Submitted to *Operations Research*
manuscript (Please, provide the mansucript number!)

$M/M/c$ Queue with Two Priority Classes

Jianfu Wang
Nanyang Business School, Nanyang Technological University, wangjf@ntu.edu.sg
Opher Baron
Rotman School of Management, University of Toronto, opher.baron@rotman.utoronto.ca
Alan Scheller-Wolf
Tepper School of Business, Carnegie Mellon University, awolf@andrew.cmu.edu

This paper provides the first exact analysis of a preemptive $M/M/c$ queue with two priority classes having different service rates. To perform our analysis, we introduce a new technique to reduce the 2-dimensionally (2D) infinite Markov Chain (MC), representing the two class state space, into a 1-dimensionally (1D) infinite MC, from which the Generating Function (GF) of the number of low-priority jobs can be derived in *closed form*. (The high-priority jobs form a simple $M/M/c$ system, and are thus easy to solve.) We demonstrate our methodology for the $c = 1, 2$ cases; when $c > 2$, the closed-form expression of the GF becomes cumbersome. We thus develop an *exact* algorithm to calculate the moments of the number of low-priority jobs for any $c \geq 2$. Numerical examples demonstrate the accuracy of our algorithm, and generate insights on: (i) the relative effect of improving the service rate of either priority class on the mean sojourn time of low-priority jobs; (ii) the performance of a system having many slow servers compared with one having fewer fast servers; and (iii) the validity of the square root staffing rule in maintaining a fixed service level for the low priority class. Finally, we demonstrate the potential of our methodology to solve other problems such as an $M/M/c$ queue with two priority classes, where the high-priority class is completely impatient.

Key words: multi-server queue, multi-class, preemptive priority, different service rates

1. Introduction

The last decade has witnessed a growing usage of prioritization in the service industry. Examples range from amusement parks, where customers with VIP tickets can skip regular lines, to cloud computing, where customers who pay the standard price have strict priority over customers who pay the discounted price, to hospital emergency departments that prioritize more urgent patients.

There are three main motivations for prioritization. The first motivation is that different customers may have different willingness to pay (or valuations) for the same product. The second

motivation is that customers may require different products or services, where some of these products are more profitable than others. The third motivation is that having different service levels may substantially affect long term profitability; for example first time customers who receive excellent service are more likely to become loyal ones, see, e.g., Afèche et al. (2012).

Modeling the effects of prioritization due to the first and third motivations can be achieved with identical service time distributions for different segments. However, appropriately characterizing the effects of prioritization due to the second motivation requires capturing differences in service times. Moreover, there are many practical applications where customers with different service requests are given different priorities. For example, contact centers prioritize phone calls over emails. Similarly, renewals of driver's licenses require a photograph and thus typically take longer than renewals of car licenses; the latter are prioritized (according to the principal of shortest processing time first). Likewise, at airports processing times of the aircrew are shorter than those of air-travelers, who have a lower priority. In all of these applications, there are several servers rather than a single one.

There are many papers that use queueing theory to derive and analyze different prioritization policies in services (e.g., Maglaras and Zeevi 2005 and references therein), inventory settings (e.g., Abouee-Mehrizi et al. 2012 and references therein), and dynamic scheduling (e.g., Van Mieghem 1995 and references therein). This literature typically focuses on characterizing the distribution of the sojourn time of different priority classes. Specifically, the distribution of sojourn times for single-server queues with priorities, such as the $M/G/1$ (see e.g., Takagi 1991) are well known. Miller (1981) gives a computationally efficient algorithm to derive the steady state probability distribution of an $M/M/1$ queue with priority by using the matrix-analytic method. But it is difficult to extend these results to multi-server priority queues. In fact, to the best of our knowledge no exact solution for the sojourn time distribution in a multi-server queueing system serving multiple priority classes with *different* service rates has appeared in the literature.

Much of the multi-server literature has focused on the $M/M/c$ queue. The $M/M/c$ queue with multiple priority classes and identical service rates was first investigated by Davis (1966), finding a closed-form expression for the Laplace transform (LT) of any priority class's waiting time. For

the same setting, Kella and Yechiali (1985) elegantly derived this LT. Buzen and Bondi (1983) gave a simple approximation for each priority class's mean sojourn time in a preemptive system with *different* service rates for each priority class. Maglaras and Zeevi (2004) used a diffusion approximation to solve a similar problem with impatient high priority customers in a heavy-traffic regime. Recently, Harchol-Balter et al. (2005) approximated the sojourn time of a preemptive $M/PH/c$ queue with different service rates. They also provide a taxonomy of relevant literature.

In this paper, we consider an $M/M/c$ queue with two priority classes under a *preemptive* discipline under either preemptive-resume or preemptive-repeat (new service times are drawn whenever preempted customers re-enter service). In particular, preemptive-resume may be an appropriate model in the emergency department and contact center contexts.

We assume that Class-i jobs arrive according to a Poisson process with rate λ_i, $i = 1, 2$. Service times for Class-i jobs are exponentially distributed with parameter μ_i, $i = 1, 2$. For stability, we require $\sum_{i=1}^{2} \frac{\lambda_i}{\mu_i} < c$. Class-1 jobs have preemptive priority, thus the analysis of Class-1 jobs is straightforward. The main goal of this paper is to develop an efficient *exact* algorithm to calculate Class-2 jobs' expected sojourn time and probability of waiting, in steady state.

To develop our algorithm, we use an approach for the analysis of continuous-time Markov chains (MCs), which we call Queue Decomposition, based on Abouee-Mehrizi et al. (2012): In many cases, the metrics of interest for a queueing system only depend on certain parts of the MC. In these cases, for the other parts of the MC, we do not need to keep track of detailed information; the transition probabilities and the time the system stays in such parts of the MC are sufficient. The queue decomposition approach is simple, yet powerful, because it allows us to focus the analysis on smaller and simpler parts of the original system. The analysis of each part of the MC jointly with a careful characterization of the transition probabilities between these parts yields an exact analysis of the original system.

In our case, we are interested in the number of Class-2 jobs in steady state, which is distributed identically to the number of Class-2 jobs seen by Class-2 departures. Such departures only occur

when there are fewer than c Class-1 jobs in the system. Thus, it is not necessary to track all the information when the MC has c or more Class-1 jobs.

Using our analysis, we derive insights on how the performances of a priority system changes as the characteristics of the jobs or servers change. These insights were unavailable before, due to the lack of exact algorithms for this preemptive system with different service rates for each priority class. In particular, we consider the following three questions:

1. How does changing μ_1 or μ_2 affect the expected sojourn time of Class-2 jobs?
2. Do Class-2 jobs prefer few fast servers or many slow servers? Why?
3. Does the square root staffing rule hold for Class-2 jobs?

After introducing the model and background results in Section 2, we present the key ideas of our methodology in Section 3. We demonstrate the methodology on the single-server case in Section 4. We discuss the $c \geq 2$ servers case in Section 5. We provide an efficient exact numerical method for systems with $c \geq 2$ in Section 6. Numerical results, insights, and extensions are given in Section 7. We summarize the paper in Section 8. All proofs are in the Appendix.

2. Model and Preliminary Results

We consider an $M/M/c$ queue with two priority classes. Let q_i, $i = 1, 2$ be the number of Class-i jobs in the system, and S_i and W_i, $i = 1, 2$ be the random variables representing the steady state sojourn time (from arrival until departure) and waiting time of Class-i jobs in the system respectively. Note that, due to Class-1 jobs' preemptive priority, Class-2 jobs might be preempted from service. In this case, we consider the difference between a Class-2 job's sojourn time and its total service time as its waiting time, i.e., $E[W_2] = E[S_2] - \frac{1}{\mu_2}$.

For the $\mu_1 = \mu_2$ case, the sojourn time distribution of each priority class is given in Buzen and Bondi (1983). We, however, consider this problem when Class-1 and Class-2 have different service requirements (i.e., $\mu_1 \neq \mu_2$). Figure 1 illustrates the Markov Chain (MC) for the number of jobs from different priority classes, (q_1, q_2), in the system. This MC is infinite in two dimensions, complicating the analysis.

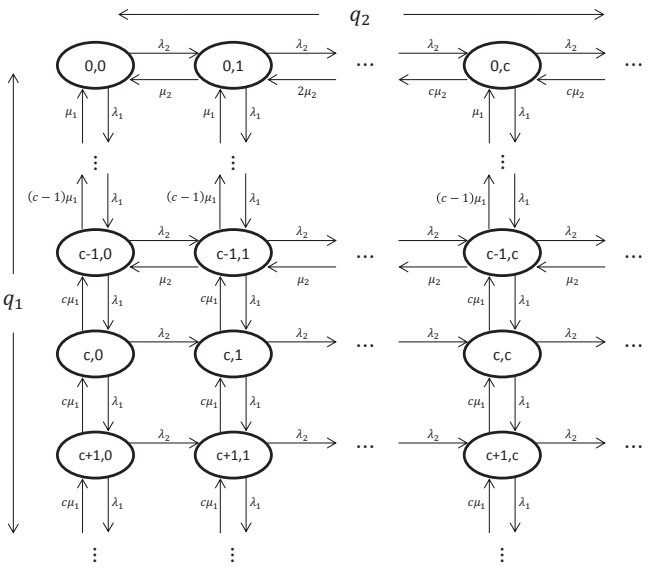

Figure 1 MC of the $M/M/c$ queue with two priority classes.

Due to the preemptive priority, Class-1 jobs see a classic $M/M/c$ queue. Their service rate at state (q_1, q_2) is $\mu_1 \min(q_1, c)$; independent of q_2. Thus, the distribution of Class-1 jobs' sojourn time is (e.g., Section 3.4, Buzacott and Shanthikumar 1993)

$$P\{S_1 < t\} = 1 - e^{-\mu_1 t} - \frac{(e^{-(c\mu_1-\lambda_1)t} - e^{-\mu_1 t})}{1-(c-\frac{\lambda_1}{\mu_1})} \frac{\lambda_1^c}{\mu_1^c c!}((1-\frac{\lambda_1}{c\mu_1})\sum_{i=0}^{c-1}\frac{\lambda_1^i}{\mu_1^i i!} + \frac{\lambda_1^c}{\mu_1^c c!})^{-1}.$$

Therefore, we focus on deriving the sojourn time and probability of no wait for Class-2 jobs.

Let r_{q_1,q_2} be Class-2 jobs' service rate when the MC is at state (q_1, q_2). In state (q_1, q_2), the number of servers available to Class-2 jobs is $c - \min(q_1, c)$. Thus,

$$r_{q_1,q_2} = \mu_2 \min(c - \min(q_1, c), q_2). \tag{1}$$

Let R_{q_2} denote Class-2 jobs' service rate vector when there are q_2 Class-2 jobs in the system, i.e., R_{q_2} includes all r_{q_1,q_2} for states in the q_2^{th} column of the MC in Figure 1. When $q_1 \geq c$, Class-2 jobs' service rate is always zero, so R_{q_2} does not include r_{q_1,q_2} for $q_1 \geq c$. Using (1),

$$R_{q_2} = (r_{0,q_2}, \ldots, r_{c-1,q_2}). \tag{2}$$

Note that we have an identical service rate vector $R_{q_2} = (c\mu_2, (c-1)\mu_2, \ldots, \mu_2)$ for any $q_2 \geq c$.

Let $v(q_1, q_2)$ denote the total rate at which the MC moves out of state (q_1, q_2). Then

$$v(q_1, q_2) = \lambda_1 + \lambda_2 + \mu_1 \min(q_1, c) + \mu_2 \min(c - \min(q_1, c), q_2). \tag{3}$$

Before moving to the next section, we recall several results and define several special matrices that are used extensively in the paper. Let t be a random time interval with Laplace transform $LT^t(s)$; let X be the number of Poisson(λ) arrivals during t, and $G_X(z)$ be the generating function (GF) of X. The distribution of X as a function of $LT^t(s)$ is given as:

$$P\{X = x\} = \frac{(-\lambda)^x}{x!} LT^{t(x)}(\lambda); \tag{4}$$

$$G_X(z) = LT^t(\lambda - \lambda z), \tag{5}$$

where $LT^{t(x)}(\lambda)$ denotes the x^{th} derivative of $LT^t(s)$ evaluated at λ (see e.g., (3.58) and (3.67) respectively in Buzacott and Shanthikumar 1993).

We write any column vector as the transpose of a corresponding row vector. Let $\mathbf{0}_{i\times j}$ and $\mathbf{1}_{i\times j}$ denote $i \times j$ matrices with all elements zero or one, respectively, and let I denote the identity matrix. The following Lemma is important for derivations in Sections 4 and 5.

LEMMA 1. *Assume a MC's state space is composed of two sets: a transient set, $\mathbb{T}$ and an absorbing set, $\mathbb{A}$. Let $\Gamma_{\mathbb{T}\to\mathbb{T}}$ and $\Gamma_{\mathbb{T}\to\mathbb{A}}$ be the one step transition matrices from $\mathbb{T}$ to $\mathbb{T}$ and $\mathbb{T}$ to $\mathbb{A}$ respectively. Then, $P\{\mathbb{A}_j \mid \mathbb{T}_i\}$, the probability of being absorbed in state $\mathbb{A}_j \in \mathbb{A}$ starting at state $\mathbb{T}_i \in \mathbb{T}$ is*

$$[P\{\mathbb{A}_j \mid \mathbb{T}_i\}]_{\mathbb{T}_i\in\mathbb{T},\mathbb{A}_j\in\mathbb{A}} = (I - \Gamma_{\mathbb{T}\to\mathbb{T}})^{-1}\Gamma_{\mathbb{T}\to\mathbb{A}}. \tag{6}$$

3. Simplification - The 1D-Infinite MC

Finding the distribution of S_2 is challenging because the Markov chain (MC) in Figure 1 is 2D-infinite. We transform the 2D-infinite continuous-time MC into a 1D-infinite discrete-time MC. We first simplify the MC by aggregating the behavior of the system during a *Class-1 busy period* (BP), which starts when there are c or more Class-1 jobs in the system (i.e., once q_1 increases to c) and ends when the number of Class-1 jobs q_1 drops to $c-1$.

Clearly, during each BP the service rate of Class-1 jobs is $c\mu_1$ and the arrival rate of Class-1 jobs is λ_1. Thus, during this BP, the MC of Class-1 jobs is identical to the busy period of an $M/M/1$ queue with arrival rate λ_1 and service rate $c\mu_1$ (see e.g., Harchol-Balter et al. 2005). Thus, the Laplace transform (LT) of this BP is (see, e.g., Takagi 1991, Chapter 1)

$$LT^{BP}(s) = \frac{1}{2\lambda_1}(\lambda_1 + c\mu_1 + s - \sqrt{(\lambda_1 + c\mu_1 + s)^2 - 4c\lambda_1\mu_1}). \tag{7}$$

Next, using (4), we obtain the probability of having l Class-2 arrivals during the BP

$$\alpha_l^{BP} = \frac{(-\lambda_2)^l}{l!} LT^{BP^{(l)}}(\lambda_2),\ l = 0, 1, 2, \ldots. \tag{8}$$

Let $G_{\alpha^{BP}}(z)$ be the generating function (GF) of α^{BP}; then from (5), we have

$$G_{\alpha^{BP}}(z) = LT^{BP}(\lambda_2 - \lambda_2 z). \tag{9}$$

During the BP, all Class-2 arrivals join the queue. When the BP is over, q_1 becomes $c-1$ and the distribution of the number of Class-2 arrivals in the BP can be calculated from (7) and (8). Specifically, let BP_i denote a Class-1 busy period that starts from a Class-1 arrival at state $(c-1, i)$, $i = 0, 1, \ldots$; then BP_i ends in state $(c-1, i+j)$ with probability (w.p.) α_j^{BP}, for $j \geq 0$. Using this method, we lose information on when those Class-2 arrivals occurred during the BP, but we will establish next that this information is not necessary.

After aggregating the Class-1 busy periods in the MC into the BP_i's, we get a 1D-infinite *discrete-time* MC with $c+1$ rows: The first c rows are identical to the first c rows in the original MC, and the $(c+1)^{th}$ row is composed of BP_i's. When the system leaves BP_i, it may enter any state $(c-1, q_2)$ with $q_2 \geq i$. Figure 2 illustrates this 1D-infinite discrete-time MC.

Still, to the best of our knowledge, there are no known exact solutions for this ladder-like 1D-infinite discrete-time MC. We overcome this difficulty by observing the system state at departure epochs of Class-2 jobs, i.e., analyzing the embedded Markov chain (EMC).

From Section 5.1.3 of Gross et al. (2008), Class-2 departures in steady state observe the steady state distribution of q_2. Thus, if we can derive the steady state probability distribution of the EMC,

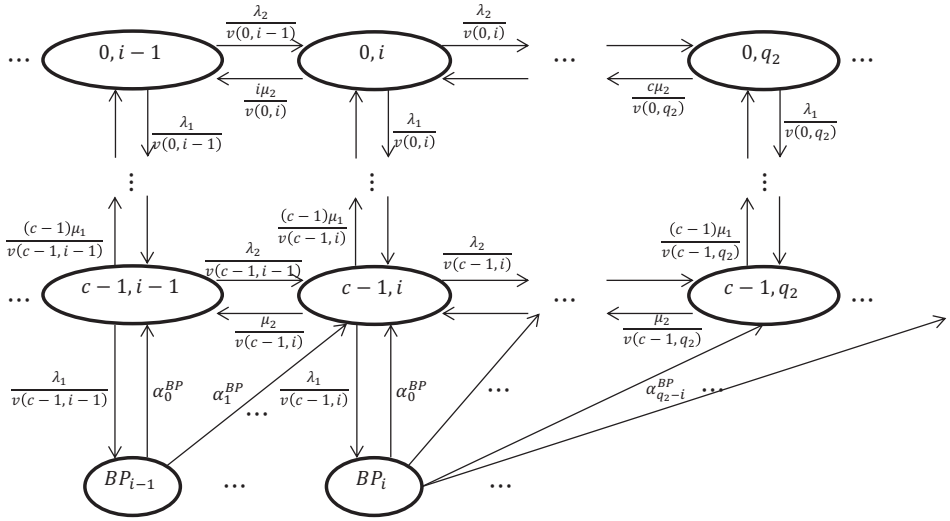

Figure 2 The Simplified MC.

we obtain the distribution of q_2. Then, we can derive different moments of q_2, and the expected sojourn time of Class-2 jobs, $E[S_2]$, from Little's Law. If we further assume that the service order of Class-2 follows FIFO (e.g., when items are made to order), we can use Distributional Little's Law from Bertsimas and Nakazato (1995) to express the sojourn time distribution of Class-2 jobs.

To determine the steady state distribution of the EMC, we can follow the three steps used to analyze the EMC of the standard $M/G/1$ model (see e.g., Section 3.3.2, Buzacott and Shanthikumar 1993): 1) derive the one-step transition matrix of the EMC, 2) characterize the generating function (GF) of the number of jobs seen by a departure in steady state, and 3) derive the unknown constant in the expression of this GF.

4. The Single-server Case

To develop some intuition for our analytical procedure, we first demonstrate it in the single-server setting. The solution for the sojourn time of Class-2 jobs in this case is known (see e.g., Takagi (1991) Chapter 3):

$$LT^{\hat{S}_2}(s) = \frac{2(\lambda_1\mu_2 + \lambda_2\mu_1 - \mu_1\mu_2)}{(\mu_2 - 2\mu_1)s + \lambda_1\mu_2 + 2\lambda_2\mu_1 - \mu_1\mu_2 - \mu_2\sqrt{(s+\lambda_1+\mu_1)^2 - 4\lambda_1\mu_1}}. \tag{10}$$

Our methodology provides an alternative proof, and more importantly it can be used in the multi-server case. For convenience, we denote quantities related to the $c = 1$ case with a "hat" ($\hat{}$).

Let $\hat{L}_k^2$, the number of Class-2 jobs seen by the k^{th} Class-2 departure, be the state of the EMC. Note that the system has one server, so Class-2 departures always see no Class-1 jobs. Let $\hat{M}$ be the EMC's transition matrix, i.e., the entry $\hat{m}_{i\to j}$ in $\hat{M}$ is defined as $\hat{m}_{i\to j} = P\{\hat{L}_{k+1}^2 = j \,\big|\, \hat{L}_k^2 = i\}$.

We next derive an equation relating $\hat{L}_k^2$ to $\hat{L}_{k+1}^2$. Let $\hat{D}_k$ be the k^{th} inter-departure time of Class-2 jobs (the time between the k^{th} and the $(k+1)^{st}$ Class-2 departure). Let the random variable $\alpha^{\hat{D}_k}$ be the number of $Poisson(\lambda_2)$ arrivals during $\hat{D}_k$. The number of Class-2 jobs seen by the $(k+1)^{st}$ Class-2 departure equals the number of Class-2 jobs seen by the k^{th} Class-2 departure minus one (the $(k+1)^{st}$ Class-2 departure) plus the number of Class-2 jobs that arrived during $\hat{D}_k$:

$$\hat{L}_{k+1}^2 = \hat{L}_k^2 - 1 + \alpha^{\hat{D}_k}. \tag{11}$$

From (11), we know that $\hat{L}_{k+1}^2 \geq \hat{L}_k^2 - 1$, so $\hat{m}_{i\to j}$ is zero, if $j < i-1$.

Thus, the transition matrix has the form illustrated in (12). Each row and column is labeled by the corresponding state $\hat{L}_k^2$. All elements of the lower triangle below the second row in $\hat{M}$ are zero.

$$\hat{M} = \begin{array}{c|ccccc|c} & 0 & 1 & 2 & 3 & 4 & \cdots \\ \hline 0 & \hat{m}_{0\to0} & \hat{m}_{0\to1} & \hat{m}_{0\to2} & \hat{m}_{0\to3} & \hat{m}_{0\to4} & \cdots \\ 1 & \hat{m}_{1\to0} & \hat{m}_{1\to1} & \hat{m}_{1\to2} & \hat{m}_{1\to3} & \hat{m}_{1\to4} & \cdots \\ 2 & 0 & \hat{m}_{2\to1} & \hat{m}_{2\to2} & \hat{m}_{2\to3} & \hat{m}_{2\to4} & \cdots \\ 3 & 0 & 0 & \hat{m}_{3\to2} & \hat{m}_{3\to3} & \hat{m}_{3\to4} & \cdots \\ 4 & 0 & 0 & 0 & \hat{m}_{4\to3} & \hat{m}_{4\to4} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}. \tag{12}$$

For $n \geq 0$, let $\hat{d}_n = P\left\{\hat{L}^2 = n\right\} = \lim_{k\to\infty} P\left\{\hat{L}_k^2 = n\right\}$, i.e., $\hat{L}^2$ is the time-stationary limiting random variable of $\hat{L}_k^2$ and $\hat{d}_n$ is the steady state probability that a Class-2 departure sees n Class-2 jobs. Let $G_{\hat{L}^2}(z) = \sum_{n=0}^{\infty} \hat{d}_n z^n$ be the generating function of $\hat{L}^2$.

