C_1)^{(i - R_{K-1} - 1)} \\ &+ \left(\frac{\mu_K \sum_{k=1}^{R_{K-1}+1} \pi_K(k, \boldsymbol{n})}{\sum_{j=1}^{K} \mu_j}\right) \frac{\left(1 - (C_1)^{(i - R_{K-1} - 1)}\right)}{1 - C_1} \\ &\qquad \text{for } (R_{K-1} + 1 < i < F_{K-1} + 2) \wedge (\boldsymbol{j} \in \{1\}^K) \end{aligned} \tag{48}$$

For all states where the number of customers is i, where $K < i \le R_{K-1} + 1$:

$$\pi_K(i-1, \boldsymbol{j})\lambda + \mu_K \sum_{k=1}^{i-1} \pi_K(k, \boldsymbol{n}) = \pi_K(i, \boldsymbol{j}) \sum_{j=1}^{K-1} \mu_j$$

$$\pi_K(i,\boldsymbol{j}) = \pi_K(K,\boldsymbol{j})\left(\frac{\lambda}{\sum_{j=1}^{K-1}\mu_j}\right)^{(i-K)} + \sum_{n=0}^{i-K-1}\left[\frac{\mu_K}{\sum_{j=1}^{K-1}\mu_j}\left(\sum_{k=1}^{i-n-1}\pi_K(k,\boldsymbol{n})\right)\left(\frac{\lambda}{\sum_{j=1}^{K-1}\mu_j}\right)^{n}\right]$$

If we let $C_2 = \frac{\lambda}{\sum_{j=1}^{K-1}\mu_j}$ and assume that $C_2 \neq 1$ (a similar derivation can be given for $C_2 = 1$), then we can simplify the above equation to

$$\begin{aligned}\pi_K(i,\boldsymbol{j}) &= \pi_K(K,\boldsymbol{j})(C_2)^{(i-k)} \\ &+ \frac{\mu_K}{\sum_{j=1}^{K-1}\mu_j}\left[\left(\frac{1-(C_2)^{(i-K)}}{1-C_2}\right)\sum_{k=1}^{K}\pi_K(k,\boldsymbol{n}) + \sum_{k=K+1}^{i-1}\pi_K(k,\boldsymbol{n})\left(\frac{1-C_2^{(i-k)}}{1-C_2}\right)\right] \qquad (49)\\ &\qquad\qquad \text{for } (K < i \leq R_{K-1}+1) \wedge (\boldsymbol{j} \in \{1\}^K)\end{aligned}$$

At this point, all that remains is to determine expressions for conditional steady state probabilities for the states in $\mathcal{S}_K$ with K or fewer customers. Let $\mathcal{S}_K^{'}$ be a subset of $\mathcal{S}_K$ containing all states $(i,\boldsymbol{j})$ where $1 \leq i \leq K$ and $\boldsymbol{j} \in \{0,1\}^{(K-1)}\{1\}$. Since the flow balance equations for states in $\mathcal{S}_K^{'}$ do not have "nice" structure, we will determine the conditional steady state probabilities for states in $\mathcal{S}_K^{'}$ by employing the concept of stochastic complementation one last time. The rate of μ_K out of any state in $\mathcal{S}_K^{'}$, which corresponds to a transition to state $(F_{K-1}+1,\{1\}^K)$ in $\{\mathcal{S}_K - \mathcal{S}_K^{'}\}$, can be "folded" back into state $(K,\{1\}^K)$ in $\mathcal{S}_K^{'}$, since this is the only entry state into $\mathcal{S}_K^{'}$ from $\{\mathcal{S}_K - \mathcal{S}_K^{'}\}$. In other words, we can compute a stochastic complement of the states in $\mathcal{S}_K^{'}$ using Theorem 3 and solve this relatively small subset of states using any chosen solution technique, as described in [21].

At this point the conditional steady state probability for each state i, $\pi_K(i,\boldsymbol{j})$, is expressed as a function of conditional steady state probabilities of states with j customers, where $j < i$. Once we compute the conditional steady state probabilities for the states in $\mathcal{S}_K^{'}$ (using any chosen solution technique, as described in [21]), we can express all other conditional steady state probabilities in Equations (47)-(50) in terms of $\pi_K(K,\boldsymbol{j})$, and $\pi_K(K,\boldsymbol{j})$ can be computed using the following equation:

$$\sum_{i=1}^{\infty}\pi_K(i,\boldsymbol{n}) = 1$$

Thus, we have determined the conditional steady state probability vector, $\boldsymbol{\pi}_{\mathcal{M}_K}$, for all states in $\mathcal{S}_K$, given that the original process $\mathcal{M}_h$ is in $\mathcal{S}_K$.

