State-Dependent $M/G/1$ Queueing Systems

Hossein Abouee-Mehrizi
Department of Management Sciences, University of Waterloo, Waterloo, ON, N2L 3G1, CANADA,
HAboueeMehrizi@UWaterloo.ca

Opher Baron
Joseph L. Rotman School of Management, University of Toronto, Toronto, M5S 3E6, CANADA,
Opher.Baron@Rotman.Utoronto.Ca

We consider a state-dependent $M_n/G_n/1$ queueing system with both finite and infinite buffer sizes. We allow the arrival rate of customers to depend on the number of people in the system. Service times are also state-dependent and service rates can be modified at both arrivals and departures of customers. We show that the steady-state solution of this system at arbitrary times can be derived using the supplementary variable method, and that the system's state at arrival epochs can be analyzed using an embedded Markov chain. For the system with infinite buffer size, we first obtain an expression for the steady-state distribution of the number of customers in the system at both arbitrary and arrival times. Then, we derive the average service time of a customer observed at both arbitrary times and arrival epochs. We show that our state-dependent queueing system is equivalent to a Markovian birth-and-death process. This equivalency demonstrates our main insight that the $M_n/G_n/1$ system can be decomposed at any given state as a Markovian queue. Thus, many of the existing results for systems modeled as $M/M/1$ queue can be carried through to the much more practical $M/G/1$ model with state-dependent arrival and service rates. Then, we extend the results to the $M_n/G_n/1$ queueing systems with finite buffer size.

Key words: $M_n/G_n/1$ queue, birth-and-death process, state-dependent service times, state-dependent arrivals

1. Introduction

Markovian queues have been used to model and analyze congested queueing systems in a vast body of literature. The beauty of the single stage Markovian queues is that the system can be modeled as a Birth-and-Death (B&D) process and decomposed to new Markovian queues at any given state. These properties make the problem tractable even when the arrivals and service rates are state-dependent or when the control of the queue dynamically changes. This tractability made the $M/M/1$ queue a preferable model for many theoretical studies in management science.

However, in many applications assuming Markovian service times is not realistic. Thus, much attention has been given to analysis of queues with generally distributed service times. An important limitation of such queueing systems however is that their analysis is not straightforward. This limitation is especially true when the arrival rates and the service times depend on the state of the system. Moreover, since $M/G/1$ systems are not regenerative at arrival epochs, the solution of the system is not simple when the control of the queue is adjusted at arrival epochs. This apparent difficulty in the analysis of such queueing systems limited their study, even when they are more appropriate for the application considered.

In this paper we consider a queueing system with state-dependent arrival and service rates, denoted as $M_n/G_n/1$ with infinite and finite buffer sizes. Since the arrival process depends on the number of customers in the system, the Poisson Arrival Sees Time Average (PASTA) property does not hold. Therefore, we consider the system in both continuous time and at arrival epochs. To analyze the system in continuous time, we use the supplementary variable method to model the system as a continuous time Markov Chain (MC). Using this method, we derive a closed form expression for the steady-state distribution of the number of customers in the system (assuming, of course, that the system is stable). Then, we obtain the steady-state rate at which the $M_n/G_n/1$ system moves from state n to state $n-1$. These transition rates can be used to decompose the state-dependent queueing systems as a B&D process, i.e., show that the $M_n/G_n/1$ system can be decomposed into several $M_n/G_n/1$ queues.

We also analyze the system at arrival epochs. We define an embedded MC and obtain the transition probabilities in these MCs. Then, we derive the steady-state distribution of the number of customers in the systems observed by an arrival. We show that the probability of having n customers in the system at an arbitrary time and the probability of observing n people in the system by an arrival are closely related. Specifically, the ratio between these two distributions is identical to the ratio of the arrival rate when there are no customers in the system to the arrival rate when there are n customers in the system. We also derive the steady-state service rate of a customer in the $M_n/G_n/1$ system observed at arrival epochs.

Using these results, we finally analyze the $M_n/G_n/1$ state-dependent queueing systems with finite buffer size and obtain the steady-state distribution of the number of people in both continuous time and at arrival epochs.

The rest of the paper is organized as follows. In Section 2 we briefly review the related literature. In Section 3 we define the problem and discuss some preliminary results. In Section 4 we model and analyze the $M_n/G_n/1$ system at both arbitrary times and arrival epochs. We extend the results to $M_n/G_n/1/K$ in Section 5. We summarize the paper in Section 6. All proofs not in the body of the paper appear in the Appendix.

2. Literature Review

In classical single server queueing systems it is assumed that customers who are looking for a particular service arrive to the system according to a stochastic process and service times are uncertain. One of the main assumptions in this literature is that the arrival and service rates are constants and do not depend on the state of the system (see e.g., Kleinrock, 1975, Cohen, 1982, Asmussen, 1991, Tijms, 2003, Gross and Harris, 2011).

Queueing systems with state-dependent arrival and service rates have been studied in the literature since Harris (1967). He provides the probability distribution of the number of people in the system for $M/M_n/1$ queueing systems where the rate of the service is $\mu_n = n\mu$. He also derives the probability distribution of the number of people in two-state $M/M/1$ systems where the service time of a customer depends on whether there are any other customers in the system or not at the onset of their service. Shanthikumar (1979) considers a two-state state-dependent $M/G/1$ queueing system and obtains the Laplace Transform (LT) of the steady-state waiting time distribution in such a system. Regterschot and De Smit (1986) analyze an $M/G/1$ queueing system with Markov modulated arrivals and service times. Gupta and Rao (1998) consider a queueing system with finite buffer where the arrival rates and service times depend on the number of people in the system. They assume that the service times can be adjusted only at the beginning of the service and obtain the distribution of the number of people in the system in continuous time and at arrival epochs. Kerner (2008) considers a state-dependent $M_n/G/1$ queueing system and provides a closed form expression for the probability distribution of the number of people in the system as a function of the probability that the server is idle. But in contrast to our results, he could not derive this probability and doesn't allow state-dependent service times. The derivations in both Gupta and Rao (1998) and Kerner (2008) are special restrictive cases of our results.

Workload-dependent queueing systems in which the arrivals and service times depend on the workload of the system rather than the number of people in the system have been also studied in the literature since early 1960's (see e.g., Gaver and Miller 1962). For example, Bekker et al. (2004) consider a work-load dependent queueing system where both arrival rate and service speed depend on the workload of the system. Assuming that the ratio of arrival rate and service speed is equal in an $M/G/1$ system, they derive the steady-state distribution of the workload. Bekker and Boxma (2007) consider an $M/G/1$ system where the service speed only changes at discrete points of arrivals. Considering the case of an N-step service speed function, they obtain the distribution of the workload right after and right before arrival epochs. Bekker et al. (2011) derive the steady-state waiting time distribution of Markovian systems in which the service speed depends on the waiting time of the first customer in line.

Another stream of related literature is papers using the supplementary variable method to derive the analytical results. For example, Hokstad (1975a) considers an $M/G/1$ system and obtains the joint distribution of the number of customers present in the system and the residual service time using the supplementary variable technique. Hokstad (1975b) extends his results to an $G/M/m$ system. (See Cohen, 1982, and Kerner, 2008, for more detail about this literature.)

3. State-Dependent Queueing System

In this section we first define the problem considered in this paper precisely and explain some preliminaries of this state-dependent queueing system. Then, we present known results for Markovian systems to highlight the parallelism between the results we obtain for the $M_n/G_n/1$ and those for the $M_n/M_n/1$.

3.1. Problem Definition and Preliminaries

We study a single server queue with state-dependent arrival rates and service times. Customers arrive to the system according to a Poisson process with a rate λ_n when there are n customers in the system. The service times are also state-dependent. Specifically, when there are n customers in the system and a new service time starts, it is generally distributed with a mean $1/\mu_n$, a density function $b_n(\cdot)$, and Laplace Transform (LT) $\tilde{b}_n(\cdot)$. We assume that $b_n(\cdot)$ is absolutely continuous. The density function $b_n(\cdot)$ may be completely different than $b_{n+1}(\cdot)$, e.g., $b_{n+1}(\cdot)$ can be uniformly distributed while $b_n(\cdot)$ is exponentially distributed.

We also allow the service rate to change when a new customer arrives to the system as follows: when there are $n \geq 1$ customers in the system and a new arrival occurs, the rate of the service changes by a factor of $\alpha_{n+1} > 0$. This will lead to a reduction in the residual service of the customer in service by this factor. (Note that if $\alpha_{n+1} < 1$ the service rate decreases and the residual service time actually increases.) Specifically, an arrival that sees n customers in the system and a remaining service time of η for the customer in service, causes the service rate to immediately change such that the remaining service time becomes $\frac{\eta}{\alpha_{n+1}}$ (assuming no future arrivals before the end of the current service, as such arrivals will again change the residual service time by α_{n+2}). That is, if the residual service time when a new customer arrives, η, has a density $f(\eta)$, the remaining service time after this arrival will change to $\frac{1}{\alpha_{n+1}}\eta$ and therefore it will have a density of

$$\frac{1}{\alpha_{n+1}} f(\frac{1}{\alpha_{n+1}}\eta). \tag{1}$$

We assume that both μ_n and α_n are finite and greater than zero, which implies that the means of all service times are finite. We further assume a non-idling policy, i.e., the server starts working as soon as there is a customer in the system and is only idle when the system is empty.

Assuming the system detailed above is stable, let $P(i)$ denote the steady-state probability of having i customers in the system, and $\frac{1}{\hat{\mu}_i}$ denote the expected service time given there are i customers in the system. Note that expressing $\hat{\mu}_i$ in our setting is not trivial because of the service rate changes allowed. We will characterize it in Section 4. Moreover, assuming the system is stable is equivalent to assuming that the utilization of the system is less than 1 or that the probability that the server is idle, $P(0)$, is greater than zero (e.g., Asmussen, 1991)

$$P(0) = 1 - \sum_{i=0}^{\infty} P(i) \frac{\lambda_i}{\hat{\mu}_{i+1}} > 0. \tag{2}$$

Condition (2) is based on the steady-state probabilities $P(i)$ that require the stability condition to be obtained. For the $M_n/G_n/1$ system, we present the necessary and sufficient condition for the stability of this system. A sufficient condition for the stability of this system is to have $\frac{\lambda_i}{\hat{\mu}_{i+1}} < 1$ for every $i \geq C$ where C is a finite positive number (see e.g., Wang, 1994). The reason is that $\frac{\lambda_i}{\hat{\mu}_{i+1}} < 1$ ensures that the transient probability of being in state C of the system is positive (since $\frac{1}{\hat{\mu}_i} < \infty$ and the arrival process is Poisson, the transit probability of being in any state $i < C$ is positive as well).