4.1. Transition Matrix of the EMC

The transition rate $\hat{m}_{\hat{L}_k^2 \to \hat{L}_{k+1}^2}$ is closely related to the Class-2 jobs' service rate vector during the inter-departure time $\hat{D}_k$. The service rate vector, by (2), is only defined when no Class-1 jobs are in the system, and, since $c = 1$, has only one element. Furthermore,

- If $\hat{L}_k^2 \geq 1$: The Class-2 jobs' service rate vector remains μ_2 until the $(k+1)^{st}$ Class-2 departure. The service rate vector is independent of Class-2 arrivals during $\hat{D}_k$.

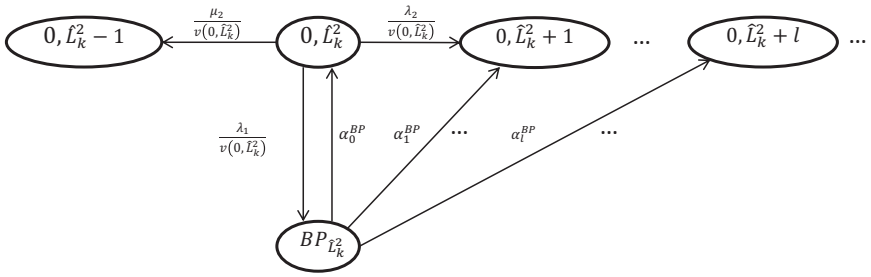

Figure 3 MC for the single server case where $\hat{L}_k^2 \geq c = 1$.

- If $\hat{L}_k^2 = 0$: The Class-2 jobs' service rate vector is zero until the next Class-2 arrival, and then it becomes μ_2.

4.1.1. The Transition Probabilities for $\hat{L}_k^2 \geq 1$ We know from (11) that the transition probabilities of the EMC are determined by $\alpha^{\hat{D}_k}$, which depends on $\hat{D}_k$. Thus, we first derive the Laplace transform of $\hat{D}_k$, $LT^{\hat{D}_k}(s)$. Then, using $LT^{\hat{D}_k}(s)$ and (4), we express the distribution of $\alpha^{\hat{D}_k}$, and then write the transition probabilities of the EMC using (11).

Figure 3 illustrates the service process of the $(k+1)^{st}$ Class-2 departure at the MC (not the EMC). At the k^{th} Class-2 departure, the MC enters state $(0, \hat{L}_k^2)$. Because $\hat{L}_k^2 \geq 1$, the rate of exiting from state $(0, \hat{L}_k^2)$ is $v(0, \hat{L}_k^2) = \lambda_1 + \lambda_2 + \mu_2$, thus after an $\exp(\lambda_1 + \lambda_2 + \mu_2)$ distributed time, the system would go to one of the following three states:

- State $BP_{\hat{L}_k^2}$, w.p. $\frac{\lambda_1}{v(0,\hat{L}_k^2)}$. The MC stays in the BP for a time period with an LT of $LT^{BP}(s)$. After this BP, the MC goes to state $(0, \hat{L}_k^2 + l)$ (with $l \geq 0$ Class-2 arrivals during the $BP_{\hat{L}_k^2}$ calculated from (8)). Due to the memoryless property and the fact that the Class-2 jobs' service rate vector stays the same, the LT of the time period from when the MC enters $(0, \hat{L}_k^2 + l)$ until the next Class-2 departure is identical to $LT^{\hat{D}_k}(s)$. Therefore, w.p. $\frac{\lambda_1}{v(0,\hat{L}_k^2)}$, $LT^{\hat{D}_k}(s)$ equals the LT of the sum of the time until the next event, the length of a BP, and $\hat{D}_k$: $\frac{\lambda_1+\lambda_2+\mu_2}{\lambda_1+\lambda_2+\mu_2+s} LT^{BP}(s) LT^{\hat{D}_k}(s)$.
- State $(0, \hat{L}_k^2 + 1)$, w.p. $\frac{\lambda_2}{v(0,\hat{L}_k^2)}$. Here, using similar reasoning as above: $LT^{\hat{D}_k}(s) = \frac{\lambda_1+\lambda_2+\mu_2}{\lambda_1+\lambda_2+\mu_2+s} LT^{\hat{D}_k}(s)$.
- State $(0, \hat{L}_k^2 - 1)$, w.p. $\frac{\mu_2}{v(0,\hat{L}_k^2)}$. The $(k+1)^{st}$ Class-2 departure occurs: $LT^{\hat{D}_k}(s) = \frac{\lambda_1+\lambda_2+\mu_2}{\lambda_1+\lambda_2+\mu_2+s}$.

Using the Total Probability Theorem (see, e.g., Papoulis 1984) and multiplying by $(\lambda_1 + \lambda_2 + \mu_2 + s)$, we get

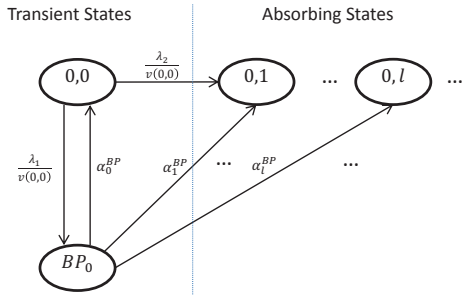

Figure 4 MC for the single server case where $\hat{L}_k^2 = 0$.

$$(\lambda_1 + \lambda_2 + \mu_2 + s)LT^{\hat{D}_k}(s) = \lambda_1 LT^{BP}(s)LT^{\hat{D}_k}(s) + \lambda_2 LT^{\hat{D}_k}(s) + \mu_2, \tag{13}$$

solving which gives

$$LT^{\hat{D}_k}(s) = \frac{\mu_2}{\lambda_1 + \mu_2 + s - \lambda_1 LT^{BP}(s)}.$$

We now return to the EMC. To simplify the notation, we let $\alpha_l^{\hat{D}_k} = \frac{(-\lambda_2)^l}{l!} LT^{\hat{D}_k}{}^{(l)}(\lambda_2)$. Then, using (4) and (11), we get the transition probabilities of the EMC from $\hat{L}_k^2 \geq 1$ to any $\hat{L}_{k+1}^2 \geq 0$:

$$\hat{m}_{\hat{L}_k^2 \to \hat{L}_{k+1}^2} = \begin{cases} 0 & \text{for } \hat{L}_{k+1}^2 < \hat{L}_k^2 - 1 \\ \alpha^{\hat{D}_k}_{\hat{L}_{k+1}^2 - \hat{L}_k^2 + 1} & \text{for } \hat{L}_{k+1}^2 \geq \hat{L}_k^2 - 1 \end{cases}, \tag{14}$$

which characterizes the rows of $\hat{M}$ in (12) corresponding to any $i \geq 1$.

4.1.2. The Transition Probabilities for $\hat{L}_k^2 = 0$ If $\hat{L}_k^2 = 0$ when the k^{th} Class-2 departure occurs, the next Class-2 event must be an arrival. This arrival may occur during BP_0, and there may be other Class-2 arrivals during BP_0. Taking this possibility into account, assume that when the service of the next Class-2 arrival is initiated, there are $l \geq 1$ Class-2 jobs in the system, i.e., the system enters state $(0, l)$ for $l \geq 1$. There are no transitions in the EMC until then.

Due to the memoryless property, the distribution of $\hat{L}_{k+1}^2$ given the system is in state $(0, l)$ is the same as the distribution of $\hat{L}_{k+1}^2$ given $\hat{L}_k^2 = l$, as given in (14) for $l \geq 1$. Thus, we require the first-passage probability distribution from state $(0, 0)$ to states $\{(0, l) \mid l \geq 1\}$. To find this probability, we consider the system after the k^{th} Class-2 departure as a MC with transient states $\{(0, 0), BP_0\}$, and absorbing states $\{(0, l) \mid l \geq 1\}$. Let $\hat{\Gamma}_{0 \to 0}$ and $\hat{\Gamma}_{0 \to 1^+}$ be the one-step transition matrices from $\{(0, 0), BP_0\}$ to $\{(0, 0), BP_0\}$ and $\{(0, l) \mid l \geq 1\}$, respectively.

In Figure 4, we use $v(0,0)=\lambda_1+\lambda_2$, depict the arrival process of jobs in $(0,0)$, and omit details that are not relevant to the development of this case. From Figure 4, we can get $\hat{\Gamma}_{0\to 0}$ and $\hat{\Gamma}_{0\to 1^+}$:

$$\hat{\Gamma}_{0\to 0}=\begin{array}{c} \\ (0,0) \\ BP_0 \end{array}\begin{array}{c} \begin{array}{cc} (0,0) & BP_0 \end{array} \\ \boxed{\begin{array}{cc} 0 & \frac{\lambda_1}{\lambda_1+\lambda_2} \\ \alpha_0^{BP} & 0 \end{array}} \end{array},\quad \hat{\Gamma}_{0\to 1^+}=\begin{array}{c} \\ (0,0) \\ BP_0 \end{array}\begin{array}{c} \begin{array}{cccc} (0,1) & (0,2) & (0,3) & \cdots \end{array} \\ \boxed{\begin{array}{cccc} \frac{\lambda_2}{\lambda_1+\lambda_2} & 0 & 0 & \cdots \\ \alpha_1^{BP} & \alpha_2^{BP} & \alpha_3^{BP} & \cdots \end{array}} \end{array}.$$

Let $\hat{\Psi}_{01}$ be the $1\times\infty$ absorbing distribution matrix from $\{(0,0)\}$ to $\{(0,l)\mid l\geq 1\}$. Using Lemma 1, we can calculate $\hat{\Psi}_{01}$ as:

$$\hat{\Psi}_{01}=[1\ 0]\,(I_{2\times 2}-\hat{\Gamma}_{0\to 0})^{-1}\hat{\Gamma}_{0\to 1^+}. \tag{15}$$

Then, we use conditional probability to calculate the transition probabilities for $\hat{L}_k^2=0$:

$$\hat{m}_{0\to\hat{L}_{k+1}^2}=\sum_{l=1}^{\hat{L}_{k+1}^2+1}\hat{m}_{l\to\hat{L}_{k+1}^2}P\left\{(0,l)\mid(0,0)\right\},\ \forall\hat{L}_{k+1}^2\geq 0, \tag{16}$$

in which $\hat{m}_{l\to\hat{L}_{k+1}^2}$ is given by (14), and $P\left\{(0,l)\mid(0,0)\right\}$ is the corresponding probability of absorption in $\{(0,l)\mid l\geq 1\}$ given in (15). Note that for the $(k+1)^{st}$ Class-2 departure to see $\hat{L}_{k+1}^2$ Class-2 jobs, l can be at most $\hat{L}_{k+1}^2+1$; thus $l\in\left[1,\hat{L}_{k+1}^2+1\right]$.

Using (14) and (16), we can write $\hat{m}_{0\to\hat{L}_{k+1}^2}$ for $\hat{L}_{k+1}^2\geq 0$ as the product of two matrices:

$$\hat{m}_{0\to\hat{L}_{k+1}^2}=\hat{\Psi}_{01}\begin{bmatrix}\alpha_{\hat{L}_{k+1}^2}^{\hat{D}_k} & \cdots & \alpha_1^{\hat{D}_k} & \alpha_0^{\hat{D}_k} & \mathbf{0}_{1\times\infty}\end{bmatrix}^T. \tag{17}$$

Note that (17) characterizes the $i=0$ row of $\hat{M}$ in (12). Thus, using (14) and (17), we obtain the transition matrix of the EMC in (12) as:

$$\hat{M}=\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ \vdots \end{array}\begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & \cdots & n & \cdots \end{array} \\ \begin{array}{|ccccccc|} \hline \hat{\Psi}_{01}\begin{bmatrix}\alpha_0^{\hat{D}_k}\\ \mathbf{0}_{\infty\times 1}\end{bmatrix} & \hat{\Psi}_{01}\begin{bmatrix}\alpha_1^{\hat{D}_k}\\ \alpha_0^{\hat{D}_k}\\ \mathbf{0}_{\infty\times 1}\end{bmatrix} & \hat{\Psi}_{01}\begin{bmatrix}\alpha_2^{\hat{D}_k}\\ \alpha_1^{\hat{D}_k}\\ \alpha_0^{\hat{D}_k}\\ \mathbf{0}_{\infty\times 1}\end{bmatrix} & \cdots & \hat{\Psi}_{01}\begin{bmatrix}\alpha_n^{\hat{D}_k}\\ \vdots\\ \alpha_1^{\hat{D}_k}\\ \alpha_0^{\hat{D}_k}\\ \mathbf{0}_{\infty\times 1}\end{bmatrix} & \cdots \\ \hline \alpha_0^{\hat{D}_k} & \alpha_1^{\hat{D}_k} & \alpha_2^{\hat{D}_k} & \cdots & \alpha_n^{\hat{D}_k} & \cdots \\ 0 & \alpha_0^{\hat{D}_k} & \alpha_1^{\hat{D}_k} & \cdots & \alpha_{n-1}^{\hat{D}_k} & \cdots \\ 0 & 0 & \alpha_0^{\hat{D}_k} & \ddots & \alpha_{n-2}^{\hat{D}_k} & \cdots \\ 0 & 0 & 0 & \vdots & \ddots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\ \hline \end{array} \end{array}. \tag{18}$$

4.2. Generating Function Approach

In this section we derive the steady state distribution of the EMC: $\hat{d}_n$, for $n \geq 0$. The equilibrium equations are given by $\left[\hat{d}_0, \hat{d}_1, \ldots\right] \hat{M} = \left[\hat{d}_0, \hat{d}_1, \ldots\right]$. Hence, from (18) we get

$$\hat{d}_n = (\left[\hat{d}_1, \hat{d}_2, \ldots\right] + \hat{d}_0 \hat{\Psi}_{01}) \left[\alpha_n^{\hat{D}_k} \ \cdots \ \alpha_1^{\hat{D}_k} \ \alpha_0^{\hat{D}_k} \ \mathbf{0}_{1\times\infty}\right]^T \quad \text{for } \forall n \geq 0. \tag{19}$$

Note that (19) has an infinite number of unknowns appearing in an (identical) infinite number of equations. To find these unknowns, we calculate the GF, as in the standard $M/G/1$ model (see e.g., Buzacott and Shanthikumar (1993), Section 3.3.2). Multiplying the n^{th} equation in (19) by z^n and summing over all n gives

$$G_{\hat{L}^2}(z) = (\left[\hat{d}_1, \hat{d}_2, \ldots\right] + \hat{d}_0 \hat{\Psi}_{01}) \sum_{n=0}^{\infty} \left[\alpha_n^{\hat{D}_k} \ \cdots \ \alpha_1^{\hat{D}_k} \ \alpha_0^{\hat{D}_k} \ \mathbf{0}_{1\times\infty}\right]^T z^n.$$

Let $G_{\alpha^{\hat{D}_k}}(z)$ be the GF of $\alpha^{\hat{D}_k}$ that can be calculated from (5) as: $G_{\alpha^{\hat{D}_k}}(z) = LT^{\hat{D}_k}(\lambda_2 - \lambda_2 z)$. Then, after some matrix algebra (see Appendix A1.1 for details), we get:

$$G_{\hat{L}^2}(z) = -\frac{\hat{d}_0}{(\lambda_1 + \lambda_2 - \alpha_0^{BP}\lambda_1)} \frac{(\lambda_1 + \lambda_2 - z\lambda_2 - \lambda_1 G_{\alpha^{BP}}(z)) G_{\alpha^{\hat{D}_k}}(z)}{z - G_{\alpha^{\hat{D}_k}}(z)}. \tag{20}$$

Note that, other than $\hat{d}_0$, all expressions in (20) are given in closed form. Therefore, all that is required to express $G_{\hat{L}^2}(z)$ in closed form is a closed-form expression for $\hat{d}_0$, which is derived next.

4.3. Finding the Idle Rate: $\hat{d}_0$

To obtain $\hat{d}_0$, we let $z \to 1$ in (20) and get (note that $z - G_{\alpha^{\hat{D}_k}}(z)$ is zero when $z \to 1$, so we need to apply L'Hopital's rule to calculate the limit on the right-hand side of (20)):

$$1 = -\frac{2\hat{d}_0}{\lambda_1 + \lambda_2 - \mu_1 + \sqrt{(\lambda_1 + \mu_1 + \lambda_2)^2 - 4\lambda_1\mu_1}} \frac{\lambda_2\mu_1\mu_2}{\lambda_1\mu_2 + \lambda_2\mu_1 - \mu_1\mu_2}, \tag{21}$$

solving which gives $\hat{d}_0$:

$$\hat{d}_0 = -\frac{\lambda_1\mu_2 + \lambda_2\mu_1 - \mu_1\mu_2}{2\lambda_2\mu_1\mu_2}(\lambda_1 + \lambda_2 - \mu_1 + \sqrt{(\lambda_1 + \lambda_2 + \mu_1)^2 - 4\lambda_1\mu_1}).$$

Substituting $\hat{d}_0$ in (20) gives us $G_{\hat{L}^2}(z)$ in closed form:

$$G_{\hat{L}^2}(z) = \frac{2(\lambda_1\mu_2 + \lambda_2\mu_1 - \mu_1\mu_2)}{\mu_2(\lambda_1 + \lambda_2 - \mu_1) + \lambda_2(2\mu_1 - \mu_2)z - \mu_2\sqrt{(\lambda_1 + \lambda_2 + \mu_1 - \lambda_2 z)^2 - 4\lambda_1\mu_1}}.$$

In a single-server queue, the service order in each priority class follows the FIFO rule, so we can use Distributional Little's Law (Bertsimas and Nakazato 1995) to get the LT of Class-2 jobs' sojourn time: $LT^{\hat{S}_2}(s) = G_{\hat{L}^2}(1 - \frac{s}{\lambda_2})$, which, of course, leads to (10).

5. General case: $c \geq 2$

The derivation of the general $c \geq 2$ servers case is very similar to the single-server case, but it is more complicated because Class-2 departures may see different numbers (i.e., $0, 1, \ldots, c-1$) of Class-1 jobs. Let (L_k^1, L_k^2) denote the state of the embedded Markov chain (EMC), i.e., L_k^1 and L_k^2 are the number of Class-1 and Class-2 jobs seen by the k^{th} Class-2 departure, respectively.

To display the one-dimensionally infinite transition matrix of the EMC for $c \geq 2$, we order the states: $\{(0,0), \ldots, (c-1,0), (0,1), \ldots, (c-1,1), \ldots, (0,n), \ldots, (c-1,n), \ldots\}$. Let Q_n be the set of states with $L_k^2 = n$ in the EMC, i.e., $Q_n = \{(0,n), \ldots, (c-1,n)\}$. When no confusion arises we also use Q_n to denote the set of states with $q_2 = n$ in the MC.

Using the ordering defined above, we specify the infinite dimensional transition matrix of the EMC, M. Let entry $m_{(L_k^1, L_k^2) \to (L_{k+1}^1, L_{k+1}^2)}$ be the probability that the $(k+1)^{st}$ Class-2 departure sees (L_{k+1}^1, L_{k+1}^2) given the k^{th} Class-2 departure left behind (L_k^1, L_k^2), and let $M_{i \to j}$ be the $c \times c$ transition matrix from Q_i to Q_j in the EMC. We illustrate $M_{i \to j}$ here:

$$M_{i \to j} = \begin{array}{c|cccc} & (0,j) & (1,j) & \cdots & (c-1,j) \\ \hline (0,i) & m_{(0,i)\to(0,j)} & m_{(0,i)\to(1,j)} & \cdots & m_{(0,i)\to(c-1,j)} \\ (1,i) & m_{(1,i)\to(0,j)} & m_{(1,i)\to(1,j)} & \cdots & m_{(1,i)\to(c-1,j)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (c-1,i) & m_{(c-1,i)\to(0,j)} & m_{(c-1,i)\to(1,j)} & \cdots & m_{(c-1,i)\to(c-1,j)} \end{array}. \tag{22}$$

Class-2 jobs are only served when there are fewer than c Class-1 jobs (i.e., $q_1 < c$) in the system, so the number of Class-1 jobs observed by the k^{th} Class-2 departure must be smaller than c, i.e., $L_k^1 = 0, 1, \ldots, c-1$. Similar to $\hat{D}_k$ and $\alpha^{\hat{D}_k}$ in Section 4, let D_k be the k^{th} inter-departure time of Class-2 jobs and α^{D_k} be the number of Class-2 arrivals during D_k. Analogous to (11):

$$L_{k+1}^2 = L_k^2 - 1 + \alpha^{D_k}. \tag{23}$$

The transition matrix M has the form illustrated in (24). Each row and column is labeled by the

corresponding set Q_i. Every block $M_{i\to j}$ is as illustrated in (22). Given (23), we have $M_{i\to j} = 0_{c\times c}$ for $j < i-1$, i.e., all blocks of the lower triangle below the row Q_1 in M are zero.

$$M = \begin{array}{c|ccccc|c} & Q_0 & Q_1 & Q_2 & Q_3 & Q_4 & \cdots \\ Q_0 & M_{0\to 0} & M_{0\to 1} & M_{0\to 2} & M_{0\to 3} & M_{0\to 4} & \cdots \\ Q_1 & M_{1\to 0} & M_{1\to 1} & M_{1\to 2} & M_{1\to 3} & M_{1\to 4} & \cdots \\ Q_2 & 0_{c\times c} & M_{2\to 1} & M_{2\to 2} & M_{2\to 3} & M_{2\to 4} & \cdots \\ Q_3 & 0_{c\times c} & 0_{c\times c} & M_{3\to 2} & M_{3\to 3} & M_{3\to 4} & \cdots \\ Q_4 & 0_{c\times c} & 0_{c\times c} & 0_{c\times c} & M_{4\to 3} & M_{4\to 4} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}. \tag{24}$$

For $i = 0, \ldots, c-1$ and $n \geq 0$, let $d_{in} = P\{(L^1, L^2) = (i,n)\} = \lim_{k\to\infty} P\{(L_k^1, L_k^2) = (i,n)\}$, so that (L^1, L^2) is the time-stationary limiting random variable of (L_k^1, L_k^2), and d_{in} is the steady state probability that a Class-2 job sees i Class-1 and n Class-2 jobs at departure.

Let $\vec{d_n} = (d_{0n}, \ldots, d_{(c-1)n})$: $\vec{d_n}$ is the $1 \times c$ row vector of steady state probabilities that the EMC is in Q_n. Let $\vec{d} = \left[\vec{d_0}\ \vec{d_1}\ \vec{d_2}\ \cdots\right]$, i.e., $\vec{d}$ is the $1 \times \infty$ row vector composed of $\vec{d_n}$, $n \geq 0$.