We can now proceed to computing $\boldsymbol{\pi}_{\mathcal{M}_l}$ for $1 \leq l < K$. The transition structure of partition $\mathcal{S}_l$, $1 \leq l < K$, is depicted in Figure 10. As can be seen from Figure 10, unfortunately, none of the other partitions, $\mathcal{S}_l$, $1 \leq l \leq K-1$, are as "well-structured" as $\mathcal{S}_K$; fortunately, they are all finite and thus we can compute all $\boldsymbol{\pi}_{\mathcal{M}_l}$, $1 \leq l \leq K-1$, using any chosen solution technique, as described in [21]. At this point, we have obtained all the conditional steady state probabilities[9] for each set $\mathcal{S}_l$, $1 \leq l \leq K-1$.

[9] As pointed out in Section 5.3.1, all that remains is to solve the aggregated process of Figure 11 and adjust the conditional steady state probabilities accordingly. Thus, we have a complete solution for the steady state probabilities of a heterogeneous multi-server threshold queueing system with hysteresis.

6 Numerical Examples

In this section we present numerical examples of the performance of the different variations of the threshold-based queueing system with hysteresis, using expected system response time as the performance measure of interest.

We first consider the *homogeneous* servers case. Figures 12 and 13 illustrate examples of homogeneous server systems with Poisson arrivals, where $K = 2$ and 5, respectively, and $\mu = 1.0$. In

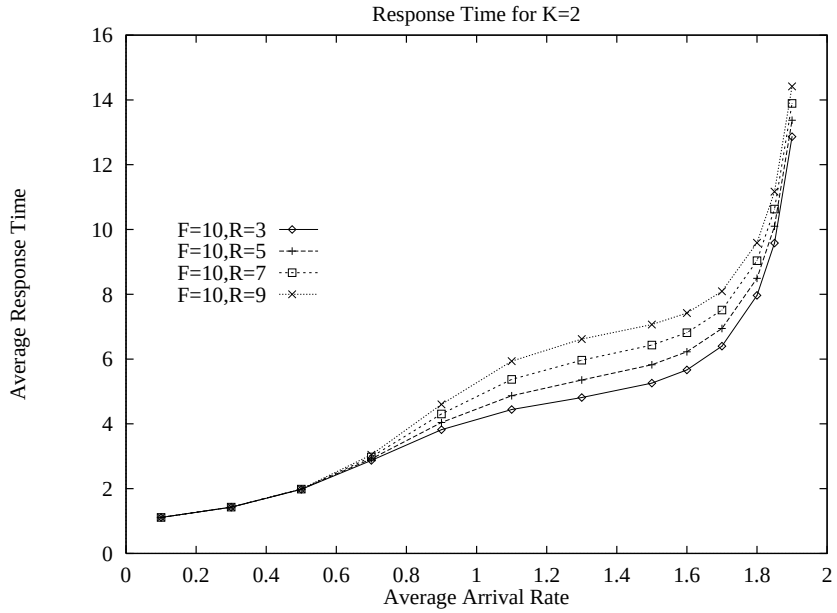

Figure 12: Results for $K = 2, \mu = 1.0$ and different forward and reverse thresholds.

both cases, we vary the forward and reverse threshold vectors and observe how the queueing system with hysteresis maintains a level of expected system response time for large input ranges. From these figures, we observe that the average response time curves are different from the *usual* queueing (e.g., M/M/K) response time curves. This is due to a combination of threshold values and workload – that is, the response time may decrease with higher loads (as can be seen in figures below) since at higher loads we may cross some threshold(s) "more frequently" (on the average) which allows us to operate with a greater number of servers more frequently (on the average). Note that, in these experiments, the difference in expected response time between systems with different threshold vectors (with all other things being equal) is relatively small. Of course, the expected cost of those systems would necessarily have to be different, since (as was mentioned in Section 1) it is a function of various factors, including threshold vector values. This indicates that there is room for improvement of the cost/performance ratio of the system — a topic (although outside the scope of this paper) we intend to pursue in our future work.