Note that the PASTA property does not hold in the $M_n/G_n/1$ system that we consider. The reason is that the rate of the arrival process depends on the number of people in the system, so future arrivals are not independent of the past and present states of the system. Therefore, the Lack of Anticipation Assumption (LAA) required for PASTA does not hold in such systems (For more detail on LAA and its essentiality to PASTA, see Medhi, 2002). However, conditioning on the number of customers in the system, PASTA does hold. This property that is called conditional PASTA (Van Doorn and Regterschot, 1988) will help us to analyze the systems at arrival epochs.

3.2. Markovian Systems

To highlight the parallelism between the $M_n/G_n/1$ and $M_n/M_n/1$ queues, we present the well known analysis of the standard $M_n/M_n/1$ queueing system in this section.

Suppose that customers arrive to the system according to a Poisson process with rate λ_j when there are j customers. Service times are exponentially distributed with a rate μ_j whenever there are j customers and let $\alpha_j = \frac{\mu_j}{\mu_{j-1}}$. This queue can be modeled as a standard B&D process and analyzed using the basic relation

$$\lambda_i P(i) = \mu_{i+1} P(i+1) \quad i \geq 0, \tag{3}$$

see e.g., Gross and Harris (2011), leading to:

OBSERVATION 1. Consider the $M_n/M_n/1$ system with arrival rate λ_j and service rate μ_j when there are j people in the system. Suppose that

$$\sum_{i=1}^{\infty} \frac{\lambda_0}{\lambda_i} \prod_{j=0}^{i-1} \frac{\lambda_{j+1}}{\mu_{j+1}} < \infty. \tag{4}$$

Then, the steady-state distribution of the number of people in the system is

$$P(i) = \frac{\lambda_0 P(0)}{\lambda_i} \prod_{j=0}^{i-1} \frac{\lambda_{j+1}}{\mu_{j+1}}, \tag{5}$$

where from (5) and the fact that $\sum_{i=0}^{\infty} P(i) = 1$, we obtain

$$P(0) = \frac{1}{1 + \sum_{i=1}^{\infty} \frac{\lambda_0}{\lambda_i} \prod_{j=0}^{i-1} \frac{\lambda_{j+1}}{\mu_{j+1}}}. \tag{6}$$

Note that (3) indicates that the average rate of moving from state i to state $i+1$ while the process is in state i is equal to the average rate of moving from state $i+1$ to state i while the process is in state $i+1$. This interpretation is sufficient to express the steady-state probabilities $P(i)$. Indeed from Level Crossing Theory (LCT) (Perry and Posner, 1990) for any stable system that its states change by jumps of -1, 0, and 1, we have

$$\textit{average arrival rate while in state } i * P(i) = \textit{average service rate while in state } i+1 * P(i+1). \tag{7}$$

Equation (7) implies that the solution for any queueing system where service and arrivals happen one at a time can be decomposed as in any B&D processes. Specifically, let λ_i and $\hat{\mu}_i$ denote the average arrival and service rates while the process is in state i, respectively. Then, (7) is equivalent to (3) for the $M_n/M_n/1$ system.

An important implication of the B&D representation for a Markovian queueing system is that the system can be easily decomposed to new Markovian queues at any given state. This means that the time periods in which there are at least κ people in the system can be modeled as an $M_n/M_n/1$ queue with arrival rate $\lambda_{\kappa+i}$ and service rate $\mu_{\kappa+i}$ (when there are i customers in the new system). We call this $M_n/M_n/1$ queue the **A**uxiliary queue. Let $P^A(i)$ denote the steady-state probability of having i people in the auxiliary queue. Let also $F(i) := \sum_{j=0}^{i} P(j)$. Then,

OBSERVATION 2. **Queue Decomposition (QD) in Markovian systems:** The steady-state probability of having $\kappa+i$ ($i \geq 0$) customers in the original queue and the steady-state probability of having $i \geq 0$ customers in the auxiliary queue are related as:

$$P(\kappa+i) = (1 - F(\kappa-1))\, P^A(i), \;\; i = 0, 1, ... \tag{8}$$

We obtain a similar result for the $M_n/G_n/1$ system in Section 4.1.3.

4. Analysis of $M_n/G_n/1$ Queues

In this section, we first analyze the state-dependent queueing systems with general service time distribution at both arbitrary times and arrival epochs. Then, we demonstrate that the number of customers in the $M_n/G_n/1$ system is identical in distribution to this number in a specific B&D process and obtain the transition rates of this process.

4.1. Time Average Analysis

To analyze the $M_n/G_n/1$ system at an arbitrary time, we use the method of supplementary variable introduced by Cox (1955), see also Chapter II.6 in Cohen (1982) and Hokstad (1975a). To model the system using this method, we consider a pair of variables n_t and η_t where n_t and η_t denote the number of customers in the system and the remaining service time of the customer in service, respectively. Note that η_t is called the supplementary variable; but as you will see next, it has an important role in characterizing the distribution of the number of customers in the system. Let $p_n(\eta, t)$ denote the probability-density of having n customers in the system when the residual service time of the customer in service is η at time t so that:

$$p_0(t) = P(n_t = 0), \tag{9}$$

$$p_n(\eta,t)d\eta = P(n_t = n) \cap [(\eta < \eta_t \leq \eta + d\eta)] \quad n = 1, 2, 3, ... \tag{10}$$

We have:

THEOREM 1. *The Chapman-Kolmogorov equations that describe the dynamic of the pair* $\{(n_t, \eta_t), t \in [0, \infty)\}$ *in our* $M_n/G_n/1$ *system are given by:*

$$p_0(t+dt) = p_0(t)(1-\lambda_0 dt) + p_1(0,t)dt + o(dt). \tag{11}$$

$$p_1(\eta - dt, t+dt) = p_1(\eta,t)(1-\lambda_1 dt) + p_2(0,t)b_1(\eta)dt + p_0(t)\lambda_0 b_1(\eta)dt + o(dt). \tag{12}$$

$$p_j(\eta - dt, t+dt) = p_j(\eta,t)(1-\lambda_j dt) + p_{j+1}(0,t)b_j(\eta)dt + \alpha_j p_{j-1}(\eta\alpha_j, t)\lambda_{j-1}dt + o(dt). \tag{13}$$

Proof.

Based on the definition of the supplementary variable that we use, we follow Hokstad (1975a) to derive a set of relations for a small time interval $(t, t+dt)$ considering the pair $\{(n_t, \eta_t), t \in [0, \infty)\}$.

First, consider $p_0(t+dt)$: Note that the continuous time MC for state 0 is identical to the one in the $M/G/1$ systems (the system is idle in both cases) and is given in Hokstad (1975a) as (11).

Next consider state $(1, \eta - dt)$: If at time $t + dt$ the system is in state $(1, \eta - dt)$ where $\eta - dt \geq 0$, at time t one of the following has happened: 1) the system was in state $(1, \eta)$ and no new customer arrived during the next dt units of time (with probability of $1 - \lambda_1 dt + o(dt)$); 2) the system was in state $(2, 0)$ and the service time of the next customer to enter service was η (with probability of $b_1(\eta)dt$); 3) the system was in state (0) and a new customer arrived during the next dt units of time (with probability of $\lambda_0 dt + o(dt)$); or 4) other possible events with probability $o(dt)$ or lower. Therefore,

$$p_1(\eta - dt, t + dt) = p_1(\eta, t)\left(1 - \lambda_1 dt + o(dt)\right) + p_2(0, t)b_1(\eta)dt + p_0(t)\left(\lambda_0 dt + o(dt)\right)b_1(\eta) + o(dt).$$

Combining all terms with order of dt^2 in $o(dt)$, we get (12).

Finally consider state (j, η) for $j > 1$: If at time $t + dt$ the system state is $(j, \eta - dt)$, at time t one of the following has happened: 1) the system was in state (j, η) and no new customer arrived during the next dt units of time (with probability of $1 - \lambda_j dt + o(dt)$); 2) the system was in state $(j+1, 0)$ and the service time of the next customer was η (with probability of $b_j(\eta)dt$); 3) the system was in state $(j-1, \eta\alpha_j)$ and a new customer arrived during the next dt units of time (with probability of $\lambda_{j-1} dt + o(dt)$); or 4) other possible events with probability $o(dt)$ or lower. Thus,

$$p_j(\eta - dt, t + dt) = p_j(\eta, t)(1 - \lambda_j dt + o(dt)) + p_{j+1}(0, t)b_j(\eta)dt + \alpha_j p_{j-1}(t, \eta\alpha_j)\left(\lambda_{j-1} dt + o(dt)\right) + o(dt).$$

Note that since $p_{j-1}(t, \eta)$ is a probability density function, the coefficient α_j in $\alpha_j p_{j-1}(t, \eta\alpha_j)$ ensures $\alpha_j p_{j-1}(t, \eta\alpha_j)$ is also a probability density function. Combining all terms with order of dt^2 in $o(dt)$, we get (13).

The pair (n_t, η_t) is a vector valued Markov process that represents the state of the system at any given time $t \in [0, \infty)$. Therefore, (11 -13) present a continuous time Markov Chain (MC) for the $M_n/G_n/1$ system. ■

Theorem 1 generalized the continuous time MC for the $M_n/G/1$ from Kerner (2008) to the $M_n/G_n/1$ system we consider.

4.1.1. Distribution of Number of People in the System Using the MC presented in Theorem 1, we derive the steady-state distribution of the number of people in the corresponding $M_n/G_n/1$ system. Let $P(i)$ denote the steady-state probability of having i customers in the $M_n/G_n/1$ system. Also, let $\tilde{h}_j(\cdot)$ denote the LT of the steady-state residual service time given that there are $j \geq 0$ customers in the system where $\tilde{h}_0(\cdot) = \tilde{b}_1(\cdot)$. Assuming $\tilde{h}_j(\cdot)$ are given and setting $\mu_0 = \mu_1$, $\alpha_1 = 1$, $P(i)$ are derived in the following theorem.

Theorem 2. *Suppose that*

$$\sum_{i=1}^{\infty}\frac{\lambda_0}{\lambda_i}\prod_{j=0}^{i-1}\frac{1-\tilde{h}_j\left(\frac{\lambda_{j+1}}{\alpha_{j+1}}\right)}{\tilde{b}_{j+1}(\lambda_{j+1})}<\infty. \tag{14}$$

Then, the steady-state distribution of the number of people in an $M_n/G_n/1$ queue is

$$P(i)=\frac{\lambda_0 P(0)}{\lambda_i}\prod_{j=0}^{i-1}\frac{1-\tilde{h}_j\left(\frac{\lambda_{j+1}}{\alpha_{j+1}}\right)}{\tilde{b}_{j+1}(\lambda_{j+1})}, \tag{15}$$

where from (15) and $\sum_{i=0}^{\infty}P(i)=1$, we have

$$P(0)=\frac{1}{1+\sum_{i=1}^{\infty}\frac{\lambda_0}{\lambda_i}\prod_{j=0}^{i-1}\frac{1-\tilde{h}_j\left(\frac{\lambda_{j+1}}{\alpha_{j+1}}\right)}{\tilde{b}_{j+1}(\lambda_{j+1})}}. \tag{16}$$

Condition (14) guarantees that the steady-state probability of having no customers in the system, $P(0)$, is larger than zero; it is necessary and sufficient for the stability of our $M_n/G_n/1$ system.