As in Section 4.1, we derive the transition matrix of the EMC based on the observation that the Class-2 jobs' service rate vector in (2) depends on Class-2 arrivals in D_k as follows:

- If $L_k^2 \geq c$: The Class-2 jobs' service rate vector remains $R_c = (c\mu_2, (c-1)\mu_2, \ldots, \mu_2)$ at least until the $(k+1)^{st}$ Class-2 departure, independent of Class-2 arrivals during D_k.
- If $L_k^2 = 1, \ldots, c-1$: The Class-2 jobs' service rate vector remains $R_{L_k^2}$ (as defined in (2)) until either the $(k+1)^{st}$ Class-2 departure or a Class-2 arrival. If there is a Class-2 arrival, this vector becomes $R_{L_k^2+1}$. (If this Class-2 arrival occurs during $BP_{L_k^2}$ together with other l Class-2 arrivals, then when the MC leaves $BP_{L_k^2}$, the service rate vector would be $R_{L_k^2+l+1}$, $l \geq 0$.)
- If $L_k^2 = 0$: The Class-2 jobs' service rate vector is $R_0 = (0, \ldots, 0)$, and remains R_0 until the next Class-2 arrival. It then becomes R_1 (or R_{l+1}, $l \geq 0$; see the discussion in previous bullet point).

We next demonstrate the derivation of M for $c = 2$. The $c > 2$ case can be analyzed similarly.

5.1. Transition Matrix of the EMC

In Section 4.1.1, we derived the Laplace transform of $\hat{D}_k$, $LT^{\hat{D}_k}$, expressed the distribution of $\alpha^{\hat{D}_k}$ using (4), and then wrote the transition probabilities of the EMC at the moment of the $(k+1)^{st}$ Class-2 departure using (23). We follow the same process here, for $c = 2$. We first derive LT^{D_k} when $L_k^2 \geq 2$, and then $L_k^2 = 1$, and finally $L_k^2 = 0$.

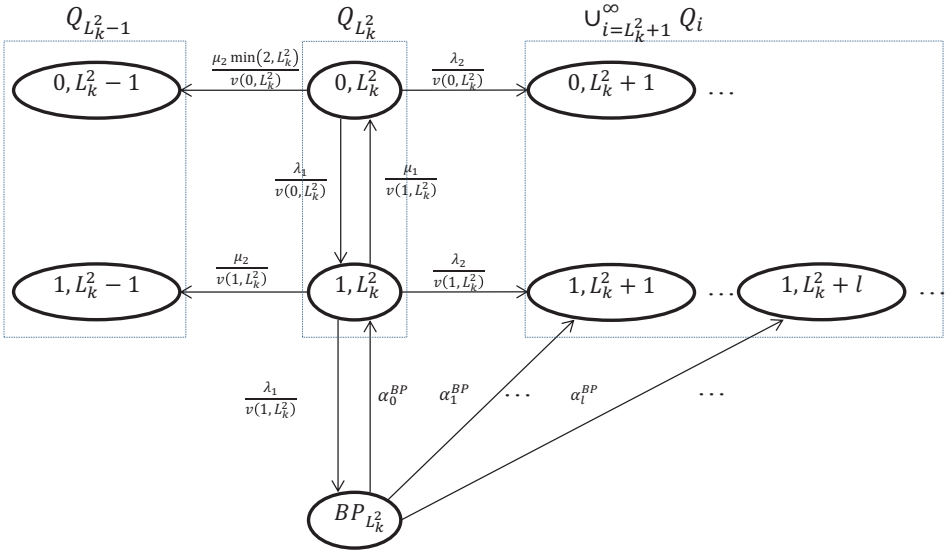

Figure 5 MC of the $c=2$ servers case where $L_k^2 \geq 1$ for Sections 5.1.1 and 5.1.2.

5.1.1. The Transition Probabilities for $L_k^2 \geq 2$ Since r_{q_1,q_2} depends on the number of Class-1 jobs in the network, D_k depends on the values of L_k^1 and L_{k+1}^1. For every $L_k^2 \geq 2$, there are four feasible combinations of L_k^1 and L_{k+1}^1: $0 \to 0$, $0 \to 1$, $1 \to 0$ and $1 \to 1$. Thus, we have 2^2 different inter-departure time distributions in the EMC. (For general $c > 2$ we have c^2 different inter-departure times when $L_k^2 \geq c$.)

Let $LT^{L_k^1, L_{k+1}^1}(s)$ be the LT of D_k conditioning on L_k^1 and L_{k+1}^1, given $L_k^2 \geq 2$ (we omit the latter dependency for notational convenience). For example, $LT^{00}(s)$ is the LT of D_k when the k^{th} and $(k+1)^{st}$ Class-2 departures see no Class-1 jobs in the network at their departures.

Figure 5 illustrates the service and arrival process of the Class-2 jobs in the MC after the k^{th} Class-2 departure when $L_k^2 \geq 1$, omitting details that are not relevant.

We next discuss the possible steps of the MC after the k^{th} Class-2 departure to express $LT^{00}(s)$, $LT^{01}(s)$, $LT^{10}(s)$, and $LT^{11}(s)$. Consider $LT^{10}(s)$ for example. The rate of exiting from state $(1, L_k^2)$ is $v(1, L_k^2) = \lambda_1 + \lambda_2 + \mu_1 + \mu_2$, thus after an $\exp(\lambda_1 + \lambda_2 + \mu_1 + \mu_2)$ distributed time, the MC would move to one of the following four states:

- State $BP_{L_k^2}$, w.p. $\frac{\lambda_1}{v(1, L_k^2)}$. Similar reasoning as in Section 4.1.1 gives: $LT^{10}(s) = \frac{\lambda_1+\lambda_2+\mu_1+\mu_2}{\lambda_1+\lambda_2+\mu_1+\mu_2+s} LT^{BP}(s) LT^{10}(s)$.

- State $(1, L_k^2+1)$, w.p. $\frac{\lambda_2}{v(1,L_k^2)}$. Similar reasoning gives: $LT^{10}(s) = \frac{\lambda_1+\lambda_2+\mu_1+\mu_2}{\lambda_1+\lambda_2+\mu_1+\mu_2+s} LT^{10}(s)$.
- State $(0, L_k^2)$, w.p. $\frac{\mu_1}{v(1,L_k^2)}$. From the memoryless property, the LT of the time from when the MC enters state $(0, L_k^2)$ until the next Class-2 departure occurs (with $L_{k+1}^1 = 0$) is $LT^{00}(s)$. Thus, w.p. $\frac{\mu_1}{v(1,L_k^2)}$, $LT^{10}(s)$ is $\frac{\lambda_1+\lambda_2+\mu_1+\mu_2}{\lambda_1+\lambda_2+\mu_1+\mu_2+s} LT^{00}(s)$.
- State $(1, L_k^2-1)$, w.p. $\frac{\mu_2}{v(1,L_k^2)}$. The next Class-2 departure occurs, but L_{k+1}^1 is not 0, so transition in the EMC from $L_k^1 = 1$ to $L_{k+1}^1 = 0$ is infeasible. Therefore, $LT^{10}(s) = 0$.

Using the Total Probability Theorem (see, e.g., Papoulis 1984) and multiplying by $\lambda_1 + \lambda_2 + \mu_1 + \mu_2 + s$, we get

$$(\lambda_1+\lambda_2+\mu_1+\mu_2+s)LT^{10}(s) = \lambda_1 LT^{BP}(s)LT^{10}(s) + \lambda_2 LT^{10}(s) + \mu_1 LT^{00}(s). \tag{25}$$

Using similar logic, we derive the following three additional equations:

$$(\lambda_1+\lambda_2+2\mu_2+s)LT^{00}(s) = \lambda_1 LT^{10}(s) + \lambda_2 LT^{00}(s) + 2\mu_2; \tag{26}$$

$$(\lambda_1+\lambda_2+2\mu_2+s)LT^{01}(s) = \lambda_1 LT^{11}(s) + \lambda_2 LT^{01}(s); \tag{27}$$

$$(\lambda_1+\lambda_2+\mu_1+\mu_2+s)LT^{11}(s) = \lambda_1 LT^{BP}(s)LT^{11}(s) + \lambda_2 LT^{11}(s) + \mu_1 LT^{01}(s) + \mu_2. \tag{28}$$

Thus, $(25-28)$ give four equations with four unknowns, which can be solved in closed form. Using $\Theta(s) = ((\lambda_1+2\mu_2+s)(\lambda_1+\mu_1+\mu_2+s-\lambda_1 LT^{BP}(s)) - \lambda_1\mu_1)^{-1}$, we get:

$$\begin{array}{ll} LT^{00}(s) = 2\mu_2(\lambda_1+\mu_1+\mu_2+s-\lambda_1 LT^{BP}(s))\Theta(s); & LT^{01}(s) = \lambda_1\mu_2\Theta(s); \\ LT^{11}(s) = \mu_2(\lambda_1+2\mu_2+s)\Theta(s); & LT^{10}(s) = 2\mu_1\mu_2\Theta(s). \end{array}$$

Let $\alpha_l^{L_k^1, L_{k+1}^1} = \frac{(-\lambda_2)^l}{l!} LT^{L_k^1, L_{k+1}^1\ (l)}(\lambda_2)$ be the probability of having l Class-2 arrivals in D_k that starts with L_k^1 and ends with L_{k+1}^1 Class-1 jobs. Then, using (4) and (23), we get, for $L_k^2 \geq 2$:

$$m_{(L_k^1, L_k^2)\to(L_{k+1}^1, L_{k+1}^2)} = \begin{cases} 0 & \text{if } L_{k+1}^2 < L_k^2 - 1 \\ \alpha_{L_{k+1}^2 - L_k^2 + 1}^{L_k^1, L_{k+1}^1} & \text{if } L_{k+1}^2 \geq L_k^2 - 1 \end{cases}. \tag{29}$$

Letting $A_l = \begin{bmatrix} \alpha_l^{00} & \alpha_l^{01} \\ \alpha_l^{10} & \alpha_l^{11} \end{bmatrix}$ be the 2×2 matrix of the probability that $\alpha^{D_k} = l$, as a function of the four different D_k, we get the matrices $M_{L_k^2 \to L_{k+1}^2}$ for $L_k^2 \geq 2$ and $L_{k+1}^2 \geq 0$:

$$M_{L_k^2 \to L_{k+1}^2} = \begin{cases} 0_{2\times 2} & \text{if } L_{k+1}^2 < L_k^2 - 1 \\ A_{L_{k+1}^2 - L_k^2 + 1} & \text{if } L_{k+1}^2 \geq L_k^2 - 1 \end{cases}. \tag{30}$$

Note that (30) characterizes the rows of M in (24) that correspond to any Q_i with $i \geq 2$.

5.1.2. The Transition Probabilities for $L_k^2 = 1$ Here, when the k^{th} Class-2 departure occurs, the MC moves into Q_1. Before the next Class-2 arrival or departure, there may be many Class-1 arrivals and departures, so the MC may move among states in $Q_1 \cup BP_1$. When the MC leaves $Q_1 \cup BP_1$, it may move to Q_0 (Class-2 departure) or to $\cup_{i=2}^{\infty} Q_i$ (Class-2 arrival). In both of these cases we can establish the conditional distribution of (L_{k+1}^1, L_{k+1}^2). Thus, we need to find the absorbing distribution matrices from Q_1 to Q_0 and $\cup_{i=2}^{\infty} Q_i$.

We again consider the MC after the k^{th} Class-2 departure as a MC with a transient set: $Q_1 \cup BP_1$, and absorbing sets: $\cup_{i=2}^{\infty} Q_i \cup Q_0$. In the MC, let $\Gamma_{1\to 1}$, $\Gamma_{1\to 0}$ and $\Gamma_{1\to 2^+}$ be the one-step transition matrices from $Q_1 \cup BP_1$ to $Q_1 \cup BP_1$, Q_0, and $\cup_{i=2}^{\infty} Q_i$, respectively.

From Figure 5, we can see that $\Gamma_{1\to 1}$, $\Gamma_{1\to 0}$ and $\Gamma_{1\to 2^+}$ are:

$$\Gamma_{1\to 1} = \begin{array}{c} \\ (0,1) \\ (1,1) \\ BP_1 \end{array} \begin{array}{c} \begin{array}{ccc} (0,1) & (1,1) & BP_1 \end{array} \\ \begin{bmatrix} 0 & \frac{\lambda_1}{v(0,1)} & 0 \\ \frac{\mu_1}{v(1,1)} & 0 & \frac{\lambda_1}{v(1,1)} \\ 0 & \alpha_0^{BP} & 0 \end{bmatrix} \end{array}, \quad \Gamma_{1\to 0} = \begin{array}{c} \\ (0,1) \\ (1,1) \\ BP_1 \end{array} \begin{array}{c} \begin{array}{cc} (0,0) & (1,0) \end{array} \\ \begin{bmatrix} \frac{\mu_2}{v(0,1)} & 0 \\ 0 & \frac{\mu_2}{v(1,1)} \\ 0 & 0 \end{bmatrix} \end{array},$$

$$\text{and } \Gamma_{1\to 2^+} = \begin{array}{c} \\ (0,1) \\ (1,1) \\ BP_1 \end{array} \begin{array}{c} \begin{array}{ccccc} (0,2) & (1,2) & (0,3) & (1,3) & \cdots \end{array} \\ \begin{bmatrix} \frac{\lambda_2}{v(0,1)} & 0 & 0 & 0 & \cdots \\ 0 & \frac{\lambda_2}{v(1,1)} & 0 & 0 & \cdots \\ 0 & \alpha_1^{BP} & 0 & \alpha_2^{BP} & \cdots \end{bmatrix} \end{array}.$$

We next discuss the possible steps of the MC, when it leaves the set $Q_1 \cup BP_1$.

- If the MC moves to Q_0, then the $(k+1)^{st}$ Class-2 departure happens before the next Class-2 arrival. Using Lemma 1, the probability of absorption in Q_0 (starting at Q_1) is

$$\Psi_{10} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \cdot (I_{3\times 3} - \Gamma_{1\to 1})^{-1} \Gamma_{1\to 0} = \frac{\mu_2 \begin{bmatrix} \lambda_1 + \lambda_2 + \mu_1 + \mu_2 - \alpha_0^{BP}\lambda_1 & \lambda_1 \\ \mu_1 & \lambda_1 + \lambda_2 + \mu_2 \end{bmatrix}}{(\lambda_1 + \lambda_2 + \mu_2 - \alpha_0^{BP}\lambda_1)(\lambda_1 + \lambda_2 + \mu_2) + \lambda_2\mu_1 + \mu_1\mu_2}. \quad (31)$$

At this absorption time the EMC moves into a state $(L_{k+1}^1, L_{k+1}^2) \in Q_0$. Thus, the transition matrix from Q_1 to Q_0 in the EMC, $M_{1\to 0}$, is Ψ_{10}.

- If the MC moves to $\cup_{i=2}^{\infty} Q_i$, then a Class-2 arrival happens before the $(k+1)^{st}$ Class-2 departure. (Again, this Class-2 arrival may have occurred during BP_1; the number of Class-2 arrivals during the BP_1 can be calculated from (8).) From Lemma 1, the absorbing distribution matrix from Q_1 to $\cup_{i=2}^{\infty} Q_i$ is

$$\Psi_{12} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \cdot (I_{3\times 3} - \Gamma_{1\to 1})^{-1} \Gamma_{1\to 2^+}. \quad (32)$$

After the MC enters states within $\cup_{i=2}^{\infty} Q_i$, the Class-2 jobs' service rate vector is identical to the one for $L_k^2 \geq 2$. Using the memoryless property, the distribution of (L_{k+1}^1, L_{k+1}^2) given the MC is in $\cup_{i=2}^{\infty} Q_i$ is identical to the distribution of (L_{k+1}^1, L_{k+1}^2) given $(L_k^1, L_k^2) \in \cup_{i=2}^{\infty} Q_i$, (29). Then, we use conditional probability to calculate transition probabilities of the EMC:

$$m_{(L_k^1,1)\to(L_{k+1}^1,L_{k+1}^2)} = \sum_{(q_1,q_2)\in\cup_{i=2}^{L_{k+1}^2+1} Q_i} m_{(q_1,q_2)\to(L_{k+1}^1,L_{k+1}^2)} P\left\{(q_1,q_2) \mid (L_k^1,1)\right\}, \tag{33}$$

in which $m_{(q_1,q_2)\to(L_{k+1}^1,L_{k+1}^2)}$ is given in (29) and $P\{(q_1,q_2) \mid (L_k^1,1)\}$ is the corresponding probability of absorption in $\cup_{i=2}^{\infty} Q_i$ given in (32). The upper bound of q_2 is L_{k+1}^2+1, because for the $(k+1)^{st}$ Class-2 departure to see L_{k+1}^2 Class-2 jobs, q_2 can be at most L_{k+1}^2+1. The lower bound of q_2 is 2, because (q_1,q_2) is in $\cup_{i=2}^{\infty} Q_i$.

From (31) and (33), we get matrices $M_{1\to L_{k+1}^2}$ for $L_{k+1}^2 \geq 0$, expressing the Q_1 row of M in (24)

$$M_{1\to L_{k+1}^2} = \begin{cases} \Psi_{10} & \text{if } L_{k+1}^2 = 0 \\ \Psi_{12}\left[A_{L_{k+1}^2-1}^T \ \cdots \ A_1^T \ A_0^T \ \mathbf{0}_{2\times\infty}\right]^T & \text{if } L_{k+1}^2 \geq 1 \end{cases}. \tag{34}$$

5.1.3. The Transition Probabilities for $L_k^2 = 0$ Using a similar analysis, we obtain the matrices $M_{0\to L_{k+1}^2}$ for $L_{k+1}^2 \geq 0$, characterizing the Q_0 row of M in (24) (see Appendix A2.1):

$$M_{0\to L_{k+1}^2} = \begin{cases} \Psi_{01}\Psi_{10} & \text{if } L_{k+1}^2 = 0 \\ (\Psi_{01}\Psi_{12}+\Psi_{02})\left[A_{n-1}^T \ \cdots \ A_1^T \ A_0^T \ \mathbf{0}_{2\times\infty}\right]^T & \text{if } L_{k+1}^2 \geq 1 \end{cases}. \tag{35}$$

Thus, using (30), (34) and (35), we obtain the transition matrix of the EMC in (24) as:

$$M = \begin{array}{c|c|c|c|c|c|} & Q_0 & Q_1 & Q_2 & \cdots & Q_n & \cdots \\ \hline Q_0 & \Psi_{01}\Psi_{10} & (\Psi_{01}\Psi_{12}+\Psi_{02})\begin{bmatrix} A_0 \\ \mathbf{0}_{\infty\times 2}\end{bmatrix} & (\Psi_{01}\Psi_{12}+\Psi_{02})\begin{bmatrix} A_1 \\ A_0 \\ \mathbf{0}_{\infty\times 2}\end{bmatrix} & \cdots & (\Psi_{01}\Psi_{12}+\Psi_{02})\begin{bmatrix} A_{n-1} \\ \vdots \\ A_1 \\ A_0 \\ \mathbf{0}_{\infty\times 2}\end{bmatrix} & \cdots \\ \hline Q_1 & \Psi_{10} & \Psi_{12}\begin{bmatrix} A_0 \\ \mathbf{0}_{\infty\times 2}\end{bmatrix} & \Psi_{12}\begin{bmatrix} A_1 \\ A_0 \\ \mathbf{0}_{\infty\times 2}\end{bmatrix} & \cdots & \Psi_{12}\begin{bmatrix} A_{n-1} \\ \vdots \\ A_1 \\ A_0 \\ \mathbf{0}_{\infty\times 2}\end{bmatrix} & \cdots \\ \hline Q_2 & 0 & A_0 & A_1 & \cdots & A_{n-1} & \cdots \\ \hline Q_3 & 0 & 0 & A_0 & \ddots & A_{n-2} & \cdots \\ \hline Q_4 & 0 & 0 & 0 & \vdots & \ddots & \cdots \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\ \hline \end{array}. \tag{36}$$

Using the transition matrix of the EMC, we can employ a similar exact analysis to the one in Sections 4.2 and 4.3 to obtain the closed-form expression of Laplace Transform of the Class-2 jobs'

sojourn time for $c=2$ case. However, the process becomes more cumbersome (see Appendix A3 for details). In the following section we focus on providing an efficient exact numerical method for the general $c \geq 2$ case.

6. Numerical Method

From the structure of the transition matrix of the EMC in (36), we see that it is an $M/G/1$-*type* Markov chain. Riska and Smirni (2002) gives an exact aggregate method to derive the steady state probability distribution of the MC and different moments of the number of Class-2 jobs in the system. As an example, we derive the first moment (see Algorithm 1 in Appendix A5). The main steps of this numerical procedure, which is the basis for our results in Section 7, are:

- Transform the 2D-infinite continuous-time MC (in Figure 1) into an $M/G/1$-type MC (with the transition matrix (24)) by: (i) using the Class-1 busy period to simplify the original MC to the MC in Figure 2; (ii) deriving the transition matrix of the EMC by observing the system state at Class-2 departures; deriving $M_{i\to j}$ $(1 \leq i \leq c+1)$ for three cases: $L_k^2 \geq c$, $L_k^2 = 0, 1, \ldots, c-1$ and $L_k^2 = 0$ as done in Section 5.1; and inserting $M_{i\to j}$ into M according to (23). The derivation of A_i in Section 5.1.1 becomes cumbersome as the number of servers c increases. We discuss the main difficulty and give an efficient exact numerical method to compute A_i in Appendix A2.2.
- Use Theorem 3.1, (18), and (21) from Riska and Smirni (2002) to derive the average number of Class-2 jobs in the system.

We next discuss the relation between steady state probability distributions of the original and embedded MCs, and the probability of no wait for Class-2 jobs. These quantities are important for our numerical results.

6.1. Relation between Original and Embedded Markov Chains

Let p_{ij} for $i, j = 0, 1, \ldots$ be the steady state probability distribution of the original MC. Recall that d_{ij} for $i = 0, \ldots, c-1$ and $j = 0, 1, \ldots$ is the steady state probability distribution of the embedded Markov chain (EMC). We show how to derive either distribution from the other.

We start by deriving d_{ij} using p_{ij}. For a Class-2 departure to leave state (i,j) behind (w.p. d_{ij}), there must be a Class-2 service completion at state $(i,j+1)$ with $i<c$, which happens with rate $\mu_2 \min(c-\min(i,c),j+1)$. Therefore, we have

LEMMA 2. *For the steady state probability distributions of both the original MC and the EMC, we have*

$$d_{ij}=\frac{p_{i(j+1)}\min(c-\min(i,c),j+1)}{\sum_{q_1=0}^{c-1}\sum_{q_2=0}^{\infty}p_{q_1q_2}\min(c-\min(q_1,c),q_2)}$$

for $i=0,\ldots,c-1$ *and* $j=0,1,\ldots$.

Note that d_{ij} for $i=0,\ldots,c-1$ and $j=0,1,\ldots$, is independent of p_{ij} for $i\geq c$ and $j=0,1,\ldots$.