We next consider the *bulk arrivals* variation of the problem. Figure 14 illustrates an example of a homogeneous servers system with a *bulk* arrival process, that is, the arrival process is Poisson where each arrival corresponds to an arrival of i customers with probability g_i, where $1 \leq i \leq 3$; in this

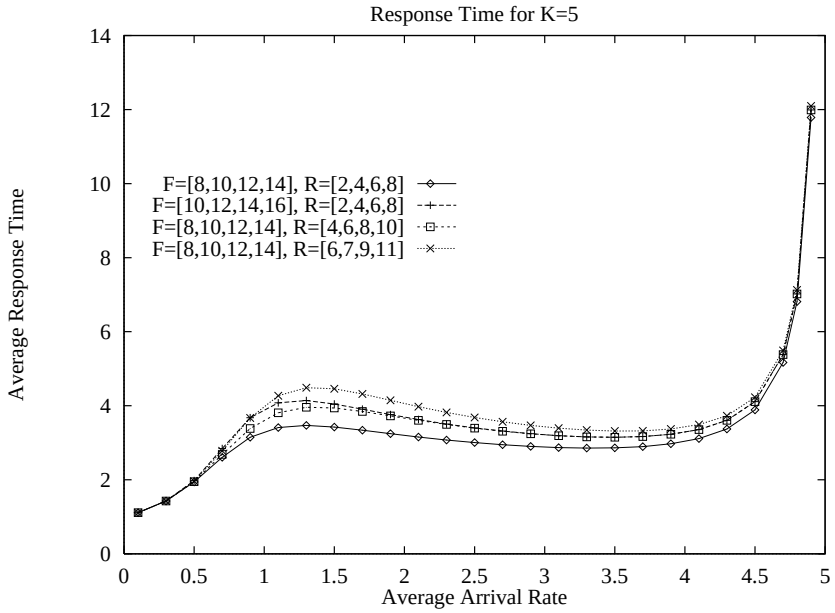

Figure 13: Results for $K = 5, \mu = 1.0$ and different forward and reverse thresholds.

figure $K = 4$ and $\mu = 1.0$. We vary the bulk size probabilities and observe that the expected response time increases slowly as we increase the bulk arrival rate. As already mentioned, similar "unusual" behavior of response time curves can also be observed here, as in the homogeneous servers case.

Finally, we consider the *heterogeneous* servers variation of problem. Figures 15 and 16 illustrate examples of heterogeneous server systems with Poisson arrivals. In Figure 15, $K = 2$, $\mu_1 = 1.5$, $\mu_2 = 1.0$, $F_1 = 10$, and we vary the reverse thresholds, R. In this figure, we can again observe the "unusual" behavior of expected response time curves, as in the case of the homogeneous servers systems. In Figure 16, we consider the case of $K = 3$, and we vary the threshold values and the service rates of the K servers. This system exhibits a more interesting behavior, when experimenting with different relative values of the heterogeneous service rates – again a topic of future work, i.e., determining appropriate combination of threshold values and service rates so as to achieve an optimal cost/performance ratio.

7 Conclusions

We considered and solved several variations of a multi-server threshold-based queueing system with hysteresis whose behavior is governed by a set of forward and reverse thresholds, namely: (1) homogeneous servers with Poisson arrivals, (2) homogeneous servers with bulk (Poisson) arrivals, and (3) heterogeneous servers with Poisson arrivals. We placed no restrictions on the number of servers or the bulk sizes, and we solved all variations of the problem using the concept of stochastic complementation. The contributions of our work are as follows. We presented a more intuitive and extensible method (than in the case of [8]) for obtaining a closed-form solution to the multi-server threshold queueing problem with hysteresis, when the servers are homogeneous and there is no restriction on the

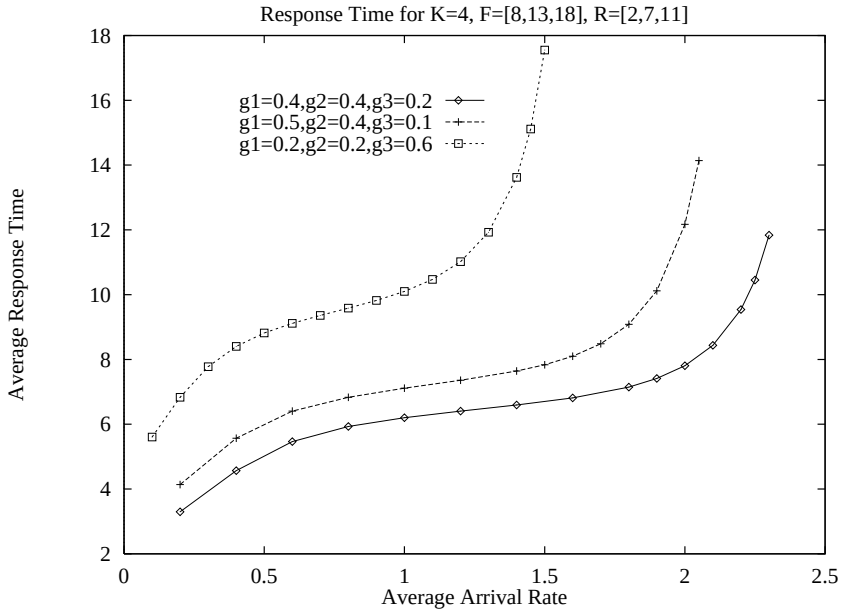

Figure 14: Results for bulk arrivals with $K = 4, \mu = 1.0$.