We next obtain $\tilde{h}_i(\cdot)$ recursively assuming that the system is stable.

Theorem 3. *Suppose that the $M_n/G_n/1$ system is stable. Then, the LT of the steady-state distribution of the residual service time given that there are i customers in the system can be calculated recursively from:*

$$\tilde{h}_i(s) = \frac{\lambda_i}{s-\lambda_i}[\tilde{b}_i(\lambda_i)\frac{1-\tilde{h}_{i-1}(\frac{s}{\alpha_i})}{1-\tilde{h}_{i-1}(\frac{\lambda_i}{\alpha_i})}-\tilde{b}_i(s)] \quad i\geq 1, \tag{17}$$

where $\tilde{h}_0(\cdot)=\tilde{b}_1(\cdot)$.

To characterize $\tilde{h}_i(\cdot)$ we assume the system is stable. But, this condition in (14) is a function of $\tilde{h}_i(\cdot)$. To verify the stability condition, one should assume that $\tilde{h}_i(\cdot)$ obtained in Theorem 3 are well defined, calculate them recursively, and then check if (14) holds. If (14) holds, then the system is stable and therefore $\tilde{h}_i(\cdot)$ are well defined.

The calculation of the LT in (17) depends on the distribution of the service time. In general, when the expected residual service time approaches zero, this calculation takes longer. For example, for a system with deterministic service time, obtaining the probability of having a large number of people in the system may require a higher accuracy of the software used and take longer (as the residual service times approach zero) compare to calculating the probability of having a small number of people in the system.

4.1.2. Modeling the $M_n/G_n/1$ Queues as a Birth-and-Death Process Let $\hat{\mu}_i$ denote the steady-state rate at which the system moves from state i to $i-1$ when i is the number of customers in the system. In the following observation, we obtain this rate.

Observation 3. The steady-state number of customers in the $M_n/G_n/1$ system has the same distribution as the steady-state number of customers in a B&D $M_n/M_n/1$ process with arrival rate λ_i and service rate

$$\hat{\mu}_i = \frac{\lambda_i \tilde{b}_i(\lambda_i)}{1 - \tilde{h}_{i-1}\left(\frac{\lambda_i}{\alpha_i}\right)} \tag{18}$$

when there are i people in the system.

Using Observation 3 we rewrite the stability condition (14) as

$$\sum_{i=1}^{\infty} \frac{\lambda_0}{\lambda_i} \prod_{j=0}^{i-1} \frac{\lambda_{j+1}}{\hat{\mu}_{j+1}} < \infty, \tag{19}$$

the probability that there are i customers in the system as

$$P(i) = \frac{\lambda_0 P(0)}{\lambda_i} \prod_{j=0}^{i-1} \frac{\lambda_{j+1}}{\hat{\mu}_{j+1}}, \tag{20}$$

and the probability that the $M_n/G_n/1$ system is idle as

$$P(0) = \frac{1}{1 + \sum_{i=1}^{\infty} \frac{\lambda_0}{\lambda_i} \prod_{j=0}^{i-1} \frac{\lambda_{j+1}}{\hat{\mu}_{j+1}}}. \tag{21}$$

Comparing (19-21) with (4-6) we observe that the stability condition and the distribution of the number of people in the $M_n/G_n/1$ has the same structure as the one in the $M_n/M_n/1$. This similarity indicates that the solution of the $M_n/G_n/1$ can be decomposed as in any B&D processes.

The expressions in (19-21) generalize the ones in equation (12) of Kerner (2008) for the $M_n/G/1$ to the $M_n/G_n/1$ queueing system.

4.1.3. Queue Decomposition (QD) at a Given State In Observation 2, we showed that the $M_n/M_n/1$ queueing system can be decomposed at any given state. In this section we extend this result to the $M_n/G_n/1$ system. To decompose this system, as in Abouee-Mehrizi et al. (2012), we define an auxiliary queue such that the steady-state probability of having $\kappa+i$ customers in the original system given that there are κ or more people in this system is identical to the steady-sate distribution of having i jobs in this auxiliary queue. To distinguish between the original queue and the auxiliary queue, we use the term "job" in the auxiliary queue. We define the auxiliary queue as an $M_n/G_n/1$ queue with the following (a) arrival and (b) service processes:

Step (a): jobs arrive to the auxiliary queue according to a Poisson process with rate $\lambda_{\kappa+i}$ for $i \geq 0$ when there are i people in the auxiliary queue.

Step (b): the distribution of the first service time in each busy period of the auxiliary queue is the distribution of the conditional residual service time in the original queue given that there are κ customers in the system, i.e., the equilibrium remaining service times given that there are κ customers in the system. The distribution of the rest of the service times in each busy period of the auxiliary queue is identical to the original queue, i.e., $b_{\kappa+i}(\cdot)$ for $i \geq 0$ when there are i people in the auxiliary queue. Moreover, when there are i customers in the auxiliary queue and a new arrival occurs, the rate of the service changes by a factor of $\alpha_{\kappa+i+1} > 0$.

Since the auxiliary queue is an $M_n/G_n/1$ queue, the steady-state distribution of the number of jobs in this queue, $P^A(i)$, can be obtained using Theorem 2 as:

$$P^A(i) = P^A(0) \prod_{j=0}^{i-1} \frac{1 - \tilde{h}_{\kappa+j}(\lambda_\kappa)}{\tilde{b}_\kappa(\lambda_\kappa)}. \tag{22}$$

Recall that in the original $M_n/G_n/1$ system we let $F(i) := \sum_{j=0}^{i} P(j)$. Then,

Observation 4. **QD in $M_n/G_n/1$ systems:** The steady-state probability of having $\kappa + i$ ($i \geq 0$) customers in the original $M_n/G_n/1$ system and the steady-state probability of having $i \geq 0$ customers in the auxiliary queue are related as:

$$P(\kappa + i) = (1 - F(\kappa - 1))\, P^A(i), \;\; i = 0, 1, \ldots \tag{23}$$

Observation 4 states that the steady-state number of jobs in the auxiliary queue is identical to the steady-state number of customers in the original state-dependent queue during the time intervals when there are more than $\kappa - 1$ customers in the original $M_n/G_n/1$ system.

4.2. Analysis at Arrival Epochs

In this section, we analyze the $M_n/G_n/1$ at arrival epochs. We remind that since in this system the arrival process is state-dependent, the PASTA property does not hold. Therefore, the steady-state distribution of the number of customers seen by an arrival is typically different than the steady-state distribution of the number of customers in the system at an arbitrary time.

Let $\bar{P}^a(n)$ denote the steady-state probability that an **A**rrival observes n customers in the system. To obtain $\bar{P}^a(n)$, we recall that in the $M_n/G_n/1$ the distribution of the number of customers seen by an arrival is identical to the steady-state distribution of the number of customers seen by a departure (this easily follows by a level crossing argument as in Buzacott and Shanthikumar, 1993). Therefore, we analyze the system at departure epochs by defining an embedded MC.

Let M_n denote the number of customers left behind by the n^{th} departing customer in the system. M_n can be found by the MC embedded at the **D**epartures. Let $\bar{P}^d(n)$ denote the steady-state distribution of being in state n of this MC. Then,

$$\bar{P}^a(n) = \bar{P}^d(n). \tag{24}$$

To derive $\bar{P}^d(n)$, we need to determine the one-step transition probabilities of M_n,

$$p_{jk} = P\left(M_{n+1} = k | M_n = j\right). \tag{25}$$

For state $j \geq 1$, let v_k^j denote the probability of having $k \geq 0$ arrivals during the service time that starts when there are j customers in the system. For $j = 0$, the relevant service time starts when there is one customer in the system. So we let $v_k^0 = v_k^1$. Then, the probability that the next departure leaves k customers behind given that there are no customers in the system, p_{0k}, is v_k^0. Similarly, the probability that the next departure leaves k customers behind given that there are $j \geq 1$ customers in the system, p_{jk}, is v_{k-j+1}^j. To summarize, we denote the one-step transition probabilities of M_n, the embedded MC, as

$$p_{jk} = \begin{cases} v_k^0, & j = 0; \\ v_{k-j+1}^j, & j \geq 1. \end{cases} \tag{26}$$

In the standard $M/G/1$, these probabilities can be easily obtained since the distribution of the service times is identical for all customers. Let λ and $\tilde{b}(\cdot)$ denote the arrival rate and the LT of the service time distribution in the standard $M/G/1$, respectively. Then, v_k^j is independent of j and the probability generating function of $v_k := v_k^j$ $(j = 0, 1, ...)$ is (see e.g., Takagi 1991)

$$V(z) = \sum_{k=0}^{\infty} v_k z^k = \tilde{b}(\lambda(1 - z)). \tag{27}$$

To obtain v_k^j in the $M_n/G_n/1$ queue, we consider the probability of a new arrival during the residual service time observed by any customer upon arrival. Fortunately, this distribution possesses the conditional PASTA property:

Corollary 1. *The conditional steady-state distribution of the residual service time observed by arrivals that find j customers in the system is identical to the conditional steady-state distribution of the residual service time at an arbitrary time in the $M_n/G_n/1$ system that also observes j customers in the system.*

We next determine the transition probabilities, p_{jk}, for $j > 0$. (The derivation for p_{0k} is similar but more detailed and it is provided in the proof of Theorem 4.) Consider $p_{j,j-1} = P(M_{n+1} = j-1 | M_n = j)$, the probability that the next departing customer leaves one less customer behind given that the last departing customer left. This is equal to the probability of no arrival during the next service time, $\tilde{b}_j(\lambda_j)$ (e.g., Conway 1967, page 171). Therefore,

$$p_{j,j-1} = \tilde{b}_j(\lambda_j), \quad j \geq 1. \tag{28}$$

With similar logic, the probability that the next departing customer leaves k customers behind given that the last departing customer left $j \geq 1$ customers behind, $P(M_{n+1} = k | M_n = j)$, is equal to the probability of $k-j+1$ arrivals during the next service time. This probability is equal to the probability that a customer arrives after the next service time starts, $\left(1 - \tilde{b}_j(\lambda_j)\right)$, and a customer arrives during the remaining service time of all arrivals that see $i < k$ customers in the system, $\left(1 - \tilde{h}_i(\frac{\lambda_{i+1}}{\alpha_{i+1}})\right)$, and no arrival during the remaining service time once there are k customers in the system, $\tilde{h}_k(\frac{\lambda_{k+1}}{\alpha_{k+1}})$:

$$p_{jk} = \tilde{h}_k\left(\frac{\lambda_{k+1}}{\alpha_{k+1}}\right)\left(1 - \tilde{b}_j(\lambda_j)\right)\prod_{i=j}^{k-1}\left(1 - \tilde{h}_i\left(\frac{\lambda_{i+1}}{\alpha_{i+1}}\right)\right), \quad j \geq 1, k \geq j. \tag{29}$$

Note that for the $M/M/1$ queue we have $\tilde{h}_j(\lambda_j) = \tilde{b}_i(\lambda_i) = \tilde{b}(\lambda) = \frac{\mu}{\mu+\lambda}$ so that

$$p_{jk} = \frac{\mu}{\mu+\lambda}\left(\frac{\lambda}{\mu+\lambda}\right)^{k-j+1}, \quad j \geq 1, k \geq j. \tag{30}$$

As expected due to memoryless property, PASTA, and that the minimum of two independent exponential random variables is an exponential random variable, in the $M/M/1$ setting p_{jk} has a Geometric distribution with parameter $\frac{\mu}{\mu+\lambda}$.