Next, we derive p_{ij} from d_{ij}. From Poisson arrivals see time average (PASTA) and departures see what arrivals do, the probability of having no Class-2 jobs in the system in steady state is identical to the probability that a Class-2 departure sees no Class-2 jobs in the system:

$$\sum_{l=0}^{\infty}p_{l0}=\sum_{l=0}^{c-1}d_{l0}. \tag{37}$$

Using a similar discussion as the one used in the proof of Lemma 5 in Appendix A3, we can obtain $\frac{p_{i0}}{\sum_{l=0}^{\infty}p_{l0}}$, and thus express p_{i0} using (37). Specifically, from Figure 1, the balance equation of flow in and out of the set of states $\{(l,0)\mid l=0,1,\ldots\}$ is

$$\lambda_2\sum_{l=0}^{\infty}p_{l0}=\mu_2\sum_{l=0}^{c-1}p_{l1},$$

which gives $\sum_{l=0}^{c-1}p_{l1}=\frac{\lambda_2}{\mu_2}\sum_{l=0}^{c-1}d_{l0}$. Further, from Lemma 2, we have

$$\frac{p_{i1}}{\sum_{l=0}^{c-1}p_{l1}}=\frac{d_{i0}}{\min(c-\min(i,c),1)\sum_{l=0}^{c-1}\frac{d_{l0}}{\min(c-\min(l,c),1)}}.$$

Thus, we obtain p_{i1} for $i=0,\ldots,c-1$. In a similar fashion, we can derive p_{ij}, for $i=0,\ldots,c-1$ and $j=2,3,\ldots$.

Once p_{ij} for $i=0,\ldots,c-1$ are derived, p_{ij} for $i\geq c$ can be calculated by solving balance equations of the flows into and out of state (i,j), for $i=c,c+1,\ldots$ and $j=1,2,\ldots$. However, the results in Section 7 do not require p_{ij} for $i\geq c$. Thus, we do not further discuss their calculation.

6.2. Probability of No Wait for Class-2 jobs

Here we use p_{ij} to calculate the probability of no wait for Class-2 jobs. For a Class-2 job's waiting time to be zero, it should (i) arrive when there is at least one idle server, and (ii) not be preempted by Class-1 jobs.

Say a Class-2 job arrives at state (i,j) (w.p., p_{ij}), for $i+j<c$, i.e., there are $c-i-j$ servers available. We call this Class-2 job the *tagged* Class-2 job. Due to the first come first serve rule, the chance of this tagged Class-2 job being preempted is independent of future Class-2 arrivals.

This tagged Class-2 job's service process, until its service completion or it is preempted by Class-1 jobs, can be represented by the MC in Figure 6. The state of this MC represent the number of Class-1 jobs, and the number of Class-2 jobs, including the tagged Class-2 job, when the tagged job arrives. For example, consider the case when the tagged Class-2 job arrived at state $(0,c-1)$ sending the system into state $(0,c)$. If a Class-1 job arrives (w.p. $\frac{\lambda_1}{\lambda_1+c\mu_2}$) at this state, the tagged Class-2 job will be preempted; otherwise, if a Class-2 job finishes service (w.p. $\frac{c\mu_2}{\lambda_1+c\mu_2}$), the system moves to state $(0,c-1)$, then the tagged Class-2 job will not be preempted unless the number of Class-1 jobs reaches 2. The states on the southeast border of the MC represent the tagged job being preempted. The states on the west border represent it finishing before being preempted.

Thus, the probability that a tagged Class-2 job (which arrives at state $(i,j-1)$) finishes service without being preempted is the probability that the MC in Figure 6, starting from state (i,j), is absorbed on the west border. This probability can be derived by applying Lemma 1 on the MC. Then, using the Total Probability Theorem (see, e.g., Papoulis 1984), the probability of no wait for Class-2 jobs is

$$P\{W_2=0\}=\sum_{i=0}^{c-1}\sum_{j=0}^{c-i-1} p_{ij}\cdot P\{\text{the tagged Class-2 job is not preempted before being served}\}. \tag{38}$$

7. Numerical Results and Extensions

We run Algorithm 1 on a 64-bit desktop with an Intel Quad Core i5-2400 @ 3.10GHz processor. For $c\le 10$, it completes within 1 second. The processing time of Algorithm 1 increases with c; for

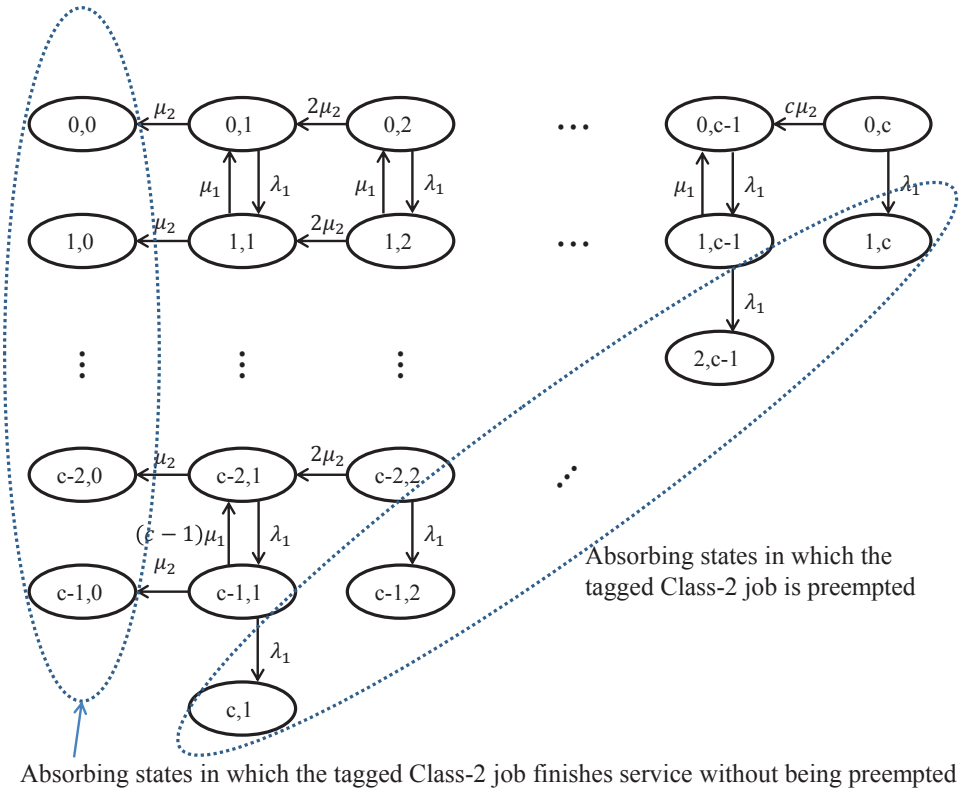

Figure 6 The Markov chain for the absorption of the tagged Class-2 job in either preemption or service completion.

$c = 50$, it takes 109 seconds. Additional details on the running times are available upon request.

Potential inaccuracies in Algorithm 1 arise from two sources. The first is that α_l^{BP} requires numerical inversion of the probability GF. Abate and Whitt (1992) give an efficient inversion algorithm with a controllable error bound. The second source is the limited storage space on any computer, so it is not practical to store an infinite number of matrices. Thus, we derive A_i for i up to $Limit = \min\{i \mid \max(A_i) \leq Tolerance\}$ where the A_is are given in $(A5)$ in Appendix A2.2. Both inaccuracy sources can be well controlled by using an accuracy tolerance 10^{-8}.

We validate Algorithm 1 in two cases where exact results are available: when $c = 2$ (based on the exact derivation in Appendix A3), and when $\mu_1 = \mu_2$ (see, e.g., Buzen and Bondi 1983). In total, we examined 280 different parameter settings for validation; all relative errors were less than 0.001%, significantly outperforming the approximation in Harchol-Balter et al. (2005), which is to our knowledge the best approximation, with a relative error within 2% compared to simulation.

Given the accuracy of Algorithm 1, we next use it to answer the three questions raised in Section

1. Then, we apply our methodology to the problem in Maglaras and Zeevi (2004) when Class-1 jobs are infinitely impatient (i.e., they leave the system if upon arrival there is no available server) by replacing the Class-1 BP in our model with a Class-1 jobs' exponential service time. Throughout this section, we use $\lambda_i = c\rho_i\mu_i$ for $i = 1, 2$, so that $\rho_1 + \rho_2 < 1$ is each server's occupation rate in the $M/M/c$ queue. Thus, once c, ρ_1, ρ_2, μ_1 and μ_2 are given, the system is determined.

7.1. Insight 1 - How Changing μ_1 or μ_2 Affects $E[S_2]$

Consider a company that operates an $M/M/2$ system to serve two priority classes where Class-1 has preemptive priority over Class-2. The company receives complaints of long sojourn times from Class-2 customers. In this section, we answer the question: When the manager is able to improve the service rate of one priority class, which service rate should she improve?

Any Class-2 customer's sojourn time is dictated by its interaction with customers of both types. All Class-1 customers present during a Class-2 customer's sojourn time may affect it, while only those Class-2 customers present when the customer arrives can affect her sojourn time. Increasing μ_1 reduces the Class-1 interference, while increasing μ_2 reduces the Class-2 interference, as well as the customer's own service time. Which of these effects dominates (and which service rate is thus preferable to improve) depends on the relation between λ_1 and λ_2.

Figure 7 illustrates the effect of improving μ_1 or μ_2 on $E[S_2]$ in different parameter settings. The solid lines show how $E[S_2]$ changes when improving μ_2 while keeping $\mu_1 = 1$ and the dashed lines show the effect of improving μ_1 while fixing $\mu_2 = 1$. In Figure 7(a) Class-1's workload is lower than Class-2's ($\lambda_1 = 0.8$ and $\lambda_2 = 1.1$), which is common in practice. In this case, upgrading μ_2 is more effective. In contrast, when $\lambda_1 = 1.1$ and $\lambda_2 = 0.8$, Figure 7(b), it is better to improve μ_1: When Class-2 customers complain about long sojourn times, it is better to improve the service rate of the *other* class. This case may occur in settings such as the contact center example, where answering phone calls comprises a larger portion of the total workload, compared to answering emails. Finally, when $\lambda_1 = 1$ and $\lambda_2 = 0.9$, Figure 7(c), we see that it is better to improve μ_1 if the maximum service rate the company can achieve is below 3.5; otherwise it is better to improve μ_2.

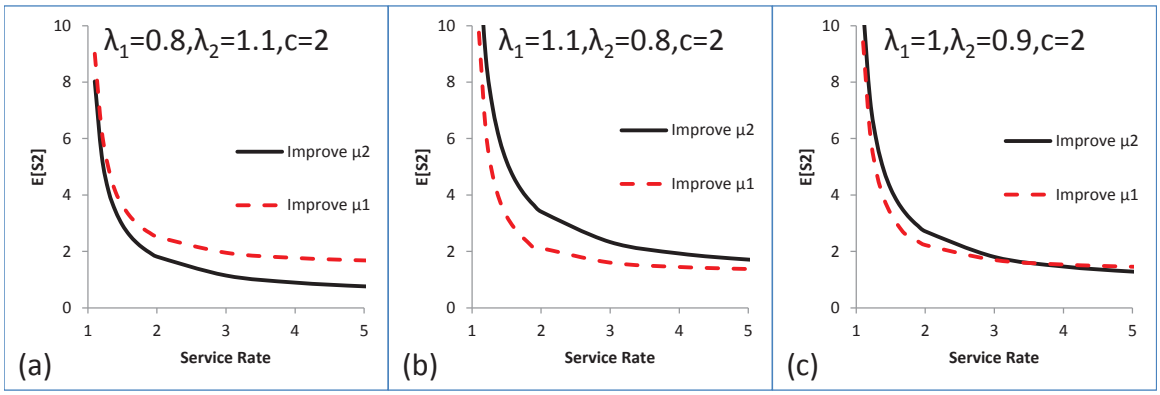

Figure 7 The effect of improving $\mu_i, i=1,2$ on $E[R_2]$ under different combinations of λ_i for $i=1,2$.

Next, we examine how the number of servers may affect the manager's decision. In Figures 8 (a) and (b) we keep the initial service and workload for both classes the same (i.e. $\mu_1=\mu_2=1$, $\rho_1=0.55$, and $\rho_2=0.4$), and change c (i.e., $c=1$ in 8(a), $c=2$ in 7(b), and $c=3$ in 8(b). When c increases both curves move downward, but the solid curve (improving μ_2) moves faster than the dashed curve (improving μ_1). When $c=3$, these two curves cross, and for $c\geq 4$ (not shown here), the solid curve is below the dashed one. However, this phenomenon does not hold for the $\rho_1=0.4$ and $\rho_2=0.55$ case (in Figure 7(a)): the solid curve is already below the dashed one when $c=2$, and increasing c only increases the gap between them. Thus, managers cannot decide on which service rate to improve simply by approximating an $M/M/c$ system as an $M/M/1$, as different c values lead to different answers. This insight holds for different combinations of λ_1 and λ_2.

Still, a simple rule of thumb is: If the conclusion from the $M/M/1$ system is to improve μ_2, then the manager can go ahead and implement it. In contrast, if the conclusion from the $M/M/1$ system is to improve μ_1, the manager needs to examine Class-2 jobs' sojourn time carefully for the $M/M/c$ system, because the number of servers affects this decision.

7.2. Insight 2 - Few Fast Servers vs. Many Slow Servers

In this section we compare systems with different numbers of servers, while keeping the arrival rates λ_i and the occupation rates ρ_i $(i=1,2)$ the same (i.e., we increase c and reduce μ_i while holding $c\mu_i=\frac{\lambda_i}{\rho_i}$ constant). That is we investigate the effect of having many slow servers compared with

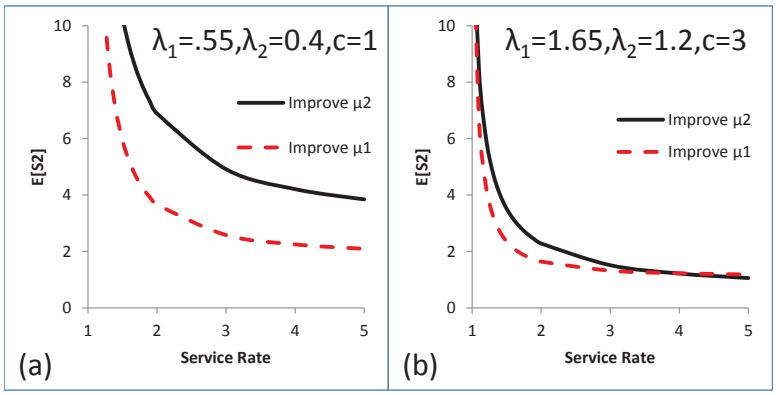

Figure 8 The effect of improving $\mu_i, i=1,2$ on $E[R_2]$ under different numbers of servers.

having fewer fast servers. We use $\rho_1=\rho_2=0.475$. (Similar result holds for different combinations of ρ_1 and ρ_2. However, the smaller $\frac{\rho_1}{\rho_1+\rho_2}$ is, the less obvious the result becomes.)

In Figure 9, we fix $\lambda_1=1$ and illustrate the effect of having more slow servers on the expected sojourn times of both classes, under different λ_2's. When comparing figures, note that since we keep $\frac{\lambda_2}{c\mu_2}=\rho_2=0.475$, for the same c, a smaller λ_2 results in smaller μ_2 and vice-versa. Also, within each figure as c increases, both μ_1 and μ_2 decrease.

We see that in most cases jobs prefer fewer fast servers. Morse (1958) observes that the optimal number of servers for a single class $M/M/c$ queue is one, thus Class-1 jobs prefer one fast server. But the number of servers affects $E[S_2]$ in different ways for different values of λ_2. When $\lambda_2=\frac{1}{3}$, Class-2 sojourn times increase faster than Class-1 jobs' as c increases, but when $\lambda_2=3$, the opposite is true. There are two competing effects here: On the one hand, reducing μ_2 increases Class-2 sojourn times due to Class-2 service time. On the other hand, higher c increases Class-2 jobs' access to servers, reducing the effect of preemption. When $\lambda_2=1$, these two effects balance and the sojourn times of both classes increase with c at similar rates. When Class-2 jobs are short (e.g., $\lambda_2=5$), the increased access is more beneficial as they are more likely to finish before being interrupted.

Another observation from Figures 9 (c) is that when $\lambda_1=1$ and $\lambda_2=3$, Class-2 jobs' average sojourn time may decrease with c, when c is small. This trend is more obvious in Figure 9 (d) when $\lambda_1=1$ and $\lambda_2=5$: $E[S_2]$ decreases by about 5% (10.4 vs. 9.9) when c increases from 2 to 15. In

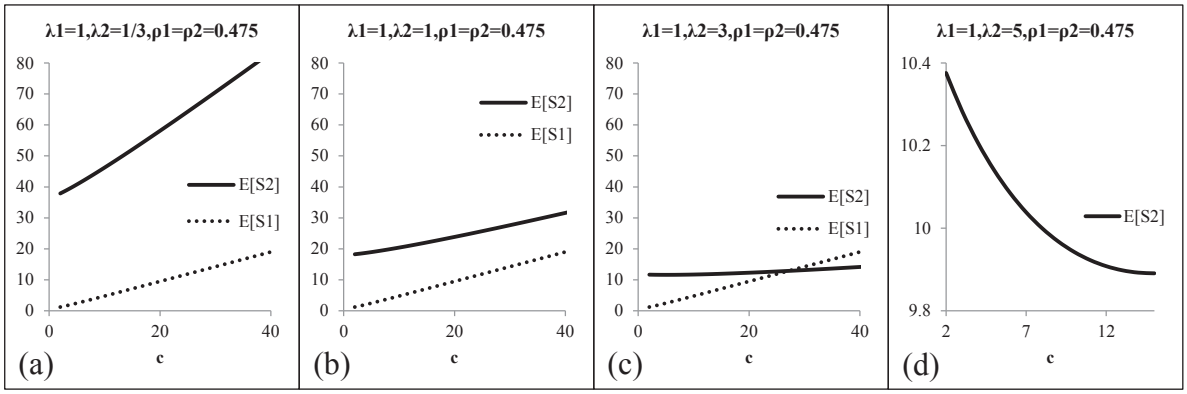

Figure 9 The effect of c on expected sojourn times of both priority classes, under different $\lambda_i, i = 1, 2$.

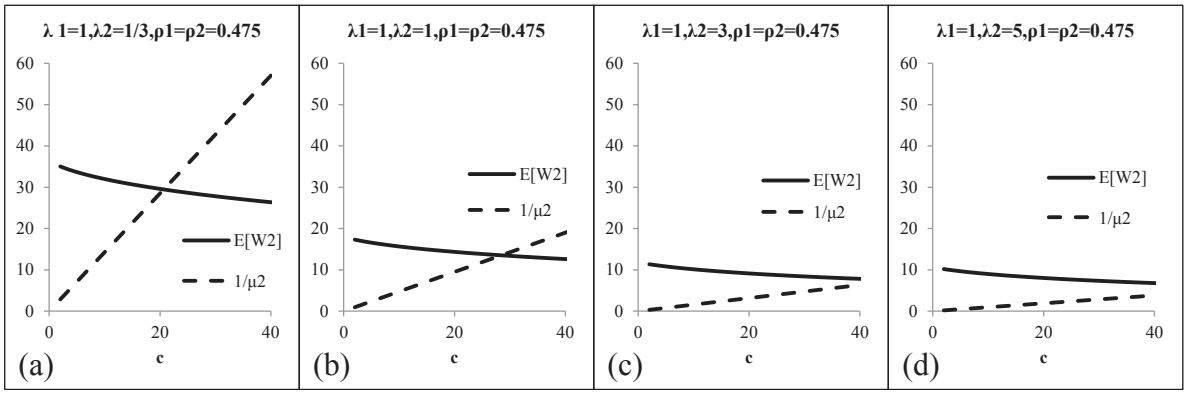

Figure 10 The effect of c on expected waiting time of Class-2 jobs, under different $\lambda_i, i = 1, 2$.

this case, the benefit of improved access to servers for Class-2 jobs dominates the negative effect of decreasing μ_2. Similar result has been shown in Wierman et al. (2006) by approximation. Our results, based on exact analysis, sharpen and provide validation of theirs.

To further investigate the different effects of increasing the number of servers, we decompose $E[S_2] = E[W_2] + \frac{1}{\mu_2}$ in Figure 10 (a-d) for the same four cases. Of course Class-2 jobs' expected service time increases linearly with c in all four cases, but with different slopes. As $\frac{1}{\mu_2} = \frac{\rho_2}{\lambda_2} c$, Class-2 jobs' expected service time increase slowly when λ_2 is large, and vice-versa. At the same time, $E[W_2]$ decreases at a similar speed in all four cases. Combining these changes, the decrease in $E[W_2]$ becomes greater than the increase in the mean service time in the $\lambda_2 = 5$ case.

To investigate how likely it is that Class-2 jobs will not preempted, we look at the probability of no wait $P\{W_2 = 0\}$ in (38), and the probability of no wait given they see at least one idle

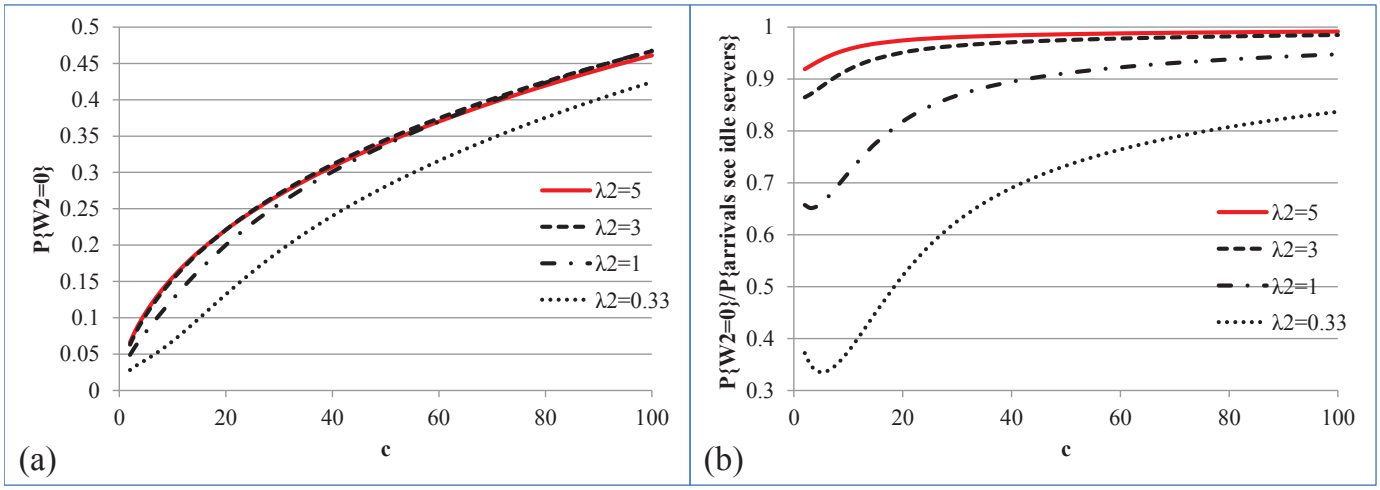

Figure 11 $P\{W_2=0\}$ and $P\{W_2=0\}/\sum_{i=0}^{c-1}\sum_{j=0}^{c-i-1}p_{ij}$ as a function of c, for $\lambda_1=1$, $\rho_1=\rho_2=0.475$.

server at arrival, i.e., $P\{W_2=0\}/\sum_{i=0}^{c-1}\sum_{j=0}^{c-i-1}p_{ij}$. Figure 11 (a-b) illustrates these two quantities, respectively, as functions of c, for the same four cases in Figure 10.