number of servers. We also presented algorithmic solutions for the bulk-arrivals and heterogeneous-servers variations of the problem (again, with no restrictions on the size of the bulk or the number of servers); to the best of our knowledge, these variations of the problem, with no restriction on the number of servers or the bulk size, have not been solved exactly in the past (except for the solution of the 2-heterogeneous-servers problem in [8]). The ease with which we were able to obtain solutions to these variations of the problem demonstrates the extensibility of our method. Note, that we can use stochastic complementation to derive closed-form solutions for some limited forms of heterogeneous-servers and bulk-arrivals variations of the problem, such as heterogeneous servers with $K = 2$ and bulk arrivals with a limited bulk size. Finally, our technique works both for systems with finite and infinite waiting rooms.

Acknowledgment: The authors would like to thank the anonymous referees for their helpful and insightful comments. Leana Golubchik's research was supported in part by the NSF CAREER grant CCR-96-25013; part of the work was done while she was with the Department of Computer Science at Columbia University. John C.S. Lui's research was supported in part by the UGC and the CUHK Grant.

References

[1] P. J. Courtois. *Decomposability : queueing and computer system applications.* ACM monograph series, Academic Press, New York, 1977.

[2] P. J. Courtois, P. Semal. *Computable Bounds for Conditional Steady-State Probabilities in Large Markov Chains and Queueing Models.* IEEE JSAC, Vol 4, number 6, September, 1986.

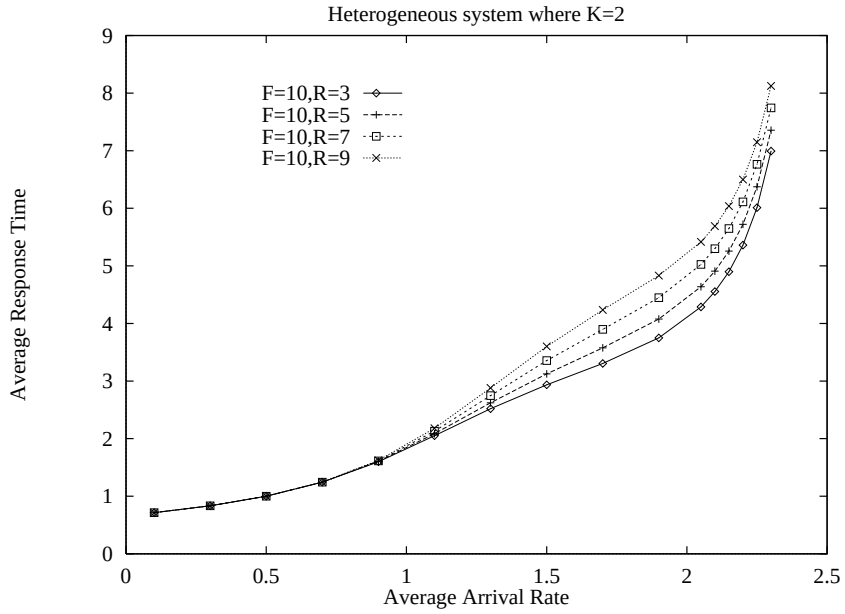

Figure 15: Results for heterogeneous servers with $K = 2, \mu = [1.5, 1.0]$.

[3] Leana Golubchik, John C.S. Lui, *Bounding of Performance Measures for a Threshold-based Queueing System with Hysteresis.* ACM SIGMETRICS '97 Conference, Vol 25, 1, pp 147-157, June 1997.

[4] W.K. Grassman. *Transient Solutions in Markovian Queueing Systems.* Computer and Operation Research, 4 pp. 47-53, 1977.

[5] S.C. Graves and J. Keilson. *The Compensation Method Applied to a One-product Production/Inventory Problem.* Journal of Math. Operational Research, Vol 6, pp 246-262, 1981.