Using the transition probabilities in (29) (and these for p_{0j}), the steady-state distribution of the number of customers in the system observed by an arrival is as follows.

THEOREM 4. *Suppose that*

$$\sum_{i=1}^{\infty}\prod_{j=0}^{i-1}\frac{1 - \tilde{h}_j\left(\frac{\lambda_{j+1}}{\alpha_{j+1}}\right)}{\tilde{b}_{j+1}(\lambda_{j+1})} < \infty. \tag{31}$$

Then, the steady-state distribution of the number of people in an $M_n/G_n/1$ queue observed by an arrival is

$$\bar{P}^a(i) = \bar{P}^a(0)\prod_{j=0}^{i-1}\frac{1 - \tilde{h}_j\left(\frac{\lambda_{j+1}}{\alpha_{j+1}}\right)}{\tilde{b}_{j+1}(\lambda_{j+1})}, \tag{32}$$

where from (32) and $\sum_{i=0}^{\infty}\bar{P}^a(i)=1$*, we have*

$$\bar{P}^a(0)=\frac{1}{1+\sum_{i=1}^{\infty}\prod_{j=0}^{i-1}\frac{1-\tilde{h}_j\left(\frac{\lambda_{j+1}}{\alpha_{j+1}}\right)}{\tilde{b}_{j+1}(\lambda_{j+1})}}. \tag{33}$$

Comparing (15) and (32), we next relate the steady-state probability of having i customers in the system at an arrival epoch to the steady-state probability of having i customers in the system at an arbitrary time.

Observation 5. In an $M_n/G_n/1$ system, the relation between the steady-state probability of having i customers in the system at arrival (or departure) epochs and at arbitrary times is given by:

$$\bar{P}^a(i)=\frac{\lambda_i P(i)}{\sum_{k=0}^{\infty}\lambda_k P(k)}. \tag{34}$$

Note that when arrival rates are identical, $\lambda_i=\lambda$ for $i=0,1,...$, the PASTA property holds and therefore $\bar{P}^a(i)=P(i)$.

Observation 5 indicates that $\bar{P}^a(i)$ can be obtained using $P(i)$ given in Theorem 2. More interestingly, Observation 5 together with Theorem 4 provide an alternative proof for Theorem 2 using that:

Lemma 1. *In an $M_n/G_n/1$ system, the relation between the steady-state probability of having i customers in the system at arbitrary times and at arrival (or departure) times is given by:*

$$P(i)=P(0)\frac{\lambda_0\bar{P}^a(i)}{\lambda_i\bar{P}^a(0)} \quad i\geq 0. \tag{35}$$

Similar to the analysis in Section 4.1.2, we next obtain the steady-state service rate of a customer observed at arrival epochs when there are i customers in the system, $\hat{\mu}_i^a$.

Observation 6. The equilibrium service rate of a customer in the $M_n/G_n/1$ system at arrival epochs when there are i customers in the system is

$$\hat{\mu}_i^a=\frac{\lambda_{i-1}}{\lambda_i}\hat{\mu}_i=\frac{\lambda_{i-1}\tilde{b}_i\left(\lambda_i\right)}{1-\tilde{h}_{i-1}\left(\frac{\lambda_i}{\alpha_i}\right)}. \tag{36}$$

Observation 6 indicates that the steady-state utilization of the system observed by arrivals who find i customers in the system, $\frac{\lambda_{i-1}}{\hat{\mu}_i^a}$, is identical to the steady-state utilization of the system when there are i customers in the system, $\frac{\lambda_i}{\hat{\mu}_i}$.

4.3. $M_n/G_n/1$ System when State-Dependence is for a Finite Number of States

In this section we obtain a closed form expression for the steady-state probability that there are no customers in the system, $P(0)$, when state-dependence is for a finite number of states. We assume that there exists a $k < \infty$ such that for any $i \geq k$, λ_i and b_i are independent of the number of people in the system, i.e., $k = \min_i\{\lambda_i, b_i(\cdot), \alpha_i : \lambda_i = \lambda_{i+1} = ..., b_i(\cdot) = b_{i+1}(\cdot) = ..., \alpha_i = \alpha_{i+1} = ... = 1\}$. This queue is state-dependent for $i < k$ and has the same arrival rate and service time distribution for $i \geq k$. We assume that $\lambda_k/\mu_k < 1$ to ensure that the system is stable.

To obtain $P(0)$, we use QD. Consider the auxiliary queue defined in Section 4.1.3 for $\kappa = k$. Let ρ_b and μ_b denote the server utilization and the rate of the first exceptional service times in this auxiliary queue, respectively. Note that the service time densities, $b_i(\cdot)$, in this system are identical for $i > k$. Therefore, the rate of the service time in the auxiliary queue is μ_k with probability ρ_b and μ_b with probability $(1 - \rho_b)$, so that

$$\rho_b = \rho_b \frac{\lambda_k}{\mu_k} + (1 - \rho_b)\frac{\lambda_k}{\mu_b}$$

leading to

$$\rho_b = \frac{\lambda_k \mu_k}{\mu_b \mu_k + \lambda_k(\mu_k - \mu_b)}. \tag{37}$$

To obtain the utilization of the auxiliary queue, ρ_b, we derive the average rate of the exceptional first service times, μ_b.

Lemma 2. *The average exceptional first service time in the busy period of the auxiliary queue is,*

$$\frac{1}{\mu_b} = \frac{1}{\mu_k} - \frac{1}{\lambda_k} + \sum_{j=1}^{k-1}\prod_{i=j}^{k-1}\frac{1}{\alpha_{i+1}}\left(\frac{1}{\mu_j} - \frac{1}{\lambda_j}\right)\frac{\tilde{b}_{i+1}(\lambda_{i+1})}{1 - \tilde{h}_i(\frac{\lambda_{i+1}}{\alpha_{i+1}})} + \frac{1}{\mu_1}\frac{\tilde{b}_1(\lambda_1)}{1 - \tilde{h}_0(\frac{\lambda_1}{\alpha_1})}\prod_{i=1}^{k-1}\frac{1}{\alpha_i}\frac{\tilde{b}_{i+1}(\lambda_{i+1})}{1 - \tilde{h}_i(\frac{\lambda_{i+1}}{\alpha_{i+1}})}. \tag{38}$$

Note that Sigman and Yechiali (2007) provide the expected conditional stationary remaining service time in an $M/G/1$ queue. Lemma 2 extends their result to the $M_n/G_n/1$ queue.

Using (15), (22), and (23) the probability of having no customers in this system can be obtained as follows.

Theorem 5. *The steady-state probability of having no customers in the $M_n/G_n/1$ system in which the arrival rates and service times are state-dependent for a finite number of states is*

$$P(0) = \frac{1 - \rho_b}{\frac{\lambda_0}{\lambda_k}\prod_{i=0}^{k-1}\frac{1 - \tilde{h}_i\left(\frac{\lambda_{i+1}}{\alpha_{i+1}}\right)}{\tilde{b}_{i+1}(\lambda_{i+1})} + (1 - \rho_b)\left(1 + \sum_{j=1}^{k-1}\frac{\lambda_0}{\lambda_j}\prod_{i=0}^{j-1}\frac{1 - \tilde{h}_i\left(\frac{\lambda_{i+1}}{\alpha_{i+1}}\right)}{\tilde{b}_{i+1}(\lambda_{i+1})}\right)}. \tag{39}$$

Substituting (39) into Theorems 2 provides a closed form expression for all steady-state probabilities for systems where the arrival rates and service times are state-dependent for a finite number of states.

5. Analysis of $M_n/G_n/1/K$

In this section, we consider a state-dependent queue with a finite buffer K, $M_n/G_n/1/K$. This queue is identical to an $M_n/G_n/1$ queueing system with $\lambda_i = 0$ for $i \geq K$ (and, of course, it has no issue of stability).

5.1. Time Average Analysis

Let $P_F(i)$ denote the steady-state probability of having i $(0 \leq i \leq K)$ customers in this system with **F**inite buffer. If the buffer size of the system is 1, $K = 1$, there is no state dependency and $\alpha_1 = 1$ by definition, thus the steady-state distribution of the number of people in the system can be obtained from $P_F(0) + P_F(1) = 1$ and $\lambda_0 P_F(0) = \mu_1 P_F(1)$ as (see e.g., Gross and Harris, 2011)

$$P_F(0) = \frac{\mu_1}{\lambda_0 + \mu_1}, \tag{40}$$

$$P_F(1) = \frac{\lambda_0}{\lambda_0 + \mu_1}. \tag{41}$$

Now consider a system with a buffer size larger than 1, $K > 1$. We make the following important observation.

OBSERVATION 7. Comparing the transitions in the $M_n/G_n/1/K$ with the ones in the $M_n/G_n/1$, we observe that the transitions up to states $i = 0, \cdots, K-1$ in both systems are identical. The difference between these two systems is that all transitions that take the $M_n/G_n/1$ to a state greater than K are lost in the $M_n/G_n/1/K$. This means that the transition rates in the states $i = 0, \cdots, K-1$ are identical in both systems.

Observation 7 emphasizes the equivalency between the transition rates in the $M_n/G_n/1/K$ and the ones in the B&D process, discussed in Observation 3, for states $i = 0, ..., K-1$. This observation enables us to solve the $M_n/G_n/1/K$ queue using the derivations for the $M_n/G_n/1$ system in a similar fashion that the $M_n/M_n/1/K$ queue is solved using the $M_n/M_n/1$ (similar to Gross and Harris, 2011, page 76):

$$P_F(i) = \frac{\lambda_0 P_F(0)}{\lambda_i} \prod_{j=0}^{i-1} \frac{\lambda_{j+1}}{\mu_{j+1}}, \quad i = 1, ..., K. \tag{42}$$

Observation 7 allows us to easily obtain the steady-state probability of having $i < K$ people in the $M_n/G_n/1/K$. We note that Gupta and Rao (1998) derive a direct solution of a special case of our finite buffer queue using the supplementary variable method that we used to solve the infinite buffer case. This method of course leads to the same solution for their special case, but is less elegant given the analysis in Section 4 and Observation 7.