From Figure 11(a), we observe that for any $c=1,\ldots,100$, $P\{W_2=0\}$ does not change much when λ_2 increases from 1/3 to 5. However, $P\{W_2=0\}/\sum_{i=0}^{c-1}\sum_{j=0}^{c-i-1}p_{ij}$ changes more dramatically, and we suspect this change causes $E[W_2]$ to decrease relatively faster to $\frac{1}{\mu_2}$ in the $\lambda_2=5$ case than in the $\lambda_2=1/3$ case. In Figure 11(b), when $\lambda_2=5$, more than 90% of Class-2 jobs *that see at least one idle server at arrival* finish service without being preempted. However, when $\lambda_2=1/3$, only 50% of those Class-2 jobs are not preempted (when $c=20$). The probability even *decreases* by 5% when c increases from 1 to 6. Thus, as λ_2 and μ_2 increases, Class-2 jobs suffer less preemption and $E[S_2]$ is lower.

7.3. Insight 3 - Square Root Staffing Rule

The square root staffing rule has been widely studied in the literature (see, e.g., Whitt 1992, and reference therein). The square root staffing rule suggests increasing the staffing level, c, relative to ρ according to $\rho=1-\frac{\gamma}{\sqrt{c}}$, where γ is a rough service grade indicator, to keep service level measures approximately the same.

In this section we investigate whether the square root staffing rule holds in the $M/M/c$ preemptive priority queue. Specifically, we consider a series of queueing systems (indexed with the

number of servers, $c = 1, 2, \ldots$) with the following parameters: the number of servers c, fixed service rates $\mu_1^c = \mu_1$ and $\mu_2^c = \mu_2$, a total workload $\rho_1^c + \rho_2^c = 1 - \frac{\gamma}{\sqrt{c}}$, and a fixed ratio of workload $w = \frac{\rho_1^c}{\rho_1^c + \rho_2^c}$, for $c = 1, 2, \ldots$. We demonstrate numerically that when $c \to \infty$, the limits of $P\{W_2^c > 0\}$, $\sqrt{c}E[W_2^c | W_2^c > 0]$, and $\sqrt{c}E[W_2^c]$ exist, which is a new result. From $E[W_2] = P\{W_2 > 0\} E[W_2 | W_2 > 0]$, we know that if either two of the above three limits exist, the other limit does as well.

First, we consider $E[W_2]$. In the special case of $\mu_1 = \mu_2 = \mu$, the overall mean waiting time for both priority classes would remain the same if the scheduling discipline were changed to First-Come-First-Serve (see, e.g., Buzen and Bondi 1983). Moreover, with regard to the total average waiting time for *all* customers, the square root staffing rule holds in a First-Come-First-Serve system with workload $\rho_1^c + \rho_2^c = 1 - \frac{\gamma}{\sqrt{c}}$, for $c = 1, 2, \ldots$, i.e., $\lim_{c\to\infty} \sqrt{c}\left(wE[W_1^c] + (1-w)E[W_2^c]\right) = \frac{\alpha}{\gamma\mu}$, where $E[W_i^c] = E[S_i^c] - \frac{1}{\mu}$ for $i = 1, 2$. Due to the preemptive priority, Class-1 jobs face a classic $M/M/c$ queue. Following the above rules of choosing parameters, we have $\rho_1^c = w\left(1 - \frac{\gamma}{\sqrt{c}}\right)$, so that $\lim_{c\to\infty} \sqrt{c}E[W_1^c] = 0$. Thus, $\lim_{c\to\infty} \sqrt{c}E[W_2^c] = \frac{\alpha}{\gamma\mu(1-w)}$ for the $\mu_1 = \mu_2$ case.

However, it is not clear whether the square root staffing rule still holds when $\mu_1 \neq \mu_2$. To explore this, we test the case of $\mu_1 = 1$, $\mu_2 = 2$, $\gamma = 1$ for three different combinations of workload: 1) $w = 0.2$; 2) $w = 0.5$; 3) $w = 0.8$. As illustrated by Figure 12(a), $\lim_{c\to\infty} \sqrt{c}E[W_2^c]$ seems to exist in all three cases and the rate of convergence is high. Moreover, it can be verified from Figure 12(b) that the step difference of $\sqrt{c}E[W_2^c]$ (i.e., $\sqrt{c+1}E\left[W_2^{c+1}\right] - \sqrt{c}E[W_2^c]$) converges to zero faster than $\frac{1}{c}$. This result suggests that $\lim_{c\to\infty} \sqrt{c}E[W_2^c]$ exists.

Numerical results suggest that the square root staffing rule holds for $P\{W_2 > 0\}$ and $E[W_2 | W_2 > 0]$ for different combinations of μ_1, μ_2, γ, and w as well (these can be derived using (38)). Due to the page limit, we do not include them here.

In view of these results, we conjecture that in the preemptive-resume $M/M/c$ queue, the square root staffing rule holds for Class-2 jobs' performance measures: $P\{W_2 > 0\}$, $E[W_2 | W_2 > 0]$, and $E[W_2]$. In practice, using the square root staffing rule can provide supreme service to Class-1 jobs while maintaining a specified service level for Class-2 jobs and keeping the utilization of all servers

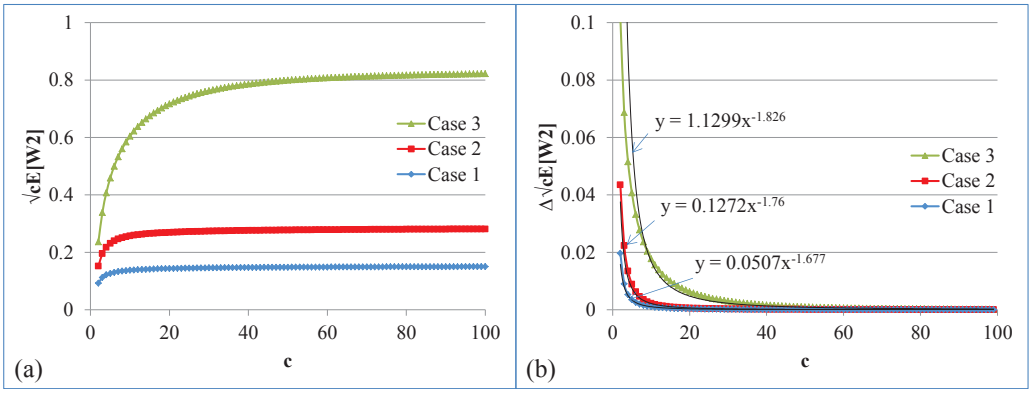

Figure 12 To test if square root safety capacity rule hold for $E[W_2]$, when $\mu_1 = 1$, $\mu_2 = 2$, $\gamma = 1$ and 1) $\pi = 0.2$; 2) $\pi = 0.5$; 3) $\pi = 0.8$.

close to one. This result is similar to the one in Pang and Perry (2014), who consider "call blending" where inbound calls are prioritized over outbound calls with infinite supply. They prove that a logarithmic safety staffing rule holds. Thus, it is possible to answer all inbound calls immediately, maintain a certain throughput rate of outbound calls, and keep all servers almost fully utilized. (Their logarithmic, rather than a square root, safety staffing rule works because the infinite supply is used to reduce demand variability.)

7.4. Extension to Impatient Class-1 Jobs

Maglaras and Zeevi (2004) considered an $M/M/c$ queue with two priority classes where the first class is completely impatient, i.e., if not served at arrival, they leave the system. They applied diffusion approximations to the problem in the asymptotic Halfin and Whitt (1981) regime. Our methodology can be applied to this system by replacing the Class-1 BP in our model with the $\exp(c\mu_1)$ busy periods caused by a Class-1 job that brings the number of Class-1 jobs in the system to c. Therefore, we can obtain a closed-form expression of the GF of L^2 when $c = 2$; and we have an efficient numerical algorithm to calculate the distribution of L^2 when $c \geq 2$.

Table 1 illustrates the accuracy of the two approximations (2D diffusion and perturbation) in Maglaras and Zeevi (2004) and of our Algorithm 1, compared with simulation under different settings in their paper. For simulation, 2D diffusion and perturbation approximations, we generate $E[L^1 + L^2]$ from their results. (Unfortunately, the confidence intervals of their simulation were not

provided.) For our algorithm, we generate $E[L^1+L^2]$ from the sum of $E[L^1]$, obtained using a single-class $M/M/c/c$ model (page 81, Gross et al. 2008), and $E[L^2]$, obtained using Algorithm 1. The results of our algorithm are typically closer to the simulation than their two approximations; in fact, since our algorithm is so accurate, errors must be due to inaccuracy of the simulation. Furthermore, Maglaras and Zeevi's approximations are only accurate in the Halfin and Whitt regime, i.e., for high ρ and c, whereas our method is accurate for all combinations of ρ and c. However, the computational burden of our algorithm increases with c: When $c=150$, our algorithm takes 30 minutes.

	Simulation	2D Diffusion		Perturbation		Our Algorithm	
(c,ρ,μ_1,μ_2)	$E\left[L^1+L^2\right]$	$E[L^1+L^2]$	$\%Error$	$E[L^1+L^2]$	$\%Error$	$E[L^1+L^2]$	$\%Error$
(100,0.95,1,2)	108.22	109.28	1.0%	112.34	3.8%	108.42	0.2%
(50,0.95,1,2)	65.71	64.88	1.3%	66.82	1.7%	64.57	1.7%
(150,0.95,1,2)	154.49	154.30	0.1%	158.63	2.7%	153.58	0.6%
(100,0.925,1,2)	99.64	98.02	1.6%	102.50	2.9%	98.13	1.5%
(100,0.975,1,2)	140.36	136.57	2.7%	139.77	0.4%	138.42	1.4%
(100,0.95,1,5)	120.46	120.02	0.4%	119.43	0.9%	118.97	1.2%
(100,0.95,2,1)	102.74	103.24	0.5%	111.39	8.4%	102.60	0.1%
(100,0.95,5,1)	101.49	101.16	0.3%	115.83	14.1%	101.31	0.2%
(100,0.95,20,10)	103.44	103.39	0.1%	112.27	8.5%	102.60	0.8%

Table 1 2D Diffusion and Perturbation in Maglaras & Zeevi 2004 v.s. Algorithm 1 in terms of $E[L^1+L^2]$ for different settings with $\rho_1=\rho_2=\frac{\rho}{2}$.

8. Summary

This paper analyzed an $M/M/c$ queue with two preemptive-resume priority classes. This problem is usually described by a 2-dimension infinite MC, representing the two class state space. We introduced a technique to reduce this 2D-infinite MC into a 1D-infinite MC, from which the Generating Function (GF) of the number of low-priority jobs can be derived in closed form. We demonstrate this methodology for the $c=1,2$ cases. When $c>2$, the closed-form expression of the GF becomes cumbersome. We thus derive an exact numerical algorithm to calculate different moments of the number of Class-2 jobs in the system for any $c\geq 2$.

We use our algorithm to generate the following insights: First, for a company serving two priority classes and receiving complaints of long sojourn times from Class-2 customers, we provide

guidelines on when the manager should improve the service rate of either customer class. Second, we demonstrated that unlike a single-class system, Class-2 jobs may prefer many slow servers to a few fast servers. Third, we numerically validated the existence of the square root staffing rule for Class-2 jobs in an $M/M/c$ queue with preemptive priority. Finally, we applied our methodology to the problem considered by Maglaras and Zeevi (2004).

For future research, it would be beneficial to extend our methodology to more than two priority classes, though this appears to be quite challenging. As priority queues have a direct application in information and communication services, it would be interesting to incorporate pricing and system design into the model and try to maximize profit.

Acknowledgement: The authors are grateful to Professor Costis Maglaras for providing data for comparison, and to Professor Hossein Abouee-Mehrizi for his help with the numerical method to express α_l^{BP}.

References

Abate, J., W. Whitt. 1992. Numerical Inversion of Probability Generating Functions. *Operations Research Letters* **12**(4) 245-251.

Abouee-Mehrizi, H., B., Balcıoglu, O. Baron. 2012. Strategies for a Centralized Single Product Multi-Class M/G/1 Make-to-Stock Queue. *Oper. Res.* **60**(4) 803-812.

Afèche, P., M. Araghi, O. Baron. 2012. Customer Acquisition, Retention, and Service Quality for a Call Center: Optimal Promotions, Priorities, and Staffing. *Working paper.*

Bertsimas, D., D. Nakazato. 1995. The distributional Little's Law and its applications. *Oper. Res.* **43**(2) 298-310.

Buzacott, J., J. Shanthikumar. 1993. Stochastic Models of Manufacturing Systems. *Prentice Hall.*

Buzen J., A. Bondi. 1983. The response times of priority classes under preemptive resume in $M/M/m$ queues. *Oper. Res.* **31**, 456-465.

Davis, R. 1966. Waiting-time distribution of a multi-server, priority queueing system. *Oper. Res. 14* (1) 133-136.

Green, D., D. Knuth. 1990. *Mathematics for the Analysis of Algorithms*, 3rd ed. Birkhāuser Boston.

Gross, D., J. Shortle, J. Thompson, C. Harris. 2008. *Fundamentals of Queueing Theory.* Wiley & Sons.

Halfin, S., W. Whitt. 1981. Heavy-traffic Limits for Queues with Many Exponential Servers. *Oper. Res.* **29**(3) 567-588.

Harchol-Balter, M., T. Osogami, A. Scheller-Wolf, A. Wierman. 2005. Multi-server queueing systems with multiple priority classes. *Queueing Systems* **51**(3) 331–360.

Jeffrey, A. 2005. *Complex Analysis and Applications*. 2nd ed. CRC.

Kella O., U. Yechiali. 1985. Waiting time in the non-preemptive priority $M/M/c$ queue. *Stochastic Models* **1**(2) 257-262.

Maglaras, C., A. Zeevi. 2004. Diffusion Approximations for a Multiclass Markovian Service System with "Guaranteed" and "Best-Effort" Service Levels. *Math. of Operations Research* **29**(4) 786-813.

Maglaras, C., A. Zeevi. 2005. Pricing and Design of Differentiated Services: Approximate Analysis and Structural Insights. *Oper. Res.* **53**(2) 242-262.

Miller, D. 1981. Computation of Steady-State Probabilities for $M/M/1$ Priority Queues. *Oper. Res.* 29(5) 945-958.

Morse, P. 1958. *Queues, Inventories, and Maintenance*, John Wiley and Sons.

Pang, G., O. Perry. 2014. A Logarithmic Safety Staffing Rule for Contact Centers with Call Blending. *Management Sci. forthcoming.*

Papoulis, A. 1984. *Probability, Random Variables, and Stochastic Processes*, 2nd ed. New York: McGraw-Hill, pp.37-38

Riska, A., E. Smirni. 2002. Exact Aggregate Solutions for M/G/1-type Markov Processes, *SIGMETRICS* 86-96.

Takagi, H. 1991. *Queueing analysis: a foundation of performance evaluation.* Amsterdam, North-Holland.

Van Mieghem, J. 1995. Dynamic Scheduling with Convex Delay Costs: The Generalized $c\mu$ Rule. *The Annals of Applied Probability* **5**(3) 808-833.

Whitt, W. 1992. Understanding the Efficiency of Multi-server Service Systems. *Management Sci.* 38(5) 708-723.

Wierman, A., M. Harchol-Balter, T. Osogami, A. Scheller-Wolf. 2006. How Many Servers are Best in a Dual-Priority M/PH/k system. *Performance Evaluation* **63**(12) 1253-1272.

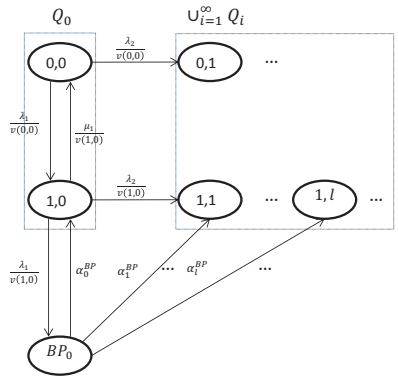

Figure A1 MC for the $c=2$ servers case where $L_k^2=0$.

Appendix

A1. Calculations

A1.1. Calculation for $G_{\hat{L}^2}(z)$ The following equation will be used in the calculation for $G_{\hat{L}^2}(z)$. The derivation is straightforward, so we skip all the details.

$$\sum_{n=0}^{\infty}\left[\alpha_n^{\hat{D}_k}\ \cdots\ \alpha_1^{\hat{D}_k}\ \alpha_0^{\hat{D}_k}\ \mathbf{0}_{1\times\infty}\right]^T z^n=\left[1\ z\ z^2\ z^3\ \cdots\right]^T G_{\alpha^{\hat{D}_k}}(z). \tag{A1}$$

With the help of (A1), we derive $G_{\hat{L}^2}(z)$:

$$\begin{aligned}
G_{\hat{L}^2}(z) &= (\left[\hat{d}_1,\hat{d}_2,...\right]+\hat{d}_0\hat{\Psi}_{01})\sum_{n=0}^{\infty}\left[\alpha_n^{\hat{D}_k}\ \cdots\ \alpha_1^{\hat{D}_k}\ \alpha_0^{\hat{D}_k}\ \mathbf{0}_{1\times\infty}\right]^T z^n,\\
G_{\hat{L}^2}(z) &= \left[\hat{d}_1,\hat{d}_2,...\right]\left[1\ z\ z^2\ z^3\ \cdots\right]^T G_{\alpha^{\hat{D}_k}}(z)+\hat{d}_0\hat{\Psi}_{01}\left[1\ z\ z^2\ z^3\ \cdots\right]^T G_{\alpha^{\hat{D}_k}}(z),\\
G_{\hat{L}^2}(z) &= \frac{G_{\hat{L}^2}(z)-\hat{d}_0}{z}G_{\alpha^{\hat{D}_k}}(z)+\frac{\hat{d}_0}{z}\frac{z\lambda_2+\lambda_1 G_{\alpha^{BP}}(z)-\alpha_0^{BP}\lambda_1}{\lambda_1+\lambda_2-\alpha_0^{BP}\lambda_1}G_{\alpha^{\hat{D}_k}}(z).
\end{aligned}$$

Solving for $G_{\hat{L}^2}(z)$ leads to (20).

A2. Transition Probabilities

A2.1. The Transition Probabilities for $L_k^2=0$ when $c=2$ As in Section 4.1.2, to find the one-step transition probabilities of the EMC, we first express the first-passage probability distribution from Q_0 to $\cup_{i=1}^{\infty}Q_i$.

We think of the MC after the k^{th} Class-2 departure as a MC with transient set: $Q_0\cup BP_0$, and absorbing sets: Q_1 and $\cup_{i=2}^{\infty}Q_i$. (Defining Q_1 and $\cup_{i=2}^{\infty}Q_i$ instead of $\cup_{i=1}^{\infty}Q_i$ is for computational

convenience.) Let $\Gamma_{0\to0}$, $\Gamma_{0\to1}$ and $\Gamma_{0\to2^+}$ be the one-step transition matrices from $Q_0 \cup BP_0$ to $Q_0 \cup BP_0$, Q_1 and $\cup_{i=2}^{\infty} Q_i$, respectively.

In Figure A1, we illustrate the arrival process of Class-2 jobs omitting details that are not relevant to the development of this case. From Figure A1, we get $\Gamma_{0\to0}$, $\Gamma_{0\to1}$ and $\Gamma_{0\to2^+}$:

$$\Gamma_{0\to0} = \begin{array}{c} \\ (0,0) \\ (1,0) \\ BP_0 \end{array} \begin{array}{c} \begin{array}{ccc} (0,0) & (1,0) & BP_0 \end{array} \\ \begin{bmatrix} 0 & \frac{\lambda_1}{v(0,0)} & 0 \\ \frac{\mu_1}{v(1,0)} & 0 & \frac{\lambda_1}{v(1,0)} \\ 0 & \alpha_0^{BP} & 0 \end{bmatrix} \end{array}, \quad \Gamma_{0\to1} = \begin{array}{c} \\ (0,0) \\ (1,0) \\ BP_0 \end{array} \begin{array}{c} \begin{array}{cc} (0,1) & (1,1) \end{array} \\ \begin{bmatrix} \frac{\lambda_2}{v(0,0)} & 0 \\ 0 & \frac{\lambda_2}{v(1,0)} \\ 0 & \alpha_1^{BP} \end{bmatrix} \end{array},$$

$$\text{and } \Gamma_{0\to2^+} = \begin{array}{c} \\ (0,0) \\ (1,0) \\ BP_0 \end{array} \begin{array}{c} \begin{array}{ccccc} (0,2) & (1,2) & (0,3) & (1,3) & \cdots \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & \cdots \\ 0 & \alpha_2^{BP} & 0 & \alpha_3^{BP} & \cdots \end{bmatrix} \end{array}.$$

Let Ψ_{01} be the absorbing distribution matrix from Q_0 to Q_1. Let Ψ_{02} be the absorbing distribution matrix from Q_0 to $\cup_{i=2}^{\infty} Q_i$. Using Lemma 1, we calculate Ψ_{01} and Ψ_{02} as:

$$\Psi_{01} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \cdot (I_{3\times3} - \Gamma_{0\to0})^{-1}\Gamma_{0\to1} = \frac{\begin{bmatrix} \lambda_2(\lambda_1+\lambda_2+\mu_1-\alpha_0^{BP}\lambda_1) & \lambda_1(\lambda_2+\alpha_1^{BP}\lambda_1) \\ \lambda_2\mu_1 & (\lambda_1+\lambda_2)(\lambda_2+\alpha_1^{BP}\lambda_1) \end{bmatrix}}{\lambda_1^2+\lambda_2^2-\alpha_0^{BP}\lambda_1^2+2\lambda_1\lambda_2+\lambda_2\mu_1-\alpha_0^{BP}\lambda_1\lambda_2}, \quad \text{(A2)}$$

and

$$\Psi_{02} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \cdot (I_{3\times3} - \Gamma_{0\to0})^{-1}\Gamma_{0\to2^+}. \quad \text{(A3)}$$

When the MC goes to $\cup_{i=1}^{\infty} Q_i$, there are one or more Class-2 jobs in the system and there are no transitions in the EMC. As in Section 4.1.2, we use conditional probability to calculate transition probabilities of the EMC:

$$m_{(L_k^1,0)\to(L_{k+1}^1,L_{k+1}^2)} = \sum_{(q_1,q_2)\in\cup_{i=1}^{L_{k+1}^2+1} Q_i} m_{(q_1,q_2)\to(L_{k+1}^1,L_{k+1}^2)} P\left\{(q_1,q_2) \mid (L_k^1,0)\right\}, \quad \text{(A4)}$$

in which $m_{(q_1,q_2)\to(L_{k+1}^1,L_{k+1}^2)}$ is given in (33) and $P\{(q_1,q_2) \mid (L_k^1,0)\}$ is the corresponding probability of absorption in Q_1 or $\cup_{i=2}^{\infty} Q_i$ given in $(A2)$ and $(A3)$ respectively. Similar to (33), we must have $q_2 \in \left[1, L_{k+1}^2+1\right]$.

From $(A4)$, we get the matrices $M_{0\to L_{k+1}^2}$ in (35) for $L_{k+1}^2 \geq 0$.