[6] O.C. Ibe. *An Approximate Analysis of a Multi-server Queueing System with a Fixed Order of Access.* IBM Research Report, RC9346, 1982.

[7] O.C. Ibe and K. Maruyama. *An Approximation Method for a Class of Queueing Systems*, Performance Evaluation Vol 5, pp 15-27, 1985.

[8] O.C. Ibe and J. Keilson. *Multi-server threshold queues with hysteresis.* Performance Evaluation, 21, page 185-212, 1995.

[9] J. Keilson. *Green's Function Methods in Probability Theory*, Charles Griffin, London, 1965.

[10] J. Keilson. *Markov Chain Models: Rarity and Exponentiality*, Springer, New York, 1979.

[11] P.J.B. King *Computer and Communication Systems Performance Modeling*, Prentice-Hall, New York, 1990.

[12] L. Kleinrock. *Queueing Systems, Volume I*, Wiley-Interscience, 1975.

[13] R.L. Larsen and A.K. Agrawala. *Control of a heterogeneous two-server exponential queueing system.* IEEE Trans. on Software Engineering, Vol 9, pp 552-526, 1983.

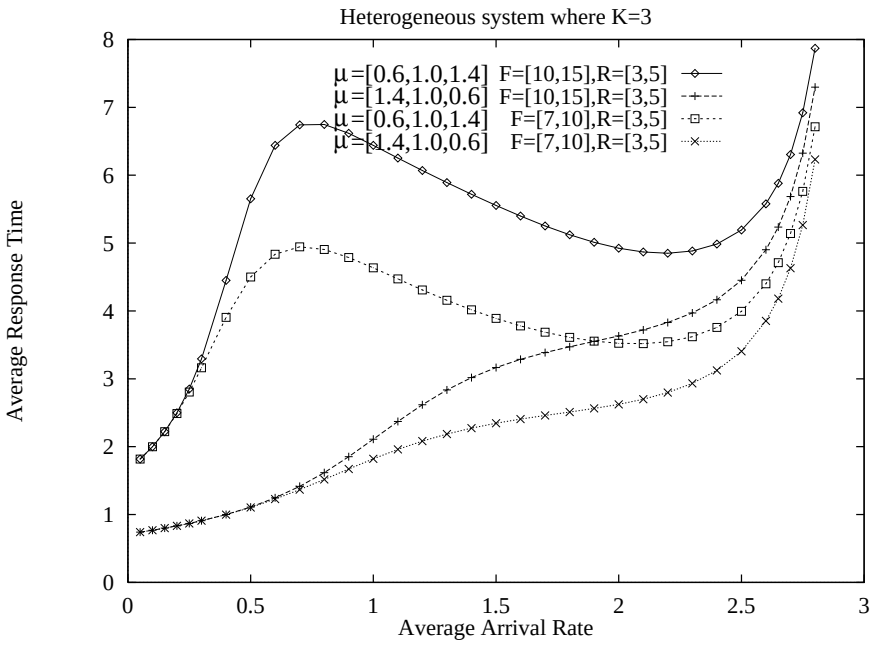

Figure 16: Results for heterogeneous servers with $K = 3$.

[14] W. Lin and P.R. Kumar. *Optimal Control of a Queueing System with two Heterogeneous Servers.* IEEE Trans. on Automatic Control, Vol 29, pp. 696-703, 1984.

[15] J.D.C Little. *A Proof of the Queueing Formula $L = \lambda W$,* Operations Research, Vol 9, 383-387, 1967.

[16] John C.S. Lui, R.R. Muntz, *Bounding Methodology for Computing Steady State Availability of Repairable Computer Systems*, Journal of ACM, pp. 676-707, July, 1994.

[17] John C.S. Lui, R.R. Muntz and D. Towsley. *Bounding the Mean Response Time of the Minimum Expected Delay Routing Policy: An Algorithmic Approach.* IEEE Trans. on Computers, 44(5), pp. 1371-1382, 1995.

[18] C.D. Meyer. *Stochastic Complementation, Uncoupling Markov Chains and the Theory of Nearly Reducible Systems.* SIAM Review, 31(2), pp. 240-272, 1989.

[19] J.A. Morrison. *Two-server queue with One Server Idle Below a Threshold* Queueing Systems, Vol 7, pp 325-336, 1990.

[20] R. Nelson and D. Towsley. *Approximating the Mean Time in System in a Multiple-server Queue that uses Threshold Scheduling.* Journal of Operation Research, Vol 35, pp 419-427, 1987.

[21] W.J. Stewart. *Introduction to the Numerical Solution of Markov Chains.* Princeton Press, 1994.