Corollary 2. *The steady-state distribution of the number of people in an $M_n/G_n/1/K$ queue, $P_F(i)$, is*

$$P_F(i) = \frac{\lambda_0 P_F(0)}{\lambda_i} \prod_{j=0}^{i-1} \frac{1-\tilde{h}_j\left(\frac{\lambda_{j+1}}{\alpha_{j+1}}\right)}{\tilde{b}_{j+1}(\lambda_{j+1})}, \quad i = 1, ..., K-1 \tag{43}$$

where $\mu_0 = \mu_1$ and $\tilde{h}_i(\cdot)$ is given in (17).

We next obtain the steady-state probability of having exactly 0 and K customers in the system using an auxiliary $M/G/1/1$ queue with the following (a) arrival and (b) service processes:

Step (a): jobs arrive to the auxiliary queue according to a Poisson process with a rate of λ_{K-1}.

Step (b): the distribution of the first service time in the busy period of the auxiliary queue is the distribution of the conditional residual service time in the original $M_n/G_n/1/K$ queue given that there are $K-1$ customers in the system, i.e., the equilibrium remaining service times given that there are $K-1$ customers in the system.

Let $1/\mu_b^F$ and $P_F^A(i)$ $(i = 0, 1)$ denote the mean service times and distribution of number of people in this auxiliary queue, respectively. Then, using (40) and (41) we have,

$$P_F^A(0) = \frac{\mu_b^F}{\lambda_{K-1} + \mu_b^F}, \tag{44}$$

$$P_F^A(1) = \frac{\lambda_{K-1}}{\lambda_{K-1} + \mu_b^F}. \tag{45}$$

Therefore, to obtain $P_F^A(i)$, we need to derive the mean of the service times, $1/\mu_b^F$. Noting that Observation 7 demonstrates that the $M_n/G_n/1$ and $M_n/G_n/1/K$ are equivalent for states $i = 0, ..., K-1$, $1/\mu_b^F$ can be obtained by substituting $k = K-2$ in (38) as:

Corollary 3. *The average of the exceptional first service time in the auxiliary queue is,*

$$\begin{aligned}\frac{1}{\mu_b^F} = & \frac{1}{\mu_{K-1}} - \frac{1}{\lambda_{K-1}} + \sum_{j=1}^{K-2} \prod_{i=j}^{K-2} \frac{1}{\alpha_{i+1}} \left(\frac{1}{\mu_j} - \frac{1}{\lambda_j}\right) \frac{\tilde{b}_{i+1}(\lambda_{i+1})}{1-\tilde{h}_i(\frac{\lambda_{i+1}}{\alpha_{i+1}})} \\ & - \frac{1}{\mu_1} \frac{\tilde{b}_1(\lambda_1)}{1-\tilde{h}_0(\frac{\lambda_1}{\alpha_1})} \prod_{i=1}^{K-2} \frac{1}{\alpha_i} \frac{\tilde{b}_{i+1}(\lambda_{i+1})}{1-\tilde{h}_i(\frac{\lambda_{i+1}}{\alpha_{i+1}})}.\end{aligned} \tag{46}$$

Let $F_F(i) := \sum_{j=0}^{i} P_F(j)$. Using Observation 4 and considering that the $M_n/G_n/1$ and $M_n/G_n/1/K$ are equivalent, the probability of having $i = 0, 1$ jobs in the auxiliary queue, $P_F^A(i)$, is identical to the probability of having $K+i-1$ customers in the original queue, $P_F(K+i-1)$, given that there are more than $K-2$ customers in the system.

Corollary 4. *The steady-state probability of having $i = K-1$, K customers in an $M_n/G_n/1/K$ is*

$$P_F(K-1) = (1 - F_F(K-2))\, P_F^A(0), \tag{47}$$

$$P_F(K) = (1 - F_F(K-2))\, P_F^A(1). \tag{48}$$

Considering (43), (44) and (47), the probability that the system is empty, $P_F(0)$, is

$$P_F(0) = \frac{P_F^A(0)}{\frac{\lambda_0}{\lambda_{K-1}} \prod_{j=0}^{K-2} \frac{1-\tilde{h}_j\left(\frac{\lambda_{j+1}}{\alpha_{j+1}}\right)}{\tilde{b}_{j+1}(\lambda_{j+1})} + P_F^A(0)\left(1 + \sum_{i=1}^{K-2} \frac{\lambda_0}{\lambda_i} \prod_{j=0}^{i-1} \frac{1-\tilde{h}_j\left(\frac{\lambda_{j+1}}{\alpha_{j+1}}\right)}{\tilde{b}_{j+1}(\lambda_{j+1})}\right)}. \tag{49}$$

Substituting (49) in (43), we can obtain $P_F(i)$ for $i = 1, ..., K-1$. Therefore, $P_F(K)$ can be obtained using (48).

5.2. Analysis at Arrival Epochs

Let $\bar{P}_F^a(n)$ denote the steady-state probability that an **A**rrival observes n customers in the system. Unlike the $M_n/G_n/1$ system, in the $M_n/G_n/1/K$ system the distribution of the number of customers seen by an arrival is not identical to the steady-state distribution of the number of customers seen by a departure. The reason is that not all arriving customers are accepted to the system. Let $\bar{P}_F^d(n)$ denote the steady-state probability that a **d**eparture observes n customers behind. We first determine the relation between $\bar{P}_F^a(n)$ and $\bar{P}_F^d(n)$.

Lemma 3. *The relation between the distribution of the number of customers in the system observed by an arrival and the one seen behind a departure is*

$$\bar{P}_F^a(i) = \bar{P}_F^d(i)(1 - \bar{P}_F^a(K)) \quad i = 0, 1, ..., K-1. \tag{50}$$

Note that if the arrival rate to the system when there are K people in the system is zero, $\lambda_K = 0$, the probability that an arrival observes K customers in the system is also zero, $\bar{P}_F^a(K) = 0$; then, $\bar{P}_F^a(i) = \bar{P}_F^d(i)$ for $i = 0, ..., K-1$.

Corollary 5. *The steady-state distribution of the number of people in an $M_n/G_n/1/K$ queue observed by an arrival is*

$$\bar{P}_F^d(i) = \bar{P}_F^d(0) \prod_{j=0}^{i-1} \frac{1 - \tilde{h}_j\left(\frac{\lambda_{j+1}}{\alpha_{j+1}}\right)}{\tilde{b}_{j+1}(\lambda_{j+1})} \quad i = 0, 1, ..., K-1 \tag{51}$$

where from (51) and $\sum_{i=0}^{K-1} P_F^d(i) = 1$, we have

$$\bar{P}_F^d(0) = \frac{1}{1 + \sum_{i=1}^{K-1} \prod_{j=0}^{i-1} \frac{1-\tilde{h}_j\left(\frac{\lambda_{j+1}}{\alpha_{j+1}}\right)}{\tilde{b}_{j+1}(\lambda_{j+1})}}. \tag{52}$$

Considering Lemma 3 and Corollary 5, we can obtain the steady-state probability that an arrival observes n ($n < K$) customers in the system, $\bar{P}_F^a(n)$. To obtain $\bar{P}_F^a(K)$, we can use the auxiliary queue defined in Section 5.1. Then, considering (48), we get

$$\bar{P}_F^a(K) = \left(1 - \bar{F}_F(K-2)\right) P_F^A(1), \tag{53}$$

where $\bar{F}_F(i) := \sum_{j=0}^{i} \bar{P}_F^a(j)$.

6. Summary

In this paper we considered state-dependent queueing systems where the arrival rates and service times depend on the number of customers in the system. We allowed the service rate to change at arrivals and the distribution of the service times to change when a new service starts. We analyzed such systems at both arbitrary times and arrival epochs and obtained the steady-state distribution of the number of customers in the system. We showed that the $M_n/G_n/1$ systems can be decomposed as in the standard B&D process. We also demonstrated that the state-dependent queueing system with general service time distribution can be decomposed to new queues at any given state. We note that obtaining the distribution of the sojourn time in the $M_n/G_n/1$ model is not straightforward. This follows because Littles law (and Littles distributional law) cannot be applied since the future arrivals affect both the arrival and service rates (see e.g., Bertsimas and Nakazato, 1995).

References

Abouee-Mehrizi, H., B. Balcioglu, O. Baron 2012. “Strategies for a Single Product Multi-Class $M/G/1$ Queueing System”, *Operations Research,* Vol. 60, 803–812.

Altman, E., K. Avrachenkov, C. Barakat, R. Nunez-Queija 2002. “State-Dependent M/G/1 Type Queueing Analysis for Congestion Control in Data Networks”, *Computer Network* Vol. 39, 789–808.

Asmussen, S. 1991. “Ladder Heights and the Markov-Modulated $M/G/1$ Queue”, *Stochastic Processes and their Applications* Vol. 37, 313–326.

Baron, O. 2008. “Regulated random walks and the LCFS backlog probability: Analysis and applications”, *Operations Research*, Vol. 56, 471-486.

Bekker, R., S. C. Borst, O. J. Boxma, O. Kella 2004. “Queues with Workload-Dependent Arrival and Service Rates”, *Queueing Systems* Vol. 46, 537–556.

Bekker, R., O. J. Boxma 2007. “An M/G/1 queue with adaptable service speed”, *Stochastic Models* Vol. 23, 373–396.

Bekker, R., G. M. Koole, B. F. Nielsen, T. B. Nielsen 2011. “Queues with waiting time dependent service”, *Queueing Systems* Vol. 68, 61–78.

Bertsimas, D., D. Nakazato, 1995. “The Distributional Little?s Law and its Applications”, *Operations Research* Vol. 43, 298–310.

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

Cohen, J. W. 1982. The Single Server Queue, North-Holland, Publishing Company.

Conway, R. W., W. L. Maxwell, L. W. Miller 1967. *Theory of Scheduling*, Dover Publications.

Cox, D. R., 1955. "A Use of Complex Probabilities in the Theory of Stochastic Processes", *Proceedings of the Cambridge Philosophical Society*, Vol. 51, 313–319.

Vb Doorn, E. A., G. J. K. Regterschot. 1988. "Conditional PASTA", *OR Letters*, Vol. 7, 229–232.

Gaver, D.P., and R.G. Miller. 1962. "Limiting distributions for some storage problems", *In: Studies in Applied Probability and Management Science*, 110–126.

Gross, D., C. M.,Harris. 2011. Fundamentals of Queueing Theory, John Wiley & Sons, New York.

Gupta, U. C., T.S.S. Srinivasa Rao 1998. "On the Analysis of Single Server Finite Queue with State Dependent Arrival and Service Processes: $M_n/G_n/1/K$", *OR Spektrum* Vol. 20, 83–89.