A2.2. Numerical Method for The Transition Probabilities for $L_k^2 \geq c$ when $c = 2$

Deriving A_i in (30), the probability of $i = 0, 1, ...$ Class-2 arrivals during different inter-departure times, is numerically complex because: (i) It is time-consuming to derive the LTs for c^2 different D_k, depending on c^2 different combinations of L_k^1 and L_{k+1}^1 – key steps in expressing the transition matrix for $L_k^2 \geq c$; (ii) The derivation of $\alpha_l^{L_k^1, L_{k+1}^1}$ using (4) and the LTs of D_k's is cumbersome. We next develop an efficient numerical algorithm to calculate A_i. Then, the techniques in Subsections 5.1.2 and 5.1.3 can be used to derive the transition matrix of the embedded Markov chain (EMC) for $L_k^2 \geq c$.

We demonstrate the algorithm for calculating A_i by deriving the transition probabilities of the EMC for $L_k^2 \geq c = 2$. The general case with $c > 2$ is similar.

As in Section 5.1.2, we first think of the MC after the k^{th} Class-2 departure as a MC with transient set: $Q_{L_k^2} \cup BP_{L_k^2}$, and absorbing sets: $Q_{L_k^2 - 1}$ and $\cup_{i=L_k^2+1}^{\infty} Q_i$. Let $\Gamma_{2\to 2}$, $\Gamma_{2\to 1}$ and $\Gamma_{2\to 3^+}$ be the one-step transition matrices from $Q_{L_k^2} \cup BP_{L_k^2}$ to $Q_{L_k^2} \cup BP_{L_k^2}$, $Q_{L_k^2-1}$ and $\cup_{i=L_k^2+1}^{\infty} Q_i$, respectively. From Figure 5, we get $\Gamma_{2\to 2}$, $\Gamma_{2\to 1}$ and $\Gamma_{2\to 3^+}$:

$$\Gamma_{2\to 2} = \begin{array}{c} \\ (0, L_k^2) \\ (1, L_k^2) \\ BP_{L_k^2} \end{array} \begin{array}{c} \begin{array}{ccc} (0, L_k^2) & (1, L_k^2) & BP_{L_k^2} \end{array} \\ \begin{bmatrix} 0 & \frac{\lambda_1}{v(0, L_k^2)} & 0 \\ \frac{\mu_1}{v(1, L_k^2)} & 0 & \frac{\lambda_1}{v(1, L_k^2)} \\ 0 & \alpha_0^{BP} & 0 \end{bmatrix} \end{array}, \quad \Gamma_{2\to 1} = \begin{array}{c} \\ (0, L_k^2) \\ (1, L_k^2) \\ BP_{L_k^2} \end{array} \begin{array}{c} \begin{array}{cc} (0, L_k^2 - 1) & (1, L_k^2 - 1) \end{array} \\ \begin{bmatrix} \frac{2\mu_2}{v(0, L_k^2)} & 0 \\ 0 & \frac{\mu_2}{v(1, L_k^2)} \\ 0 & 0 \end{bmatrix} \end{array},$$

$$\text{and } \Gamma_{2\to 3^+} = \begin{array}{c} \\ (0, L_k^2) \\ (1, L_k^2) \\ BP_{L_k^2} \end{array} \begin{array}{c} \begin{array}{ccccc} (0, L_k^2 + 1) & (1, L_k^2 + 1) & (0, L_k^2 + 2) & (1, L_k^2 + 2) & \cdots \end{array} \\ \begin{bmatrix} \frac{\lambda_2}{v(0, L_k^2)} & 0 & 0 & 0 & \cdots \\ 0 & \frac{\lambda_2}{v(1, L_k^2)} & 0 & 0 & \cdots \\ 0 & \alpha_1^{BP} & 0 & \alpha_2^{BP} & \cdots \end{bmatrix} \end{array}.$$

Then, with similar reasoning as in Section 5.1.2, we calculate A_i from:

$$A_i = \begin{cases} \Psi_{21} & \text{for } i = 0 \\ \Psi_{23^+} \left[A_{i-1}^T \cdots A_1^T \ A_0^T \ \mathbf{0}_{2\times\infty} \right]^T & \text{for } i \geq 1 \end{cases}, \tag{A5}$$

where

$$\Psi_{21} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \cdot (I - \Gamma_{2\to 2})^{-1} \Gamma_{2\to 1} = \frac{\begin{bmatrix} 2\mu_2(\lambda_1 + \lambda_2 + \mu_1 + \mu_2 - \alpha_0^{BP}\lambda_1) & \lambda_1\mu_2 \\ 2\mu_1\mu_2 & \mu_2(\lambda_1 + \lambda_2 + 2\mu_2) \end{bmatrix}}{(\lambda_1 + \lambda_2 + \mu_1 + \mu_2 - \alpha_0^{BP}\lambda_1)(\lambda_1 + \lambda_2 + 2\mu_2) - \lambda_1\mu_1}. \tag{A6}$$

and

$$\Psi_{23^+} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \cdot (I - \Gamma_{2\to 2})^{-1} \Gamma_{2\to 3^+}. \tag{A7}$$

Notice that A_i only depends on $A_0, A_1, \ldots, A_{i-1}$. Thus, A_i can be calculated recursively from A_0 (which is Ψ_{21} in $(A6)$).

Once we get A_i, we obtain the rows of M in (24) that correspond to any Q_i with $i \geq 2$ numerically. Then, using (34) and (35), we compute the transition matrix in (24). Since we cannot practically store an infinite number of matrices, we derive A_i for i up to $Limit = \min\{i \mid \max(A_i) \leq Tolerance\}$ using (A5). These matrices accurately capture the behavior of the entire system when the $Tolerance$ is small enough.

A3. Derivation of $G_{L^2}(z)$ in Closed Form

Let $G_{(i,L^2)}(z) = \sum_{n=0}^{\infty} d_{in} z^n$ be the GF of L^2 when $L^1 = i$, i.e., of the joint event $L^2 = n$ and $L^1 = i$, for $i = 0, 1, \ldots, c-1$. So $\sum_{n=0}^{\infty} \vec{d_n} z^n = \left[G_{(0,L^2)}(z), \ldots, G_{(c-1,L^2)}(z)\right]$ is the $1 \times c$ row vector of GF of L^2 for $L^1 = 0, 1, \ldots, c-1$.

Note that a Class-2 departure can only see $0, \ldots, c-1$ Class-1 jobs, so once we get $G_{(i,L^2)}(z)$, $0 \leq i \leq c-1$, using the total probability theorem (see, e.g., Papoulis 1984), we have the GF of the number of Class-2 jobs at Class-2 departures:

$$G_{L^2}(z) = \sum_{i=0}^{c-1} G_{(i,L^2)}(z). \tag{A8}$$

A3.1. Generating Function Approach We now derive the steady state distribution of the EMC for the case of $c = 2$. Recalling that $\vec{d}$ is the row vector of the steady state distribution of the EMC, the equilibrium equations are given by $\vec{d} \cdot M = \vec{d}$, so from (36)

$$\vec{d_n} = \begin{cases} (\vec{d_1} + \vec{d_0}\Psi_{01})\Psi_{10} & \text{if } n = 0 \\ (\left[\vec{d_2}, \vec{d_3}, ...\right] + \vec{d_1}\Psi_{12} + \vec{d_0}(\Psi_{01}\Psi_{12} + \Psi_{02}))\left[A_{n-1}^T \ \cdots \ A_1^T \ A_0^T \ \mathbf{0}_{2\times\infty}\right]^T & \text{if } n \geq 1 \end{cases}. \tag{A9}$$

Note that $(A9)$, just as (19), has an infinite number of unknowns appearing in an infinite (identical) number of equations. To find these unknowns, we calculate the GF as in the standard $M/G/1$ queue. Multiplying the n^{th} equation in $(A9)$ by z^n and summing over all n:

$$[G_{(0,L^2)}(z), G_{(1,L^2)}(z)]$$

$$= \vec{d_0} + (\left[\vec{d_2}, \vec{d_3}, \ldots\right] + \vec{d_1}\Psi_{12} + \vec{d_0}(\Psi_{01}\Psi_{12} + \Psi_{02}))\sum_{n=1}^{\infty}\left[A_{n-1}^T \ \cdots \ A_1^T \ A_0^T \ \mathbf{0}_{2\times\infty}\right]^T z^n.$$

With some matrix calculations (see Appendix A3.2 for details), we get:

$$\left[G_{(0,L^2)}(z), G_{(1,L^2)}(z)\right] = \vec{d_0}D(z), \tag{A10}$$

where $D(z)$ is given in closed form in Appendix A3.2.

Therefore, if we can express $\vec{d_0}$ in closed form as well, we could use $(A10)$ to express $\left[G_{(0,L^2)}(z), G_{(1,L^2)}(z)\right]$ in closed form. Then, we get the GF of L^2:

$$G_{L^2}(z) = G_{(0,L^2)}(z) + G_{(1,L^2)}(z). \tag{A11}$$

If we further assume that the service order in each priority class follows the FIFO rule, we can use the Distributional Little's Law (Bertsimas and Nakazato 1995) to get the LT of Class-2 jobs' sojourn time:

$$LT^{S_2}(s) = G_{(0,L^2)}(1 - \frac{s}{\lambda_2}) + G_{(1,L^2)}(1 - \frac{s}{\lambda_2}).$$

The next two sections are devoted to deriving $\vec{d_0}$.

A3.2. Calculation for $G_{L^2}(z)$ The following results will be used in the calculation for $D(z)$. The derivation of them is straightforward, so we skip all the details.

$$\sum_{i=1}^{\infty}\left[\ A_{n-1}^T \ \cdots \ A_1^T \ A_0^T \ \mathbf{0}_{1\times\infty}\ \right]^T z^i = z\Upsilon G_A, \tag{A12}$$

$$\text{in which } \Upsilon = \left[\ I_{2\times2} \ zI_{2\times2} \ z^2I_{2\times2} \ z^3I_{2\times2} \ \cdots\right]^T \text{ and } G_A = \begin{bmatrix} G_{\alpha^{00}}(z) & G_{\alpha^{01}}(z) \\ G_{\alpha^{10}}(z) & G_{\alpha^{11}}(z) \end{bmatrix}.$$

$$\left[\vec{d_2}, \vec{d_3}, \ldots\right] z\Upsilon = \frac{1}{z}(G_{L^2}(z) - \vec{d_0} - \vec{d_1}z) \tag{A13}$$

$$\vec{d_1} = \vec{d_0}(\Psi_{10}^{-1} - \Psi_{01}) \tag{A14}$$

$$\begin{aligned} z^2\Psi_{10}^{-1}\Psi_{12}\Upsilon &= \frac{z^2\Psi_{10}^{-1}\begin{bmatrix} \lambda_2(\lambda_1+\lambda_2+\mu_1+\mu_2-\alpha_0^B\lambda_1) & \frac{1}{z}\lambda_1(z\lambda_2+\lambda_1G_{\alpha^B}(z)-\alpha_0^B\lambda_1) \\ \lambda_2\mu_1 & \frac{1}{z}(\lambda_1+\lambda_2+\mu_2)(z\lambda_2+\lambda_1G_{\alpha^B}(z)-\alpha_0^B\lambda_1) \end{bmatrix}}{(\lambda_1+\lambda_2+\mu_2)(\lambda_1+\lambda_2+\mu_2-\lambda_1\alpha_0^B)+\mu_1(\lambda_2+\mu_2)} \\ &= \begin{bmatrix} z^2\frac{\lambda_2}{\mu_2} & 0 \\ 0 & \frac{z}{\mu_2}(z\lambda_2+\lambda_1G_{\alpha^B}(z)-\alpha_0^B\lambda_1) \end{bmatrix} \end{aligned} \tag{A15}$$

$$\Psi_{02}\Upsilon = \begin{bmatrix} 0 & \frac{1}{z^2}\frac{\lambda_1^2(G_{\alpha^B}(z)-\alpha_0^B-z\alpha_1^B)}{\lambda_1^2+\lambda_2^2-\alpha_0^B\lambda_1^2+2\lambda_1\lambda_2+\lambda_2\mu_1-\alpha_0^B\lambda_1\lambda_2} \\ 0 & \frac{1}{z^2}\frac{\lambda_1(\lambda_1+\lambda_2)(G_{\alpha^B}(z)-\alpha_0^B-z\alpha_1^B)}{\lambda_1^2+\lambda_2^2-\alpha_0^B\lambda_1^2+2\lambda_1\lambda_2+\lambda_2\mu_1-\alpha_0^B\lambda_1\lambda_2} \end{bmatrix} \tag{A16}$$

With the help of these results, we derive $D(z)$:

$$\left[G_{(0,L^2)}(z), G_{(1,L^2)}(z)\right] = \vec{d_0} + (\left[\vec{d_2}, \vec{d_3}, ...\right] + \vec{d_1}\Psi_{12} + \vec{d_0}(\Psi_{01}\Psi_{12} + \Psi_{02}))\sum_{n=1}^{\infty}\left[A_{n-1}^T \ \cdots \ A_1^T \ A_0^T \ \mathbf{0}_{1\times\infty}\right]^T z^n.$$

From $(A12)$, we have

$$\left[G_{(0,L^2)}(z), G_{(1,L^2)}(z)\right] = \vec{d_0} + z\left\{\left[\vec{d_2}, \vec{d_3}, ...\right] + \vec{d_1}\Psi_{12} + \vec{d_0}(\Psi_{01}\Psi_{12} + \Psi_{02})\right\}\Upsilon G_A.$$

From $(A13)$, we have

$$[G_{(0,L^2)}(z), G_{(1,L^2)}(z)] = \vec{d_0} + \frac{1}{z}([G_{(0,L^2)}(z), G_{(1,L^2)}(z)] - \vec{d_0} - \vec{d_1}z)G_A + z(\vec{d_1}\Psi_{12} + \vec{d_0}(\Psi_{01}\Psi_{12} + \Psi_{02}))\Upsilon G_A.$$

Moving $\left[G_{(0,L^2)}(z), G_{(1,L^2)}(z)\right]$ to the left side of the equation gives

$$\left[G_{(0,L^2)}(z), G_{(1,L^2)}(z)\right](zI_{2\times 2} - G_A) = \vec{d_0}(z^2(\Psi_{01}\Psi_{12} + \Psi_{02^+})\Upsilon G_A - G_A + zI_{2\times 2}) + \vec{d_1}(z^2\Psi_{12}\Upsilon G_A - zG_A).$$

From $(A14)$, we have

$$\left[G_{(0,L^2)}(z), G_{(1,L^2)}(z)\right](zI_{2\times 2} - G_A) = \vec{d_0}\left\{(z^2\Psi_{10}^{-1}\Psi_{12}\Upsilon + z^2\Psi_{02}\Upsilon - I_{2\times 2} - z(\Psi_{10}^{-1} - \Psi_{01}))G_A + zI_{2\times 2}\right\}.$$

From $(A15)$ and $(A16)$, we have

$$\begin{aligned} &\left[G_{(0,L^2)}(z), G_{(1,L^2)}(z)\right](zI_{2\times 2} - G_A) \\ &= \vec{d_0}\left(\begin{bmatrix} z^2\frac{\lambda_2}{\mu_2} - 1 & \lambda_1^2\frac{(G_{\alpha^B}(z)-\alpha_0^B-z\alpha_1^B)}{\lambda_1^2+\lambda_2^2-\alpha_0^B\lambda_1^2+2\lambda_1\lambda_2+\lambda_2\mu_1-\alpha_0^B\lambda_1\lambda_2} \\ 0 & \frac{z}{\mu_2}(z\lambda_2 + \lambda_1 G_{\alpha^B}(z) - \alpha_0^B\lambda_1) \\ & + \frac{\lambda_1(\lambda_1+\lambda_2)(G_{\alpha^B}(z)-\alpha_0^B-z\alpha_1^B)}{\lambda_1^2+\lambda_2^2-\alpha_0^B\lambda_1^2+2\lambda_1\lambda_2+\lambda_2\mu_1-\alpha_0^B\lambda_1\lambda_2} - 1 \end{bmatrix} G_A - z(\Psi_{10}^{-1} - \Psi_{01})G_A + zI_{2\times 2}\right). \end{aligned}$$

We know, $(zI_{2\times 2} - G_A)^{-1} = \frac{\begin{bmatrix} G_{\alpha^{11}}(z) - z & -G_{\alpha^{01}}(z) \\ -G_{\alpha^{10}}(z) & G_{\alpha^{00}}(z) - z \end{bmatrix}}{zG_{\alpha^{00}}(z) + zG_{\alpha^{11}}(z) - G_{\alpha^{00}}(z)G_{\alpha^{11}}(z) + G_{\alpha^{01}}(z)G_{\alpha^{10}}(z) - z^2}$, so we have:

$$D(z) = \frac{\left\{\begin{array}{c} \begin{bmatrix} z^2\frac{\lambda_2}{\mu_2} - 1 & \frac{\lambda_1^2(G_{\alpha^{BP}}(z)-\alpha_0^{BP}-z\alpha_1^{BP})}{\lambda_1^2+\lambda_2^2-\alpha_0^{BP}\lambda_1^2+2\lambda_1\lambda_2+\lambda_2\mu_1-\alpha_0^{BP}\lambda_1\lambda_2} \\ 0 & \frac{z}{\mu_2}(z\lambda_2 + \lambda_1 G_{\alpha^{BP}}(z) - \alpha_0^{BP}\lambda_1) \\ & + \frac{\lambda_1(\lambda_1+\lambda_2)(G_{\alpha^{BP}}(z)-\alpha_0^{BP}-z\alpha_1^{BP})}{\lambda_1^2+\lambda_2^2-\alpha_0^{BP}\lambda_1^2+2\lambda_1\lambda_2+\lambda_2\mu_1-\alpha_0^{BP}\lambda_1\lambda_2} - 1 \end{bmatrix}\mathscr{C}(z) - z(\Psi_{10}^{-1} - \Psi_{01})\mathscr{C}(z) \\ + z\begin{bmatrix} -(z - G_{\alpha^{11}}(z)) & -G_{\alpha^{01}}(z) \\ -G_{\alpha^{10}}(z) & -(z - G_{\alpha^{00}}(z)) \end{bmatrix} \end{array}\right\}}{zG_{\alpha^{00}}(z) + zG_{\alpha^{11}}(z) - G_{\alpha^{00}}(z)G_{\alpha^{11}}(z) + G_{\alpha^{01}}(z)G_{\alpha^{10}}(z) - z^2}, \tag{A17}$$

in which

$$\mathscr{C}(z)=\begin{bmatrix} G_{\alpha^{00}}(z) & G_{\alpha^{01}}(z) \\ G_{\alpha^{10}}(z) & G_{\alpha^{11}}(z) \end{bmatrix}\begin{bmatrix} G_{\alpha^{11}}(z)-z & -G_{\alpha^{01}}(z) \\ -G_{\alpha^{10}}(z) & G_{\alpha^{00}}(z)-z \end{bmatrix}.$$

Ψ_{10}^{-1} can be calculated from (31) as $\Psi_{10}^{-1}=\frac{1}{\mu_2}\begin{bmatrix} \lambda_1+\lambda_2+\mu_2 & -\lambda_1 \\ -\mu_1 & \lambda_1+\lambda_2+\mu_1+\mu_2-\alpha_0^{BP}\lambda_1 \end{bmatrix}$, and $G_{\alpha^{L_k^1,L_{k+1}^1}}(z)$ is the GF of $\alpha^{L_k^1,L_{k+1}^1}$. It can be calculated from (5) as:

$$G_{\alpha^{L_k^1,L_{k+1}^1}}(z)=LT^{L_k^1,L_{k+1}^1}(\lambda_2-\lambda_2 z). \tag{A18}$$

A3.3. Expressing $\vec{d_0}$ in Closed Form To obtain $\vec{d_0}$, we let $z\to 1$ in $(A10)$ and get

$$\left[G_{(0,L^2)}(1),G_{(1,L^2)}(1)\right]=\vec{d_0}\cdot\lim_{z\to 1}D(z). \tag{A19}$$

Notice that the denominator of $D(z)$ is zero when $z\to 1$, so we need to apply L'Hopital's rule to calculate $\lim_{z\to 1}D(z)$. The value of $\lim_{z\to 1}D(z)$ is determined by $G_{\alpha^{BP}}(z),G_{\alpha^{00}}(z),G_{\alpha^{11}}(z),G_{\alpha^{01}}(z),G_{\alpha^{10}}(z)$ and their first order derivatives, which can all be calculated from (9) and $(A18)$.

Note that $(A19)$ is composed of two equations with four unknowns: $G_{(0,L^2)}(1)$, $G_{(1,L^2)}(1)$, d_{00} and d_{10}. Another equation is the normalization requirement

$$G_{(0,L^2)}(1)+G_{(1,L^2)}(1)=1. \tag{A20}$$

Thus, to find a closed-form expression of $\left[G_{(0,L^2)}(z),G_{(1,L^2)}(z)\right]$, we need another linearly independent equation of these four variables. To find this equation, we focus on the value of $\varphi_1=\frac{d_{10}}{d_{10}+d_{00}}$.

Let a *Level-j Class-2 busy period* $(j=0,1,\ldots)$ start once a Class-2 job arrives at the system when j Class-2 jobs are present (but not necessarily in service), and terminate at the first time the number of Class-2 jobs in the system drops back to j. Let "a Level-j Class-2 busy period starts with i Class-1 jobs" denote that the first Class-2 arrival in this Level-j Class-2 busy period sees i Class-1 jobs, similarly "a Level-j Class-2 busy period ends with i Class-1 jobs" denote that the Class-2 departure that ends this Level-j Class-2 busy period sees i Class-1 jobs. Recall that, in our M/M/2 queue, a Class-2 departure sees either zero or one Class-1 job. With these definitions, φ_1 is the probability that a Level-0 Class-2 busy period ends with one Class-1 job.

Let Π_i be the probability that a Level-0 Class-2 busy period starts with $i \geq 0$ Class-1 jobs. Let F_i be the probability that a Level-0 Class-2 busy period that started with $i \geq 0$ Class-1 jobs ends with one Class-1 job. Note that, in the $c = 2$ case, the probability that a Level-j Class-2 busy period ($j = 1, 2, \ldots$) that started with a fixed $i \geq 0$ Class-1 jobs ends with one Class-1 jobs is the same for any Level-j Class-2 busy period for any $j = 1, 2, \ldots$. Let B_i be this probability.

Using the Total Probability Theorem, we have

$$\varphi_1 = \sum_{i=0}^{\infty} \Pi_i F_i \ . \tag{A21}$$

Thus, if we can find F_i and Π_i in closed form, we can also express φ_1 in closed form.

We now discuss the possible sequences of events in these busy periods, and use the memoryless property to write recursive expressions for F_i and B_i. For example, if a Level-0 Class-2 busy period starts with no Class-1 jobs (i.e., a Class-2 job arrives at an empty system), then three events may happen next in the system:

1. Class-1 arrival, w.p. $\frac{\lambda_1}{\lambda_1+\lambda_2+\mu_2}$. Thus, one Class-1 job is in the system. Then, due to the memoryless property, F_0 is identical to F_1, the probability that a Level-0 Class-2 busy period that started with one Class-1 job ends with one Class-1 job.