Harris, C.M. 1967. "Queues with State-Dependent Stochastic Service Rates," *Operations Research*, Vol. 15, 117–130.

Hokstad, P. 1975a. "THE G/M/m QUEUE WITH FINITE WAITING ROOM," *Scand J Statist*, Vol. 12, 779–792.

Hokstad, P. 1975b. "A Supplementary Variable Technique Applied to the $M/G/1$ Queue," *Scand J Statist*, Vol. 2, 95–98.

Kerner, Y. 2008. "The Conditional Distribution of the Residual Service Time in the $M_n/G/1$ Queue," *Stochastic Models*, Vol. 24, 364–375.

Kleinrock, L. 1975. Queueing Systems, John Wiley & Sons, New York.

Medhi, J. 2002. Stochastic Models in Queueing Theory, Academic Press.

Perry, D., M. J. M. Posner 1990. "Control of Input and Demand Rates in Inventory Systems of Perishable Commodities", *Operations Research* Vol. 37, 85–97.

Sigman, K., U. Yechiali. 2007. "Stationary remaining service time conditional on queue length", *OR Letters*, Vol. 35, 581–583.

Shanthikumar, J. G. 1979. "On a single-server queue with state-dependent service," *Naval Research Logistics*, Vol. 26, 305–309.

Regterschot, G. J. K., J. H. A. de Smit 1986. "The Queue $M/G/1$ with Markov Modulated Arrivals and Services," *Mathematics of Operations Research*, Vol. 11, 465–483.

Takagi, H. 1993. *Queueing Analysis*, Volume 2, Elsevier: North Holland, The Netherlands.

Tijms, H.C. 2003. A First Course in Stochastic Models, John Wiley & Sons, Chichester.

Tijms, H.C., Van Hoorn, M.H., Federgruen, A. 1981. "Approximations for the steady-state probabilities in the M/G/c queue," *Advanced Applied Probability*, Vol. 13, 186–206.

Wang, P. P. 1996. "Queueing Models with Delayed State-Dependent Service Times", *European Journal of Operational Research* Vol. 88, 614–621.

Zhao, X., He, Q., Zhang, H. 2012. “On a Queueing System with State-dependent Service Rate,” *working paper*.

7. Proofs

Proof of Observation 2. Note that the Auxiliary queue is defined as an $M_n/M_n/1$ queue with arrival rate $\lambda_{\kappa+i}$ and service rate $\mu_{\kappa+i}$ when there are i customers in this queue. Therefore, the steady-state distribution of the number of people in this queue can be obtained using (5) and (6). Using (5), the probability of having $\kappa+i$ people in the original queue, $P(\kappa+i)$, is

$$\begin{aligned} P(\kappa+i) &= \frac{\lambda_0 P(0)}{\lambda_{\kappa+i}} \prod_{j=0}^{\kappa+i-1} \frac{\lambda_{j+1}}{\mu_{j+1}} = \frac{\lambda_0 P(0)}{\lambda_{\kappa+i}} \left(\prod_{j=0}^{\kappa-1} \frac{\lambda_{j+1}}{\mu_{j+1}} \right) \left(\prod_{j=k}^{\kappa+i-1} \frac{\lambda_{j+1}}{\mu_{j+1}} \right) \\ &= \left(\frac{\lambda_0 P(0)}{\lambda_\kappa P^A(0)} \prod_{j=0}^{\kappa-1} \frac{\lambda_{j+1}}{\mu_{j+1}} \right) \left(\frac{\lambda_\kappa P^A(0)}{\lambda_{\kappa+i}} \prod_{j=0}^{i-1} \frac{\lambda_{\kappa+j+1}}{\mu_{\kappa+j+1}} \right), \end{aligned} \tag{54}$$

where $P^A(0)$ denotes the probability of having no customers in the auxiliary queue.

We next prove that the first term in (54) is equal to $(1-F(\kappa-1))$. Substituting $P^A(0)$ given in (6) into the first term in (54), we get

$$\begin{aligned} &\left(\frac{\lambda_0 P(0)}{\lambda_\kappa P^A(0)} \prod_{j=0}^{\kappa-1} \frac{\lambda_{j+1}}{\mu_{j+1}} \right) = \left(\frac{\lambda_0 P(0)}{\lambda_\kappa} \prod_{j=0}^{\kappa-1} \frac{\lambda_{j+1}}{\mu_{j+1}} \right) \left(1 + \sum_{i=1}^{\infty} \frac{\lambda_\kappa}{\lambda_{\kappa+i}} \prod_{j=0}^{i-1} \frac{\lambda_{\kappa+j+1}}{\mu_{\kappa+j+1}} \right) \\ = &\left(\frac{\lambda_0 P(0)}{\lambda_\kappa} \prod_{j=0}^{\kappa-1} \frac{\lambda_{j+1}}{\mu_{j+1}} \right) \left(1 + \sum_{i=\kappa+1}^{\infty} \frac{\lambda_\kappa}{\lambda_i} \prod_{j=\kappa}^{i-1} \frac{\lambda_{j+1}}{\mu_{j+1}} \right) \\ = &\left(\frac{\lambda_0 P(0)}{\lambda_\kappa} \prod_{j=0}^{\kappa-1} \frac{\lambda_{j+1}}{\mu_{j+1}} + \sum_{i=\kappa+1}^{\infty} \frac{\lambda_0 P(0)}{\lambda_i} \prod_{j=0}^{i-1} \frac{\lambda_{j+1}}{\mu_{j+1}} \right) = \sum_{i=\kappa}^{\infty} \frac{\lambda_0 P(0)}{\lambda_i} \prod_{j=0}^{i-1} \frac{\lambda_{j+1}}{\mu_{j+1}} = (1-F(\kappa-1)). \end{aligned}$$

Comparing the second term in (54) with (5), we find that the former is the probability of having i people in the Auxiliary queue, $P^A(i)$. Therefore, (54) is equivalent to (8),

$$P(\kappa+i) = (1-F(\kappa-1))\, P^A(i).$$

Supplementary Result for Proof of Theorem 2

Before we prove Theorem 2, we prove the following lemma. Let $p_j(\eta)$ denote the steady-state density of the residual service time of the customer in service when there are j customers in the system, assuming that such steady-state density exists. Then,

Lemma 4.

$$\int_{u=\eta}^{\infty} e^{-\lambda_1 u} p_1'(u) du = \int_{u=\eta}^{\infty} e^{-\lambda_1 u} \left(\lambda_1 p_1(u) - \lambda_0 P(0) b_1(u) - p_2(0) b_1(u) \right) du \tag{55}$$

$$\int_{u=\eta}^{\infty} e^{-\lambda_j u} p_j'(u) du = \int_{u=\eta}^{\infty} e^{-\lambda_j u} \left(\lambda_j p_j(u) - \alpha_j p_{j-1}(\alpha_j u) \lambda_{j-1} - p_{j+1}(0) b_j(u) \right) du. \tag{56}$$

Proof of Lemma 4.

As we assume that $b_j(\cdot)$ are all absolutely continuous, a sufficient condition for $p_j(\eta)$ to exist is that the $M_n/G_n/1$ system is stable. Dividing both sides of (12) and (13) by dt and rearranging them we get,

$$\frac{p_1(\eta,t)-p_1(\eta-dt,t+dt)}{dt} = p_1(\eta,t)\lambda_1 - p_2(0,t)b_1(\eta) - \lambda_0 p_0(t)b_1(\eta) + \frac{o(dt)}{dt},$$
$$\frac{p_j(\eta,t)-p_j(\eta-dt,t+dt)}{dt} = p_j(\eta,t)\lambda_j - p_{j+1}(0,t)b_j(\eta) - \alpha_j p_{j-1}(\eta\alpha_j,t)\,\lambda_{j-1} + \frac{o(dt)}{dt}.$$

Taking the limit as $dt \to 0$, we get:

$$p_1'(\eta,t) = \lambda_1 p_1(\eta,t) - \lambda_0 p_0(t) b_1(\eta) - p_2(0,t)b_1(\eta),$$
$$p_j'(\eta,t) = \lambda_j p_j(\eta,t) - \alpha_j p_{j-1}(\alpha_j\eta,t)\,\lambda_{j-1} - p_{j+1}(0,t)b_j(\eta).$$

Assuming that the steady-state exists and taking the limit $t \to \infty$, we get:

$$p_1'(\eta) = \lambda_1 p_1(\eta) - \lambda_0 P(0) b_1(\eta) - p_2(0)b_1(\eta), \tag{57}$$
$$p_j'(\eta) = \lambda_j p_j(\eta) - \alpha_j p_{j-1}(\alpha_j\eta)\,\lambda_{j-1} - p_{j+1}(0)b_j(\eta). \tag{58}$$

Multiplying both sides of (57) and (58) by $e^{-\lambda_j u}$ and taking integral, we get (55) and (56).

Proof of Theorem 2.

We note that $p_j(\infty)=0$, $p_j(0)\geq 0$ for $j>0$, and $P(0)>0$ because we assume that the system is stable. Also, $\int_{\eta=0}^{\infty} p_j(\eta)d\eta = P(j)$ and similarly from (1) $\int_{\eta=0}^{\infty} \alpha_j p_{j-1}(\alpha_j\eta)d\eta = P(j-1)$. Then, by setting $\eta=\lambda_j=0$ in Lemma 4, from (55) and (56) we get

$$-p_1(0) = \lambda_1 P(1) - \lambda_0 P(0) - p_2(0)$$
$$-p_j(0) = \lambda_j P(j) - \lambda_{j-1} P(j-1) - p_{j+1}(0).$$

Therefore, for $j \geq 1$

$$\lambda_{j-1}P(j-1) - p_j(0) = \lambda_j P(j) - p_{j+1}(0).$$

Note that $\lambda_{j-1}P(j-1) - p_j(0)$ is independent of j and must go to zero in the limit when $j \to \infty$ if the system is stable (see e.g., Kerner, 2008). Therefore,

$$p_{j+1}(0) = \lambda_j P(j), \quad j \geq 0. \tag{59}$$

Note that (59) can be explained using level crossing as well. Considering level j (number of customers in the system), $\lambda_j P(j)$ is the rate of up-crossing this level and $p(j+1,0)$ is the rate of down-crossing (i.e., the rate at which a departure leaves j customers behind).

Recalling that $h_j(\cdot)$ denotes the steady-state density of the residual service time observed by an arrival that sees j customers in the system, we have

$$h_j(\eta) = \frac{p_j(\eta)}{\int_{r=0}^{\infty} p_j(r)dr} = \frac{p_j(\eta)}{P(j)}. \tag{60}$$

Therefore,

$$\tilde{h}_j(s) = \frac{\int_{r=0}^{\infty} e^{-sr} p_j(r)dr}{P(j)}. \tag{61}$$

Substituting (59) in (57) and (58), and multiplying both sides by $e^{-\lambda_j \eta}$, we get after some algebra,

$$e^{-\lambda_1 \eta}(p_1'(\eta) - \lambda_1 p_1(\eta)) = -b_1(\eta)e^{-\lambda_1 \eta}(\lambda_0 P(0) + \lambda_1 P(1)) \tag{62}$$

$$e^{-\lambda_j \eta}(p_j'(\eta) - \lambda_j p_j(\eta)) = -e^{-\lambda_j \eta}\left(\alpha_j p_{j-1}\left(\alpha_j \eta\right)\lambda_{j-1} + \lambda_j P(j) b_j(\eta)\right). \tag{63}$$

Considering (61), Lemma 4, and recalling $b_j(\infty) = 0$ for $j \geq 1$ we get,

$$\begin{aligned}
p_1(\eta) &= e^{\lambda_1 \eta} \int_{u=\eta}^{\infty} \left(b_1(u)e^{-\lambda_1 u}(\lambda_0 P(0) + \lambda_1 P(1))\right) du \\
p_j(\eta) &= e^{\lambda_j \eta} \int_{u=\eta}^{\infty} \left(e^{-\lambda_j u}\left(\alpha_j p_{j-1}\left(\alpha_j u\right)\lambda_{j-1} + \lambda_j P(j) b_j(u)\right)\right) du.
\end{aligned}$$

Substituting (59) in the above equations, we get for $\eta = 0$

$$\begin{aligned}
\lambda_0 P(0) &= (\lambda_0 P(0) + \lambda_1 P(1)) \int_{u=0}^{\infty} \left(b_1(u)e^{-\lambda_1 u}\right) du \\
\lambda_{j-1} P(j-1) &= \int_{u=0}^{\infty} \left(e^{-\lambda_j u}\left(\alpha_j p_{j-1}\left(u\alpha_j\right)\lambda_{j-1} + \lambda_j P(j) b_j(u)\right)\right) du.
\end{aligned}$$

Solving the integral from right hand side of the above equations we get,

$$\begin{aligned}
&\lambda_0 P(0) = (\lambda_0 P(0) + \lambda_1 P(1))\tilde{b}_1\left(\lambda_1\right) \\
&\lambda_{j-1} P(j-1) = \lambda_{j-1} P(j-1)\tilde{h}_{j-1}\left(\frac{\lambda_j}{\alpha_j}\right) + \lambda_j P(j)\tilde{b}_j\left(\lambda_j\right),
\end{aligned}$$

because from (61) we have

$$\int_{u=0}^{\infty} e^{-\lambda_j u}\left(\alpha_j p_{j-1}\left(u\alpha_j\right)\right) du = P(j-1)\tilde{h}_{j-1}\left(\frac{\lambda_j}{\alpha_j}\right). \tag{64}$$

Finally, with $\tilde{h}_0(\cdot) = \tilde{b}_1(\cdot)$ and $\alpha_1 = 1$ after some algebra and using induction we get for each $j \geq 1$

$$P(i) = \frac{\lambda_0 P(0)}{\lambda_i} \prod_{j=0}^{i-1} \frac{1 - \tilde{h}_j\left(\frac{\lambda_{j+1}}{\alpha_{j+1}}\right)}{\tilde{b}_{j+1}(\lambda_{j+1})}.$$

Standard arguments establish the rest of the theorem.

Proof of Theorem 3.

Setting $\eta = 0$ in Lemma 4, from (55) and (56) we get (similar to (64))

$$P(1)s\tilde{h}_1(s) - p_1(0) = \lambda_1 P(1)\tilde{h}_1(s) - \tilde{b}_1(s)\left(p_2(0) + \lambda_0 P(0)\right) \tag{65}$$
$$P(j)s\tilde{h}_j(s) - p_j(0) = \lambda_j P(j)\tilde{h}_j(s) - \lambda_{j-1}P(j-1)\tilde{h}_{j-1}\left(\frac{s}{\alpha_j}\right) - \tilde{b}_j(s)p_{j+1}(0). \tag{66}$$

Substituting (59) in (65) and (66), we get

$$P(1)s\tilde{h}_1(s) - \lambda_0 P(0) = \lambda_1 P(1)\tilde{h}_1(s) - \tilde{b}_1(s)\left(\lambda_1 P(1) + \lambda_0 P(0)\right) \tag{67}$$
$$P(j)s\tilde{h}_j(s) - \lambda_{j-1}P(j-1) = \lambda_j P(j)\tilde{h}_j(s) - \lambda_{j-1}P(j-1)\tilde{h}_{j-1}\left(\frac{s}{\alpha_j}\right) - \tilde{b}_j(s)\lambda_j P(j). \tag{68}$$

Now considering (67) we get:

$$\tilde{h}_1(s) = \frac{\left(\lambda_0 P(0) - \tilde{b}_1(s)\left(\lambda_1 P(1) + \lambda_0 P(0)\right)\right)}{\left(P(1)s - \lambda_1 P(1)\right)}.$$

Substituting $\lambda_1 P(1)$ from (15) and recalling that $\tilde{h}_0(\cdot) = \tilde{b}_1(\cdot)$, we get

$$\tilde{h}_1(s) = \left(\frac{\lambda_1}{s - \lambda_1}\right)\left(\frac{\tilde{b}_1(\lambda_1)}{1 - \tilde{b}_1\left(\frac{\lambda_1}{\alpha_1}\right)}\left(1 - \tilde{b}_1(s)\right) - \tilde{b}_1(s)\right).$$

Considering that $\tilde{h}_0(\cdot) = \tilde{b}_1(\cdot)$ and $\alpha_1 = 1$, we get

$$\tilde{h}_1(s) = \frac{\lambda_1}{s - \lambda_1}\left[\tilde{b}_1(\lambda_1)\frac{1 - \tilde{h}_0(\frac{s}{\alpha_1})}{1 - \tilde{h}_0(\frac{\lambda_1}{\alpha_1})} - \tilde{b}_1(s)\right].$$

Similarly, we can obtain $\tilde{h}_j(s)$ for $j > 1$ using (68) as,

$$\tilde{h}_j(s) = \frac{\lambda_j}{s - \lambda_j}\left[\tilde{b}_j(\lambda_j)\frac{1 - \tilde{h}_{j-1}(\frac{s}{\alpha_j})}{1 - \tilde{h}_{j-1}(\frac{\lambda_j}{\alpha_j})} - \tilde{b}_j(s)\right], \quad j \geq 1.$$

Proof of Observation 3. From Theorem 2 we have

$$\begin{aligned} P(i) &= \frac{\lambda_0 P(0)}{\lambda_i}\Pi_{j=0}^{i-1}\frac{1 - \tilde{h}_j\left(\frac{\lambda_{j+1}}{\alpha_{j+1}}\right)}{\tilde{b}_{j+1}\left(\lambda_{j+1}\right)} \\ &= \lambda_{i-1}\frac{1 - \tilde{h}_{i-1}\left(\frac{\lambda_i}{\alpha_i}\right)}{\lambda_i \tilde{b}_i\left(\lambda_i\right)}P(i-1) \end{aligned}$$

Multiplying both sides by λ_{i-1} and comparing the result with (3), we get $\hat{\mu}_i$ as in (18).

Proof of Observation 4. Consider the auxiliary queue. We define a continuous time MC for the auxiliary queue similar to the one expressed in (12) and (13) with states (j, η) where j is the number of jobs in the auxiliary queue, while η denotes the remaining service time. We define

$g_j(\eta, t)$ as the probability that there are j jobs in the auxiliary queue, and remaining service time is η at time t. Therefore,

$$g_1(\eta - dt, t + dt) = g_1(\eta, t)(1 - \lambda_{\kappa+1}dt) + g_2(0, t)b_{\kappa+1}(\eta)dt + g_t(0, 0)\lambda_\kappa b_0^A(\eta)dt, \quad j = 1, \tag{69}$$
$$g_j(\eta - dt, t + dt) = g_j(\eta, t)(1 - \lambda_{\kappa+j}dt) + g_{j-1}(\eta, t)\lambda_{\kappa+j-1}dt + g_{j+1}(0, t)b_{\kappa+j}(\eta)dt, \; j \geq 2, \tag{70}$$

where $b_0^A(\cdot)$ is the steady-state service time density of a job that finds 0 jobs in this queue.

Now consider the original system defined in (12) and (13). Given that there are more than $\kappa - 1$ customers in the system, (12) and (13) reduced to the same equations as given in (69) and (70) where $b_0^A(\cdot)$ is the equilibrium service time densities of customers that find κ customers in the original system (Step 2 of the definition of the auxiliary queue). Thus, the steady-state distribution of the number of jobs in the auxiliary queue, $P^A(i)$, is identical to the steady-state state distribution of the number of customers in the original queue given that there are more than κ customers in the system, $P(\kappa + i)$.

Proof of Corollary 1. The proof is similar to the proof of Theorem 2.2.2 in Kerner (2008). Van Doorn and Regterschot (1988) define the adapted LAA as the LAA conditioning on the state of the system, and show that under the adapted LAA PASTA holds. We next establish that the adapted LAA holds in our settings. Let $X(t)$ denote the state of the system, and N_s be the Poisson process that generates the future arrivals when the state of the system is $X(t) = s$. Then, for every s we have $\{N_s(t + u) - N_s(t), u \geq 0\}$ and $X(t)$ are independent and the adaptive LAA holds. Therefore, Theorem 1 in Van Doorn and Regterschot (1988) holds.

Proof of Theorem 4.