2. Class-2 arrival, w.p. $\frac{\lambda_2}{\lambda_1+\lambda_2+\mu_2}$. A Level-1 Class-2 busy period is started. It ends with one Class-2 job and either zero or one Class-1 job:

   (a) One Class-1 job, w.p. B_0. Then, due to the memoryless property, a Level-0 Class-2 busy period starts with one Class-1 job, and it will end with one Class-1 job w.p. F_1.

   (b) No Class-1 jobs, w.p. $1 - B_0$. Then, due to the memoryless property, a Level-0 Class-2 busy period starts with no Class-1 jobs, and it will end with one Class-1 job w.p. F_0.

3. Class-2 departure, w.p. $\frac{\mu_2}{\lambda_1+\lambda_2+\mu_2}$. A Level-0 Class-2 busy period ends with no Class-1 jobs. That is, it ends with one Class-1 job w.p. 0.

Using the Total Probability Theorem and multiplying by $\lambda_1 + \lambda_2 + \mu_2$, we get

$$(\lambda_1 + \lambda_2 + \mu_2)F_0 = \lambda_1 F_1 + \lambda_2(B_0 F_1 + (1 - B_0)F_0) + \mu_2 \cdot 0. \tag{A22}$$

Similar logic yields

$$(\lambda_1+\lambda_2+\mu_1+\mu_2)F_1 = \lambda_1 F_2+\lambda_2(B_1F_1+(1-B_1)F_0)+\mu_1F_0+\mu_2, \tag{A23}$$

$$(\lambda_1+\lambda_2+2\mu_1)F_i = \lambda_1 F_{i+1}+\lambda_2(B_iF_1+(1-B_i)F_0)+2\mu_1F_{i-1} \text{ for } i\geq 2, \tag{A24}$$

for a Level-0 Class-2 busy period; and

$$(\lambda_1+\lambda_2+2\mu_2)B_0 = \lambda_1 B_1+\lambda_2(B_0B_1+(1-B_0)B_0)+2\mu_2\cdot 0, \tag{A25}$$

$$(\lambda_1+\lambda_2+\mu_1+\mu_2)B_1 = \lambda_1 B_2+\lambda_2(B_1B_1+(1-B_1)B_0)+\mu_1B_0+\mu_2, \tag{A26}$$

$$(\lambda_1+\lambda_2+2\mu_1)B_i = \lambda_1 B_{i+1}+\lambda_2(B_iB_1+(1-B_i)B_0)+2\mu_1B_{i-1} \text{ for } i\geq 2, \tag{A27}$$

for Level-j Class-2 busy periods, $j=1,2,\ldots$.

Note that B_i is independent of F_i, but F_i depends on B_i. Therefore, we first express B_i.

LEMMA 3. *B_i is given by*

$$B_i = \begin{cases} \frac{\lambda_1\Delta_0^B}{2\mu_2-\lambda_2\Delta_0^B} & \text{if } i=0 \\ \frac{\lambda_1\Delta_0^B}{2\mu_2-\lambda_2\Delta_0^B}+\Delta_0^B+\kappa\frac{g-g^i}{1-g} & \text{if } i\geq 1 \end{cases},$$

where $\Delta_0^B = \frac{-2\mu_1+g\lambda_1+g\lambda_2+2g\mu_1-g^2\lambda_1}{g\lambda_2}$, $\kappa = \frac{1}{\lambda_1 g}((\lambda_1+\lambda_2+\mu_1+\mu_2-\lambda_2\Delta_0^B)\Delta_0^B-\frac{\lambda_1\mu_2\Delta_0^B}{2\mu_2-\lambda_2\Delta_0^B}-\mu_2)$, *and g is the only root in* $(0,1)$ *of the following quartic function:*

$$\lambda_1^2g^4+\lambda_1(2\mu_2-\lambda_1-\lambda_2-4\mu_1)g^3+2(\mu_1(2\mu_1+4\lambda_1+\lambda_2-2\mu_2)-\lambda_1\mu_2)g^2+4\mu_1(\mu_2-\lambda_1-\lambda_2-3\mu_1)g+8\mu_1^2.$$

Then, using the same technique, we can express F_i.

LEMMA 4. *F_i is given by*

$$F_i = \begin{cases} \frac{2\lambda_1\mu_2\Delta_0^F}{\mu_2(2\mu_2-\lambda_2\Delta_0^B)} & \text{if } i=0 \\ \frac{2\lambda_1+2\mu_2-\lambda_2\Delta_0^B}{2\mu_2-\lambda_2\Delta_0^B}\Delta_0^F+\xi_1\frac{h-h^i}{1-h}+\xi_2\frac{g-g^i}{1-g} & \text{if } i\geq 1 \end{cases}, \tag{A28}$$

where $h=\frac{1}{2\lambda_1}((\lambda_1+\lambda_2+2\mu_1)-\sqrt{(\lambda_1+\lambda_2+2\mu_1)^2-8\lambda_1\mu_1})$, *and*

$$\begin{bmatrix}\xi_1\\ \xi_2\\ \Delta_0^F\end{bmatrix} = H^{-1}\begin{bmatrix}-\frac{\mu_2}{\lambda_1}\\ \frac{1}{\lambda_1}(\mu_2-\frac{1}{\lambda_1}\mu_2(\lambda_1+\lambda_2+2\mu_1))\\ (\frac{2}{\lambda_1^2}\mu_1\mu_2+\frac{1}{\lambda_1^2}(\mu_2-\frac{1}{\lambda_1}\mu_2(\lambda_1+\lambda_2+2\mu_1))(\lambda_1+\lambda_2+2\mu_1))\end{bmatrix},$$

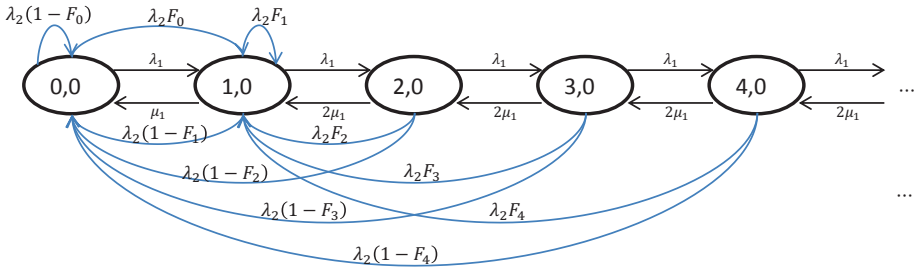

Figure A2 The MC when there are no Class-2 jobs.

in which

$$H=\begin{bmatrix} h & g & -\frac{\lambda_1+\lambda_2+\mu_1+\mu_2-\lambda_2\Delta_0^B}{\lambda_1} \\ h^2 & g^2 & \frac{\mu_1+\mu_2+g\kappa\lambda_2}{\lambda_1}-\frac{\lambda_1+\lambda_2+2\mu_1}{\lambda_1^2}(\lambda_1+\lambda_2+\mu_1+\mu_2-\lambda_2\Delta_0^B)+\frac{2\mu_2}{2\mu_2-\lambda_2\Delta_0^B} \\ h^3 & g^3 & \frac{2\mu_1}{\lambda_1^2}(\lambda_1+\lambda_2+\mu_1+\mu_2-\lambda_2\Delta_0^B)+\frac{\lambda_2\kappa g^2}{\lambda_1} \\ & & +\frac{\lambda_1+\lambda_2+2\mu_1}{\lambda_1^2}(\mu_1+\mu_2+g\kappa\lambda_2-\frac{\lambda_1+\lambda_2+2\mu_1}{\lambda_1}(\lambda_1+\lambda_2+\mu_1+\mu_2-\lambda_2\Delta_0^B)+\frac{2\lambda_1\mu_2}{2\mu_2-\lambda_2\Delta_0^B}) \end{bmatrix},$$

which is a nonsingular matrix according to the row reduction result.

Now we seek Π_i. The MC in Figure A2 tracks the number of Class-1 jobs present when a Level-0 Class-2 busy period starts; $\Pi_i, \forall i \geq 0$ is the solution to this MC. To find the Π_i, we write down the Balance Equations:

$$\lambda_2(1-\Pi_0) = \lambda_1\Pi_0 - \mu_1\Pi_1 + \lambda_2\varphi_1 \tag{A29}$$

$$\lambda_2(1-\Pi_0-\Pi_1) = \lambda_1\Pi_1 - 2\mu_1\Pi_2 \tag{A30}$$

$$\vdots$$

$$\lambda_2(1-\sum_{j=0}^{i}\Pi_j) = \lambda_1\Pi_i - 2\mu_1\Pi_{i+1} \tag{A31}$$

Again, using the same technique, we can express Π_i.

LEMMA 5. *Π_i can be expressed as a function of φ_1:*

$$\Pi_i = \begin{cases} \frac{\mu_1(1-f)+\lambda_2(1-\varphi_1)}{\lambda_1+\lambda_2+\mu_1-f\mu_1} & \text{if } i=0 \\ \frac{(1-f)(\lambda_1+\lambda_2\varphi_1)}{\lambda_1+\lambda_2+\mu_1-f\mu_1}f^{i-1} & \text{if } i\geq 1 \end{cases}, \tag{A32}$$

where $f=\frac{1}{4\mu_1}(\lambda_1+\lambda_2+2\mu_1-\sqrt{(\lambda_1+\lambda_2+2\mu_1)^2-8\lambda_1\mu_1})$.

Substituting $(A28)$ and $(A32)$ in $(A21)$ gives us an equation of φ_1, from which we can get φ_1:

$$\varphi_1 = \frac{\lambda_1(f-1)E+\Delta_0^F\frac{2\lambda_1}{2\mu_2-\lambda_2\Delta_0^B}(\lambda_2+\mu_1-f\mu_1)}{-\lambda_2(f-1)E+\Delta_0^F\frac{2\lambda_1\lambda_2}{2\mu_2-\lambda_2\Delta_0^B}+(\lambda_1+\lambda_2+\mu_1-f\mu_1)}, \tag{A33}$$

where $E = -\frac{1}{f-1}(\frac{g\xi_2}{g-1} + \frac{h\xi_1}{h-1} - \frac{\Delta_0^F(2\lambda_1+2\mu_2-\lambda_2\Delta_0^B)}{2\mu_2-\lambda_2\Delta_0^B}) + \frac{g\xi_2}{(fg-1)(g-1)} + \frac{h\xi_1}{(fh-1)(h-1)}$.

Hence, $(A19)$, $(A20)$ and $(A33)$ give four equations with four unknowns whose solution gives $\vec{d_0}$.

A4. Proofs

A4.1. Proof of Lemma 1 The one step transition probability of the MC can be written in matrix form as

$$P = \begin{matrix} \\ T \\ A \end{matrix} \begin{matrix} \begin{matrix} T & A \end{matrix} \\ \begin{bmatrix} \Gamma_{T\to T} & \Gamma_{T\to A} \\ 0 & I \end{bmatrix} \end{matrix},$$

where I is the identity matrix. Then, P^n represents the n step transition probabilities for the MC. Using induction, we obtain

$$P^n = \begin{bmatrix} \Gamma_{T\to T}^n & \sum_{i=0}^n \Gamma_{T\to T}^i \Gamma_{T\to A} \\ 0 & I \end{bmatrix}.$$

By letting n go to infinity and noting that $\sum_{i=0}^n \Gamma_{T\to T}^i = (I - \Gamma_{T\to T})^{-1}$, the probability that the system eventually reaches a state $A_i \in A$ is as given in the Lemma.

A4.2. Proof of Lemma 3 After some algebra, we can write $(A25 - A27)$ as:

$$B_1 = \frac{(\lambda_1 + 2\mu_2 + \lambda_2 B_0)B_0}{\lambda_1 + \lambda_2 B_0}, \tag{A34}$$

$$B_2 = \frac{1}{\lambda_1}((\lambda_1 + \lambda_2 + \mu_1 + \mu_2 - \lambda_2(B_1 - B_0))B_1 - (\lambda_2 + \mu_1)B_0 - \mu_2), \tag{A35}$$

$$B_{i+1} = \frac{1}{\lambda_1}((\lambda_1 + \lambda_2 + 2\mu_1 - \lambda_2(B_1 - B_0))B_i - 2\mu_1 B_{i-1} - \lambda_2 B_0) \text{ for } i \geq 2. \tag{A36}$$

Let $\Delta_i^B = B_{i+1} - B_i$ for $i \geq 0$, be the step difference of the sequence B_i. So, we have

$$B_i = B_1 + \sum_{j=1}^{i-1} \Delta_j^B \text{ for } i \geq 2. \tag{A37}$$

From the definition of Δ_i^B and $(A34)$, we get $\Delta_0^B = \frac{2\mu_2 B_0}{\lambda_1 + \lambda_2 B_0}$. Similarly, we get from $(A34 - A36)$

$$\Delta_1^B = \frac{1}{\lambda_1}((\lambda_1 + \lambda_2 + \mu_1 + \mu_2 - \lambda_2\Delta_0^B)\Delta_0^B - \frac{\lambda_1\mu_2\Delta_0^B}{2\mu_2 - \lambda_2\Delta_0^B} - \mu_2), \tag{A38}$$

$$\Delta_2^B = \frac{1}{\lambda_1}((\lambda_1 + \lambda_2 + 2\mu_1 - \lambda_2\Delta_0^B)\Delta_1^B - (\mu_1 + \mu_2)\Delta_0^B - \frac{\lambda_1\mu_2\Delta_0^B}{2\mu_2 - \lambda_2\Delta_0^B} + \mu_2), \tag{A39}$$

$$\Delta_i^B = \frac{(\lambda_1 + \lambda_2 + 2\mu_1 - \lambda_2\Delta_0^B)}{\lambda_1}\Delta_{i-1}^B - \frac{2\mu_1}{\lambda_1}\Delta_{i-2}^B \text{ for } i \geq 3. \tag{A40}$$

We notice that Δ_i^B is a linear homogeneous function of Δ_{i-1}^B and Δ_{i-2}^B , so Δ_i^B is a *linear homogeneous recurrence sequence* (see e.g., Green and Knuth (1990) Chapter 2). The solution to the recurrence sequence takes the form $\Delta_i^B = \kappa_1 g_1^i + \kappa_2 g_2^i,\ i \geq 1$, where g_1 and g_2 are roots of the *Characteristic Polynomial*: $CP(g) = \lambda_1 g^2 - (\lambda_1 + \lambda_2 + 2\mu_1 - \lambda_2 \Delta_0^B)g + 2\mu_1$. Note that because $B_i \in [0,1]$, we have

$$\lim_{i\to\infty} \Delta_i^B = 0. \tag{A41}$$

For Δ_i^B to satisfy $(A41)$, i.e., converge to zero, either $g_j < 1$ or $\kappa_j = 0$ for both $j = 1, 2$. Because B_0,$B_1 \in [0,1]$, we have $\Delta_0^B < 1$, so we have $CP(1) = \lambda_2(\Delta_0^B - 1) < 0$. Thus, $CP(g)$ has only one root that is smaller than one:

$$g = \frac{1}{2\lambda_1}(\lambda_1 + \lambda_2 + 2\mu_1 - \lambda_2\Delta_0^B - \sqrt{(\lambda_1 + \lambda_2 + 2\mu_1 - \lambda_2\Delta_0^B)^2 - 8\lambda_1\mu_1}). \tag{A42}$$

(It is also easy to verify that g is greater than zero.) For the other root that is greater than one, the corresponding κ_j must be zero. Thus, Δ_i^B takes the form

$$\Delta_i^B = \kappa g^i,\ i \geq 1. \tag{A43}$$

Notice that g is a function of Δ_0^B, so in the expression of Δ_i^B we have two unknowns: κ and Δ_0^B. Substituting $(A43)$ into $(A38)$ and $(A39)$ gives

$$\kappa g = \frac{1}{\lambda_1}((\lambda_1 + \lambda_2 + \mu_1 + \mu_2 - \lambda_2\Delta_0^B)\Delta_0^B - \frac{\lambda_1\mu_2\Delta_0^B}{2\mu_2 - \lambda_2\Delta_0^B} - \mu_2), \tag{A44}$$

$$\kappa g^2 = \frac{1}{\lambda_1}((\lambda_1 + \lambda_2 + 2\mu_1 - \lambda_2\Delta_0^B)\Delta_1^B - (\mu_1 + \mu_2)\Delta_0^B - \frac{\lambda_1\mu_2\Delta_0^B}{2\mu_2 - \lambda_2\Delta_0^B} + \mu_2). \tag{A45}$$

Dividing $(A45)$ with $(A44)$ gives:

$$g = -\frac{1}{\lambda_1}\frac{\begin{array}{c}\lambda_2^3{\Delta_0^B}^4 - \lambda_2^2(2\lambda_1 + 2\lambda_2 + 3\mu_1 + 3\mu_2){\Delta_0^B}^3 \\ +\lambda_2(\lambda_1^2 + 2\lambda_1\lambda_2 + \lambda_2^2 + 2\lambda_1\mu_1 + 3\lambda_1\mu_2 + 3\lambda_2\mu_1 + 6\lambda_2\mu_2 + 2\mu_1^2 + 8\mu_1\mu_2 + 2\mu_2^2){\Delta_0^B}^2 \\ -\mu_2(3\lambda_2^2 + 8\lambda_2\mu_1 + 4\lambda_2\mu_2 + 3\lambda_1\lambda_2 + 4\mu_1^2 + 4\mu_1\mu_2 + 2\lambda_1\mu_1)\Delta_0^B + 2\mu_2^2(\lambda_2 + 2\mu_1)\end{array}}{\lambda_2^2{\Delta_0^B}^3 - \lambda_2(\lambda_2 + \lambda_1 + \mu_1 + 3\mu_2){\Delta_0^B}^2 + \mu_2(2\mu_2 + \lambda_1 + 3\lambda_2 + 2\mu_1)\Delta_0^B - 2\mu_2^2}. \tag{A46}$$

Substituting $\Delta_0^B = -\frac{\lambda_1 g^2 - (\lambda_1 + \lambda_2 + 2\mu_1)g + 2\mu_1}{\lambda_2 g}$ into $(A46)$ gives a polynomial equation of degree six:

$0 = \lambda_1^3(\mu_1 - \mu_2)g^6 - \lambda_1^2(6\mu_1^2 + 2\mu_2^2 + 2\lambda_1\mu_1 - \lambda_1\mu_2 + 2\lambda_2\mu_1 - \lambda_2\mu_2 - 8\mu_1\mu_2)g^5 + (\lambda_1^3\mu_1 + 2\lambda_1^2\lambda_2\mu_1 +$
$16\lambda_1^2\mu_1^2 - 14\lambda_1^2\mu_1\mu_2 + 2\lambda_1^2\mu_2^2 + \lambda_1\lambda_2^2\mu_1 + 8\lambda_1\lambda_2\mu_1^2 - 6\lambda_1\lambda_2\mu_1\mu_2 + 12\lambda_1\mu_1^3 - 20\lambda_1\mu_1^2\mu_2 + 8\lambda_1\mu_1\mu_2^2)g^4 -$

$(14\lambda_1^2\mu_1^2 - 6\lambda_1^2\mu_1\mu_2 + 16\lambda_1\lambda_2\mu_1^2 - 6\lambda_1\lambda_2\mu_1\mu_2 + 40\lambda_1\mu_1^3 - 44\lambda_1\mu_1^2\mu_2 - 8\lambda_1\mu_1\mu_2^2 + 2\lambda_2^2\mu_1^2 + 8\lambda_2\mu_1^3 - 8\lambda_2\mu_1^2\mu_2 + 8\mu_1^4 - 16\mu_1^3\mu_2 + 8\mu_1^2\mu_2^2)g^3 + 4\mu_1^2(\lambda_1^2 + 2\lambda_1\lambda_2 + 11\lambda_1\mu_1 - 6\lambda_1\mu_2 + \lambda_2^2 + 6\lambda_2\mu_1 - 3\lambda_2\mu_2 + 8\mu_1^2 - 10\mu_1\mu_2 + 2\mu_2^2)g^2 - (40\mu_1^4 + 16\lambda_1\mu_1^3 + 16\lambda_2\mu_1^3 - 24\mu_1^3\mu_2)g + 16\mu_1^4$

If $\mu_1 \neq \mu_2$, then we have several possible solutions:

$$g_a = \frac{\mu_1}{2\lambda_1(\mu_1-\mu_2)}(\lambda_1+\lambda_2+2\mu_1-2\mu_2-\sqrt{(\lambda_1+\lambda_2+2\mu_1-2\mu_2)^2-8\lambda_1(\mu_1-\mu_2)}),$$
$$g_b = \frac{\mu_1}{2\lambda_1(\mu_1-\mu_2)}(\lambda_1+\lambda_2+2\mu_1-2\mu_2+\sqrt{(\lambda_1+\lambda_2+2\mu_1-2\mu_2)^2-8\lambda_1(\mu_1-\mu_2)}),$$

and roots of $\varpi(g) = a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$, where $a_4 = \lambda_1^2,\ a_3 = \lambda_1(2\mu_2 - \lambda_2 - 4\mu_1 - \lambda_1),\ a_2 = 2(2\mu_1^2 + 4\lambda_1\mu_1 - \lambda_1\mu_2 + \lambda_2\mu_1 - 2\mu_1\mu_2), a_1 = 4\mu_1(\mu_2 - \lambda_1 - \lambda_2 - 3\mu_1),\ a_0 = 8\mu_1^2$. It is easy to check that $g_a > 1$; if $\mu_1 > \mu_2$, then $g_b > g_a$ so $g_b > 1$; if $\mu_1 < \mu_2$, then $g_b < 0$. So, g cannot be g_a or g_b, and g must be one of the four roots of $\varpi(g)$.