We first obtain the transition probabilities p_{0k}. Consider p_{00}, the probability that the next departing customer leaves no customer behind given that the last departing customer left no customer behind, $P(M_{n+1} = 0 | M_n = 0)$. This probability is equal to the probability of having no arrivals during the next service time, i.e., it is $\tilde{b}_1(\lambda_1)$ (e.g., Conway 1967, page 171). Therefore,

$$p_{00} = \tilde{b}_1(\lambda_1). \tag{71}$$

Next consider P_{01}. The probability that the next departing customer leaves one customer behind given that the last departing customer left no customer behind, $P(M_{n+1} = 1 | M_n = 0)$. This probability is equal to the probability of one arrival during the next service time. The only sample path that would lead to this event is that there is exactly one arrival during the sojourn time of the departing customer. Because this sojourn time is identical to the service time of this customer, the probability p_{01} is equal to the probability that *(i)* a customer arrives after the service time starts, $\left(1-\tilde{b}_1(\lambda_1)\right)$ and *(ii)* no customers arrive during the remaining service time observed by this arrival. Noting that this arrival sees one customer in the system upon the arrival, the equilibrium remaining service time observed by this arrival is $h_1(\cdot)$. Considering that the rate of the residual service time is modified by α_2 after this customer joins the queue, the probability that no customers arrive during the remaining service time observed by this arrival is $\tilde{h}_1(\frac{\lambda_2}{\alpha_2})$. Therefore,

$$p_{01} = \left(1-\tilde{b}_1(\lambda_1)\right)\tilde{h}_1\left(\frac{\lambda_2}{\alpha_2}\right). \tag{72}$$

With a similar logic, the probability that the next departing customer leaves k customers behind given that the last departing customer left no customer behind, $P\left(M_{n+1}=k|M_n=0\right)$, is equal to the probability of k arrivals during the next service time. This probability is equal to the probability that a customer arrives after the first service time starts, $\left(1-\tilde{b}_1(\lambda_1)\right)$, followed by a customer arrival during the remaining service times of all arrivals that see $i<k$ customers in the system, each with probability $\left(1-\tilde{h}_i(\frac{\lambda_{i+1}}{\alpha_{i+1}})\right)$, and no arrival during the remaining service time once there are k customers in the system, $\tilde{h}_k(\frac{\lambda_{k+1}}{\alpha_{k+1}})$. Recalling our definition $\tilde{h}_0(\cdot)=\tilde{b}_1(\cdot)$, we have,

$$p_{0k} = \tilde{h}_k\left(\frac{\lambda_{k+1}}{\alpha_{k+1}}\right)\prod_{i=0}^{k-1}\left(1-\tilde{h}_i(\frac{\lambda_{i+1}}{\alpha_{i+1}})\right). \tag{73}$$

Using the transition probabilities in (29) derived in the body of the paper and (73) we next derive the steady-state probabilities of the embedded MC. Considering (24), we need to derive the steady-state distribution of the number of customers in system using the embedded MC. Note that the steady-state probability that a departure leaves k customers behind satisfies $P_d(k)=\sum_{j=0}^{\infty}P_d(j)p_{jk}$.

First consider $P_d(0)$:

$$\begin{aligned} P_d(0) &= P_d(0)p_{00}+P_d(1)p_{10} = P_d(0)\tilde{b}_1(\lambda_1)+P_d(1)\tilde{b}_1(\lambda_1) \\ \Rightarrow P_d(1) &= P_d(0)\frac{1-\tilde{b}_1(\lambda_1)}{\tilde{b}_1(\lambda_1)}. \end{aligned} \tag{74}$$

Next consider $P_d(1)$:

$$\begin{aligned} P_d(1) &= P_d(0)p_{01}+P_d(1)p_{11}+P_d(2)p_{21} \\ &= P_d(0)\left(1-\tilde{b}_1(\lambda_1)\right)\tilde{h}_1\left(\frac{\lambda_2}{\alpha_2}\right)+P_d(1)\left(1-\tilde{b}_1(\lambda_1)\right)\tilde{h}_1\left(\frac{\lambda_2}{\alpha_2}\right)+P_d(2)\tilde{b}_2(\lambda_2). \end{aligned} \tag{75}$$

Substituting $P_d(1)$ from (74) to (75), we get

$$P_d(2) = P_d(0)\prod_{j=0}^{1}\frac{1-\tilde{h}_j\left(\frac{\lambda_{j+1}}{\alpha_{j+1}}\right)}{\tilde{b}_{j+1}(\lambda_{j+1})}.$$

The rest of the probabilities can be obtained similarly.

Proof of Corollary 2.

Substituting (18) in (3) we get

$$\lambda_i P_F(i) = \frac{\lambda_{i+1}\tilde{b}_{i+1}(\lambda_{i+1})}{1-\tilde{h}_i\left(\frac{\lambda_{i+1}}{\alpha_{i+1}}\right)} P_F(i+1).$$

Therefore,

$$P_F(i+1) = \frac{\lambda_i}{\lambda_{i+1}} \frac{1-\tilde{h}_i\left(\frac{\lambda_{i+1}}{\alpha_{i+1}}\right)}{\tilde{b}_{i+1}(\lambda_{i+1})} P_F(i),$$

which after some algebra leads to (43).

Proof of Lemma 1. (34) is equivalent to

$$P(i) = \frac{\bar{P}^a(i)\sum_{k=0}^{\infty}\lambda_k P(k)}{\lambda_i}. \tag{76}$$

Using (76) we get $P(0)$ as,

$$P(0) = \frac{\bar{P}^a(0)\sum_{k=1}^{\infty}\frac{\lambda_k}{\lambda_0}P(k)}{1-\bar{P}^a(0)}. \tag{77}$$

Using (76) and (77), we obtain $P(i)$ as a function of $P(0)$ given in (35).

Proof of Observation 6. The proof is based on Theorem 4 and similar to the proof of Observation 3.

Proof of Lemma 2.

Noting that $1/\mu_b = -\frac{d\tilde{h}_\kappa(s)}{ds}|_{s=0}$, by taking the derivative of (17) with respect to s and setting $s=0$, we get:

$$\begin{aligned}
\frac{1}{\mu_b} = -\frac{d\tilde{h}_k(s)}{ds}|_{s=0} &= \frac{\lambda_k}{(s-\lambda_k)^2}\left[\tilde{b}_k(\lambda_k)\frac{1-\tilde{h}_{k-1}(\frac{s}{\alpha_k})}{1-\tilde{h}_{k-1}(\frac{\lambda_k}{\alpha_k})} - \tilde{b}_k(s)\right]|_{s=0} \\
&- \frac{\lambda_k}{(s-\lambda_k)}\left[-\frac{\tilde{b}_k(\lambda_k)}{1-\tilde{h}_{k-1}(\frac{\lambda_k}{\alpha_k})}\frac{d\tilde{h}_{k-1}(\frac{s}{\alpha_k})}{ds} - \frac{d\tilde{b}_k(s)}{ds}\right]|_{s=0} \\
&= -\frac{1}{\lambda_k} + \frac{1}{\mu_k} + \left[\frac{\tilde{b}_k(\lambda_k)}{1-\tilde{h}_{k-1}(\frac{\lambda_k}{\alpha_k})}\frac{d\tilde{h}_{k-1}(\frac{s}{\alpha_k})}{ds}|_{s=0}\right].
\end{aligned}$$

We prove the result by induction. For $k=1$ we have $\alpha_1 = 1$ so that

$$\frac{1}{\mu_b} = -\frac{1}{\lambda_1} + \frac{1}{\mu_1} + \left[\frac{\tilde{b}_1(\lambda_1)}{1-\tilde{b}_1(\frac{\lambda_1}{\alpha_1})}\frac{d\tilde{b}_1(\frac{s}{\alpha_1})}{ds}|_{s=0}\right] = -\frac{1}{\lambda_1} + \frac{1}{\mu_1} - \left[\frac{\tilde{b}_1(\lambda_1)}{1-\tilde{b}_1(\frac{\lambda_1}{\alpha_1})}\frac{1}{\mu_1}\right],$$

which with $\tilde{h}_0(\cdot) = \tilde{b}_1(\cdot)$ is equivalent to (38) for $k=1$. Now suppose (38) holds for $\kappa = m-1$, i.e.,

$$-\frac{d\tilde{h}_{m-1}(s)}{ds}|_{s=0} = \frac{1}{\mu_{m-1}} - \frac{1}{\lambda_{m-1}} + \sum_{j=1}^{m-2}\prod_{i=j}^{m-2}\frac{1}{\alpha_{i+1}}\left(\frac{1}{\mu_j} - \frac{1}{\lambda_j}\right)\prod_{i=j}^{m-2}\frac{\tilde{b}_{i+1}(\lambda_{i+1})}{1-\tilde{h}_i(\frac{\lambda_{i+1}}{\alpha_{i+1}})}$$

$$-\frac{1}{\mu_1}\prod_{i=1}^{m-2}\frac{1}{\alpha_i}\prod_{i=0}^{m-2}\frac{\tilde{b}_{i+1}(\lambda_{i+1})}{1-\tilde{h}_i(\frac{\lambda_{i+1}}{\alpha_{i+1}})} \tag{78}$$

We next show that it holds for $k=m$.

$$\begin{aligned}\frac{1}{\mu_b} &= -\frac{1}{\lambda_m}+\frac{1}{\mu_m}+\left[\frac{\tilde{b}_m(\lambda_m)}{1-\tilde{h}_{m-1}(\frac{\lambda_m}{\alpha_m})}\frac{d\tilde{h}_{m-1}(\frac{s}{\alpha_m})}{ds}|_{s=0}\right]\\ &= -\frac{1}{\lambda_m}+\frac{1}{\mu_m}+\left[\frac{\tilde{b}_m(\lambda_m)}{1-\tilde{h}_{m-1}(\frac{\lambda_m}{\alpha_m})}\frac{1}{\alpha_m}\frac{d\tilde{h}_{m-1}(s)}{ds}|_{s=0}\right]\end{aligned} \tag{79}$$

Substituting (78) in (79), we get (38) for $\kappa=m$, which completes the proof.

Proof of Theorem 5. Based on Observation 4, we have

$$P(k)=(1-F(k-1))\,P^A(0).$$

Substituting $P(k)$ and $F(k-1)$ from (15), we get

$$\frac{\lambda_0 P(0)}{\lambda_k}\prod_{i=0}^{k-1}\frac{1-\tilde{h}_i\left(\frac{\lambda_{i+1}}{\alpha_{i+1}}\right)}{\tilde{b}_{i+1}(\lambda_{i+1})}=\left(1-\sum_{j=0}^{k-1}\frac{\lambda_0 P(0)}{\lambda_j}\prod_{i=0}^{j-1}\frac{1-\tilde{h}_i\left(\frac{\lambda_{i+1}}{\alpha_{i+1}}\right)}{\tilde{b}_{i+1}(\lambda_{i+1})}\right)(1-\rho_b)$$

resulting in (39) after some algebra.

Proof of Lemma 3. Let N^a denote the number of customers in the system seen by an arrival. Then,

$$\bar{P}^a_F(i)=P\left(N^a=i|accepted\right)P\left(accepted\right)+P\left(N^a=i|not\,accepted\right)P\left(not\,accepted\right). \tag{80}$$

By level crossing argument for states $i<K$, the frequency of transitions from state i to state $i+1$ is equal to the frequency of transitions from state $i+1$ to state i. Therefore, the probability that an arriving customer observes i people in the system given that she is accepted to the queue, $P\left(N^a=i|accepted\right)$, is identical to the probability that a departing customer leaves i people behind, $\bar{P}^d_F(i)$. Moreover, for states $i<K$, $P\left(N^a=i|not\,accepted\right)$ is zero. Therefore,

$$\begin{aligned}\bar{P}^a_F(i) &= \bar{P}^d_F(i)\left(1-\bar{P}^a_F(K)\right)+(0)\,\bar{P}^a_F(K)\\ &= \bar{P}^d_F(i)\left(1-\bar{P}^a_F(K)\right),\quad i=0,...,K-1.\end{aligned}$$