The four roots of a quartic function (polynomial of degree four) are well known. Let $\Delta_1 = a_2^2 - 3a_3a_1 + 12a_4a_0,\ \Delta_2 = 2a_2^3 - 9a_3a_2a_1 + 27a_4a_1^2 + 27a_3^2a_0 - 72a_4a_2a_0$, and $\Delta = \frac{\sqrt[3]{2}\Delta_1}{3a_4\sqrt[3]{\Delta_2+\sqrt{-4\Delta_1^3+\Delta_2^2}}} + \frac{\sqrt[3]{\Delta_2+\sqrt{-4\Delta_1^3+\Delta_2^2}}}{3\sqrt[3]{2}a_4}$, then the four roots of $\varpi(g)$ are

$$x_1 = -\frac{a_3}{4a_4} - \frac{1}{2}\sqrt{\frac{a_3^2}{4a_4^2} - \frac{2a_2}{3a_4} + \Delta} - \frac{1}{2}\sqrt{\frac{a_3^2}{2a_4^2} - \frac{4a_2}{3a_4} - \Delta - \frac{-\frac{a_3^3}{a_4^3} + \frac{4a_3a_2}{a_4^2} - \frac{8a_1}{a_4}}{4\sqrt{\frac{a_3^2}{4a_4^2} - \frac{2a_2}{3a_4} + \Delta}}}, \tag{A47}$$

$$x_2 = -\frac{a_3}{4a_4} - \frac{1}{2}\sqrt{\frac{a_3^2}{4a_4^2} - \frac{2a_2}{3a_4} + \Delta} + \frac{1}{2}\sqrt{\frac{a_3^2}{2a_4^2} - \frac{4a_2}{3a_4} - \Delta - \frac{-\frac{a_3^3}{a_4^3} + \frac{4a_3a_2}{a_4^2} - \frac{8a_1}{a_4}}{4\sqrt{\frac{a_3^2}{4a_4^2} - \frac{2a_2}{3a_4} + \Delta}}}, \tag{A48}$$

$$x_3 = -\frac{a_3}{4a_4} + \frac{1}{2}\sqrt{\frac{a_3^2}{4a_4^2} - \frac{2a_2}{3a_4} + \Delta} - \frac{1}{2}\sqrt{\frac{a_3^2}{2a_4^2} - \frac{4a_2}{3a_4} - \Delta + \frac{-\frac{a_3^3}{a_4^3} + \frac{4a_3a_2}{a_4^2} - \frac{8a_1}{a_4}}{4\sqrt{\frac{a_3^2}{4a_4^2} - \frac{2a_2}{3a_4} + \Delta}}}, \tag{A49}$$

$$x_4 = -\frac{a_3}{4a_4} + \frac{1}{2}\sqrt{\frac{a_3^2}{4a_4^2} - \frac{2a_2}{3a_4} + \Delta} + \frac{1}{2}\sqrt{\frac{a_3^2}{2a_4^2} - \frac{4a_2}{3a_4} - \Delta + \frac{-\frac{a_3^3}{a_4^3} + \frac{4a_3a_2}{a_4^2} - \frac{8a_1}{a_4}}{4\sqrt{\frac{a_3^2}{4a_4^2} - \frac{2a_2}{3a_4} + \Delta}}}. \tag{A50}$$

Because $\varpi(1) < 0$ and $\lim_{g\to\infty}\varpi(g) = \infty$, $\varpi(g)$ has at least one root in $(1,\infty)$. Because $\varpi(0) = 8\mu_1^2 > 0$ and $\varpi(1) = -\lambda_2(\lambda_1 + 2\mu_1) < 0$, $\varpi(g)$ has at least one root in $(0,1)$. Because $\varpi(0) = 8\mu_1^2 > 0$ and $\lim_{g\to-\infty}\varpi(g) = \infty$, $\varpi(g)$ has either two or no roots in $(-\infty, 0)$. Next, we prove $\varpi(g)$ has only

one root in $(0,1)$.

From $\sum_{i=1}^{2}\frac{\lambda_i}{\mu_i}<2$, we get that $\mu_2>\frac{\lambda_2\mu_1}{2\mu_1-\lambda_1}$. Then we discuss the following three cases:

1. If $\frac{\lambda_2\mu_1}{2\mu_1-\lambda_1}\geq\frac{\lambda_2+\lambda_1+4\mu_1}{2}$, then $\mu_2>\frac{\lambda_2+\lambda_1+4\mu_1}{2}$, i.e., $a_3=\lambda_1(2\mu_2-\lambda_2-4\mu_1-\lambda_1)>0$. Note from $(A47)$ that, in this case, x_1 is either a complex root or a negative real root:

(a) If x_1 is a complex root, because of the *Complex Conjugate Root Theorem* (i.e., Jeffrey 2005), x_2 must be the other complex root. Obviously $x_4\geq x_3$, so we know $x_4\in(1,\infty)$ and $x_3\in(0,1)$.

(b) If x_1 is a negative real root, because $\varpi(g)$ has either two or no roots in $(-\infty,0)$, $\varpi(g)$ must have two negative real roots. Therefore, $\varpi(g)$ has only one root in $(0,1)$.

2. If $\frac{\lambda_2\mu_1}{2\mu_1-\lambda_1}<\frac{\lambda_2+\lambda_1+4\mu_1}{2}$ and $\mu_2>\frac{\lambda_2+\lambda_1+4\mu_1}{2}$, then as in the first case, we know $x_4\in(1,\infty)$ and $x_3\in(0,1)$.

3. If $\frac{\lambda_2\mu_1}{2\mu_1-\lambda_1}<\frac{\lambda_2+\lambda_1+4\mu_1}{2}$ and $\frac{\lambda_2\mu_1}{2\mu_1-\lambda_1}<\mu_2\leq\frac{\lambda_2+\lambda_1+4\mu_1}{2}$, we let $\epsilon=\frac{\lambda_2+\lambda_1+4\mu_1}{2}-\mu_2$ (i.e., $0\leq\epsilon<\frac{\lambda_1^2+2\lambda_1\mu_1+\lambda_2\lambda_1-8\mu_1^2}{2(\lambda_1-2\mu_1)}$), $\varpi_1(g)=-2\lambda_1g^3+2(\lambda_1+2\mu_1)g^2-4\mu_1g$ and $\varpi_2(g)=\lambda_1^2g^4+(-\lambda_1^2+2\lambda_1\mu_1-\lambda_2\lambda_1-4\mu_1^2)g^2-2\mu_1(2\mu_1+\lambda_1+\lambda_2)g+8\mu_1^2$, so that $\varpi(g)=\epsilon\varpi_1(g)+\varpi_2(g)$.

$\varpi_1(g)$ and $\varpi_2(g)$ have some properties that are easy to derive that can be used to identify the root we want.

- $\varpi_1(g)$ is a convex function on $[0,1]$ and $\varpi_1(0)=\varpi_1(1)=0$.
- $\varpi_2(g)$ is a decreasing function on $[0,1]$, $\varpi_2(0)=8\mu_1^2>0$ and $\varpi_2(1)=-\lambda_2(\lambda_1+2\mu_1)<0$.

To prove $\varpi_2(g)$ is a decreasing function on $[0,1]$, we just need to prove the first derivative of $\varpi_2(g)$ is negative, i.e., $\varpi_2'(g)=4\lambda_1^2g^3+(4\lambda_1\mu_1-2\lambda_1^2-2\lambda_2\lambda_1-8\mu_1^2)g-(4\mu_1^2+2\lambda_1\mu_1+2\lambda_2\mu_1)<0$, for $\forall g\in[0,1]$.

Obviously, $\varpi_2'(0)=-(4\mu_1^2+2\lambda_1\mu_1+2\lambda_2\mu_1)<0$ and $\varpi_2'(1)=-2(4\mu_1^2-\lambda_1^2)-2\mu_1(2\mu_1-\lambda_1)-2\lambda_2\lambda_1-2\lambda_2\mu_1<0$. We know the second derivative of $\varpi_2(g)$ is $\varpi_2''(g)=12\lambda_1^2g^2+(4\lambda_1\mu_1-2\lambda_1^2-2\lambda_2\lambda_1-8\mu_1^2)$ and $\varpi_2''(0)=-4\mu_1(2\mu_1-\lambda_1)-2\lambda_1^2-2\lambda_2\lambda_1<0$. If there exists a point $\bar{g}$ in $[0,1]$ such that $\varpi_2'(\bar{g})>0$, then $\varpi_2'(g)$ must have two critical points in $[0,1]$, i.e., $\varpi_2''(g)$ must have two roots in $[0,1]$. However, we know $\varpi_2''(g)$ has one negative root and one positive root. Therefore, $\varpi_2'(g)<0$ for $\forall g\in[0,1]$. Thus, $\varpi_2(g)$ is a decreasing function for $\forall g\in[0,1]$.

Therefore, for $\forall\epsilon\geq0$, $\varpi(g)=\epsilon\varpi_1(g)+\varpi_2(g)$ has only one root in $(0,1)$.

Hence, we proved $\varpi(g)$ has only one root in $(0,1)$. Then, we just need to pick up the root in $(0,1)$ from the four roots of $\varpi(g)$, which is not difficult. Once we get g, solving $(A42)$ and $(A44)$ gives the corresponding Δ_0^B and κ as given in Lemma 3.

A4.3. Proof of Lemma 4 As in the Proof of Lemma 3, we write $(A22-A24)$ in another form:

$$F_1 = \frac{(\lambda_1+\mu_2+\lambda_2 B_0)F_0}{\lambda_1+\lambda_2 B_0} = \frac{2\lambda_1+2\mu_2-\lambda_2\Delta_0^B}{2\lambda_1}F_0, \tag{A51}$$

$$F_2 = \frac{1}{\lambda_1}((\lambda_1+\lambda_2+\mu_1+\mu_2)F_1-(\lambda_2+\mu_1)F_0-\lambda_2 B_1(F_1-F_0)-\mu_2), \tag{A52}$$

$$F_{i+1} = \frac{1}{\lambda_1}((\lambda_1+\lambda_2+2\mu_1)F_i-2\mu_1 F_{i-1}-\lambda_2 B_i(F_1-F_0)-\lambda_2 F_0) \text{ for } i\geq 2. \tag{A53}$$

Let $\Delta_i^F = F_{i+1}-F_i$ be the step difference of the sequence F_i. So, we have

$$F_i = F_1+\sum_{j=1}^{i-1}\Delta_j^F \text{ for } i\geq 2.$$

Because $F_i\in[0,1]$, we have $\lim_{i\to\infty}\Delta_i^F=0$. Using $(A51)$, we get $\Delta_0^F=\frac{\mu_2 F_0}{\lambda_1+\lambda_2 B_0}$. Similarly, from $(A51-A53)$, we get

$$\begin{aligned}
\Delta_1^F &= \frac{1}{\lambda_1}((\lambda_1+\lambda_2+\mu_1+\mu_2-\lambda_2\Delta_0^B)\Delta_0^F-\mu_2),\\
\Delta_2^F &= \frac{1}{\lambda_1}((\lambda_1+\lambda_2+2\mu_1)\Delta_1^F-(\lambda_2\kappa g+\mu_1+\mu_2)\Delta_0^F-(\lambda_1+\lambda_2 B_0)\Delta_0^F+\mu_2),\\
\Delta_i^F &= \frac{(\lambda_1+\lambda_2+2\mu_1)}{\lambda_1}\Delta_{i-1}^F-\frac{2\mu_1}{\lambda_1}\Delta_{i-2}^F-\frac{\lambda_2\kappa\Delta_0^F}{\lambda_1 g}g^i \text{ for } i\geq 3.
\end{aligned}$$

Note that Δ_i^F is a linear non-homogeneous function of Δ_{i-1}^F and Δ_{i-2}^F, so Δ_i^F is a *non-homogeneous recurrence sequence* (see e.g., Green and Knuth (1990) Chapter 2), with solution of the form

$$\Delta_i^F = \xi_1 h_1^i+\xi_2 h_2^i+\xi_3 g^i,$$

where g is given in Lemma 3; h_1 and h_2 are roots of $\lambda_1 h^2-(\lambda_1+\lambda_2+2\mu_1)h+2\mu_1=0$. We know one of the two roots is greater than one. Because Δ_i^F converges to zero, with the same discussion in the proof of Lemma 3, we get that Δ_i^F has the form:

$$\Delta_i^F = \xi_1 h^i+\xi_2 g^i,\ i\geq 1$$

where $h = \frac{1}{2\lambda_1}((\lambda_1+\lambda_2+2\mu_1) - \sqrt{(\lambda_1+\lambda_2+2\mu_1)^2 - 8\lambda_1\mu_1})$. To find ξ_1, ξ_2 and Δ_0^F, we solve three equations

$$\Delta_1^F = \xi_1 h + \xi_2 g,\ \Delta_2^F = \xi_1 h^2 + \xi_2 g^2,\ \Delta_3^F = \xi_1 h^3 + \xi_2 g^3.$$

Notice that $\Delta_i^F, i = 1,2,3$ are all linear functions of Δ_0^F, so it is not hard to get the expression for ξ_1, ξ_2 and Δ_0^F in Lemma 4.

A4.4. Proof of Lemma 5 Subtracting the $(i-1)^{st}$ equation from the i^{th} equation given in (A30-A31) yields

$$2\mu_1\Pi_i = (\lambda_1+\lambda_2+2\mu_1)\Pi_{i-1} - \lambda_1\Pi_{i-2} \text{ for } i \geq 3.$$

This means that Π_i is a linear homogeneous recurrence sequence. The solution to the recurrence sequence takes the form

$$\Pi_i = \omega_1 f_1^i + \omega_2 f_2^i,\ i \geq 1,$$

where ω_1 and ω_2 are roots of

$$2\mu_1 f^2 - (\lambda_1+\lambda_2+2\mu_1)f + \lambda_1 = 0. \tag{A54}$$

Because $\Pi_i \in [0,1]$, we know either $f_j < 1$ or $\omega_j = 0$ for both $j = 1,2$. Equation $(A54)$ has one root greater than one and the other root smaller than one. For the root greater than one, the corresponding ω_j must be zero. Thus, Π_i takes the form

$$\Pi_i = \omega f^i \text{ for } i \geq 1, \tag{A55}$$

where $f = \frac{1}{4\mu_1}(\lambda_1+\lambda_2+2\mu_1 - \sqrt{(\lambda_1+\lambda_2+2\mu_1)^2 - 8\lambda_1\mu_1})$, which is the root smaller than one. Substituting Π_1 in $(A55)$ gives $\omega = \frac{\Pi_1}{f}$. From $(A29)$, we get

$$\Pi_1 = \frac{1}{\mu_1}((\lambda_1+\lambda_2)\Pi_0 - \lambda_2(1-\varphi_1)).$$

Therefore, from

$$1 = \sum_{i=0}^{\infty}\Pi_i = \frac{\Pi_1}{1-f} + \Pi_0 = \frac{(\lambda_1+\lambda_2)\Pi_0 - \lambda_2(1-\varphi_1)}{\mu_1(1-f)} + \Pi_0$$

we get $\Pi_0 = \frac{\mu_1(1-f)+\lambda_2(1-\varphi_1)}{\mu_1(1-f)+(\lambda_1+\lambda_2)}$. Therefore, Π_i can be expressed as a function of φ_1 as in $(A32)$.

A5. Algorithms

Algorithm 1 *Calculate the transition matrix of the EMC for* $\forall c \geq 2$.

Step 1*: Let* $\Gamma_{c\to c}$, $\Gamma_{c\to(c-1)}$ *and* $\Gamma_{c\to(c+1)^+}$ *be the one-step transition matrices from* $Q_c \cup BP_c$ *to* $Q_c \cup BP_c$, Q_{c-1} *and* $\cup_{j=c+1}^{\infty} Q_j$. *Set* $\Psi_{c1} = \left[I_{c\times c}\ \mathbf{0}_{c\times 1} \right] \cdot (I - \Gamma_{c\to c})^{-1}\Gamma_{c\to(c-1)}$ *and* $\Psi_{c2} = \left[I_{c\times c}\ \mathbf{0}_{c\times 1} \right] \cdot (I - \Gamma_{c\to c})^{-1}\Gamma_{c\to(c+1)^+}$. *Set* $A_0 = \Psi_{c1}$ *and let* $i = 1$.

Step 2*: Set* $A_i = \Psi_{c2} \left[A_{i-1}^T \cdots A_1^T\ A_0^T\ \mathbf{0}_{c\times\infty} \right]^T$.

Step 3*: Let* $i = i+1$. *If* $\max(A_i) > Tolerance$, *then go to* ***Step 2****; else set* $Limit = i$ *and* $i = c-1$, *and go to* ***Step 4***

Step 4*: Let* $\Gamma_{i\to i}$, $\Gamma_{i\to(i-1)}$ *and* $\Gamma_{i\to(i+1)^+}$ *be the one-step transition matrices from* $Q_i \cup BP_i$ *to* $Q_i \cup BP_i$, Q_{i-1} *and* $\cup_{j=i+1}^{\infty} Q_j$. *Set* $\Psi_{i1} = \left[I_{c\times c}\ \mathbf{0}_{c\times 1} \right] \cdot (I - \Gamma_{i\to i})^{-1}\Gamma_{i\to(i-1)}$, *and* $\Psi_{i2} = \left[I_{c\times c}\ \mathbf{0}_{c\times 1} \right] \cdot (I - \Gamma_{i\to i})^{-1}\Gamma_{i\to(i+1)^+}$. *Let* $j = 0$.

Step 5*: If* $j < i-1$, *then set* $M_{i\to j} = \mathbf{0}_{c\times c}$; *else if* $j = i-1$, *then set* $M_{i\to j} = \Psi_{i1}$; *else if* $i \leq j < c-1$, *then set* $M_{i\to j} = \Psi_{i2} \left[M_{i+1\to j}^T \cdots M_{c-2\to j}^T\ M_{c-1\to j}^T\ \mathbf{0}_{c\times\infty} \right]^T$; *else set* $M_{i\to j} = \Psi_{i2} \left[M_{i+1\to j}^T \cdots M_{c-1\to j}^T\ A_{j-c+1}^T \cdots A_0^T\ \mathbf{0}_{c\times\infty} \right]^T$.

Step 6*: Let* $j = j+1$. *If* $j < Limit$, *then go to* ***Step 5****; else let* $i = i-1$. *If* $i \geq 1$, *then let* $j = 0$ *and go to* ***Step 5****; else let* $i = 0$ *and go to* ***Step 7****.*

Step 7*: Let* $\Gamma_{0\to 0}$ *and* $\Gamma_{0\to 1^+}$ *be the one-step transition matrices from* $Q_0 \cup BP_0$ *to* $Q_0 \cup BP_0$, $\cup_{j=1}^{\infty} Q_j$. *Set* $\Psi_0 = \left[I_{c\times c}\ \mathbf{0}_{c\times 1} \right] \cdot (I - \Gamma_{0\to 0})^{-1}\Gamma_{0\to 1^+}$. *Let* $j = 0$.

Step 8*: If* $0 \leq j < c-1$, *then set* $M_{0\to j} = \Psi_0 \left[M_{1\to j}^T \cdots M_{c-2\to j}^T\ M_{c-1\to j}^T\ \mathbf{0}_{c\times\infty} \right]^T$; *else if* $M_{0\to j} = \Psi_0 \left[M_{i+1\to j}^T \cdots M_{c-1\to j}^T\ A_{j-c+1}^T \cdots A_0^T\ \mathbf{0}_{c\times\infty} \right]^T$.

Step 9*: Let* $j = j+1$. *If* $j < Limit$, *go to* ***Step 8****; else set* $G = A_0$ *and go to* ***Step 10****.*

Step 10*: Set* $G = \sum_{i=0}^{Limit} A_i G^i$.

Step 11*: If* $\max(G - \sum_{i=0}^{Limit} A_i G^i) > Tolerance$, *then go to* ***Step 10****; else set* $\hat{\mathbf{L}} = \begin{bmatrix} M_{0\to 0} & \cdots & M_{0\to c-1} \\ \vdots & \ddots & \vdots \\ M_{c-1\to 0} & \cdots & M_{c-1\to c-1} \end{bmatrix}$, $\hat{\mathbf{B}} = \left[\mathbf{0}_{c\times c(c-1)}\ A_0 \right]$, $\hat{\mathbf{F}}^{(i)} = \begin{bmatrix} M_{0\to i} \\ \vdots \\ M_{c-1\to i} \end{bmatrix}$ *for* $i = c, \ldots, Limit$, $\mathbf{B} = A_0$, $\mathbf{F}^{(0)} = \mathbf{L} = A_1$, $\mathbf{F}^{(i)} = A_{i+1}$ *for* $i \geq 1$, $\hat{\mathbf{S}}^{(i)} = \sum_{j=i}^{Limit} \hat{\mathbf{F}}^{(j)} G^{j-i}$ *for* $i \geq 1$ *and* $\mathbf{S}^{(i)} = \sum_{j=i}^{Limit} \mathbf{F}^{(j)} G^{j-i}$ *for* $i \geq 0$, *and go to* ***Step 12****.*

***Step 12**: Solve* $\left[\pi^{(0)}_{1\times c^2}\ \pi^{(1)}_{1\times c}\ \pi^{(*)}_{1\times c}\right]\cdot\begin{bmatrix}\mathbf{1}_{c^2\times 1} & \hat{\mathbf{L}} & \hat{\mathbf{F}}^{(1)}-\sum_{j=3}^{Limit}\hat{\mathbf{S}}^{(j)}G & \sum_{j=2}^{Limit}\hat{\mathbf{F}}^{(j)}+\sum_{j=3}^{Limit}\hat{\mathbf{S}}^{(j)}G\\ \mathbf{1}_{c\times 1} & \hat{\mathbf{B}} & \mathbf{L}-\sum_{j=2}^{Limit}\mathbf{S}^{(j)}G & \sum_{j=1}^{Limit}\mathbf{F}^{(j)}+\sum_{j=2}^{Limit}\mathbf{S}^{(j)}G\\ \mathbf{1}_{c\times 1} & \mathbf{0}_{c\times c} & \mathbf{B}-\sum_{j=1}^{Limit}\mathbf{S}^{(j)}G & \sum_{j=1}^{Limit}\mathbf{F}^{(j)}+\mathbf{L}+\sum_{j=1}^{Limit}\mathbf{S}^{(j)}G\end{bmatrix}=$ $\left[1,\mathbf{0}_{1\times(c^2+2c)}\right]$.

***Step 13**: Set* $\hat{\mathbf{F}}_{[k,i]}=\sum_{j=i}^{Limit}j^k\hat{\mathbf{F}}^{(j)}$ *for* $i\geq 1$ *and* $k=0$ *or 1,* $\mathbf{F}_{[k,i]}=\sum_{j=i}^{Limit}j^k\mathbf{F}^{(j)}$ *for* $i\geq 1$, $b^{[1]}=-\pi^{(0)}\sum_{j=1}^{Limit}(j+1)\hat{\mathbf{F}}^{(j)}-\pi^{(1)}(2\mathbf{L}+\sum_{j=1}^{Limit}(j+2)\mathbf{F}^{(j)})-\pi^{(*)}(\mathbf{L}+\sum_{j=1}^{Limit}(j+1)\mathbf{F}^{(j)})$, *and* $c^{[1]}=-\pi^{(0)}\sum_{j=2}^{Limit}j\hat{\mathbf{F}}_{[0,j]}\mathbf{1}^T-\pi^{(1)}\sum_{j=1}^{Limit}(j+1)\mathbf{F}_{[0,j]}\mathbf{1}^T-\pi^{(*)}\sum_{j=1}^{Limit}j\mathbf{F}_{[0,j]}\mathbf{1}^T$.

***Step 14**: Solve* $r^{[1]}\cdot\left[\mathbf{B}+\mathbf{L}+\sum_{j=1}^{Limit}\mathbf{F}^{(j)},(\mathbf{F}_{[1,1]}-\mathbf{B})\mathbf{1}^T\right]=\left[b^{[1]},c^{[1]}\right]$.

***Step 15**: Let* $E[L^2]=\pi^{(0)}_{1\times c^2}\cdot\left[\mathbf{0}_{1\times c}\ \mathbf{1}_{1\times c}\ \cdots\ (\mathbf{c}-\mathbf{1})_{1\times c}\right]^T+\pi^{(1)}_{1\times c}\cdot\left[\mathbf{c}_{1\times c}\right]^T+(r^{[1]}+(c-1)\pi^{(*)}_{1\times c})\cdot\mathbf{1}^T$.

Stop.