Stochastic Systems
2016, Vol. 6, No. 2, 420–458
DOI: 10.1214/15-SSY183

CLEARING ANALYSIS ON PHASES: EXACT LIMITING PROBABILITIES FOR SKIP-FREE, UNIDIRECTIONAL, QUASI-BIRTH-DEATH PROCESSES

By Sherwin Doroudi[*], Brian Fralix[§,†], and Mor Harchol-Balter[¶,‡]

University of Minnesota[*], Clemson University[†], and Carnegie Mellon University[‡]

A variety of problems in computing, service, and manufacturing systems can be modeled via infinite repeating Markov chains with an infinite number of levels and a finite number of phases. Many such chains are quasi-birth-death processes with transitions that are *skip-free in level*, in that one can only transition between consecutive levels, and *unidirectional in phase*, in that one can only transition from lower-numbered phases to higher-numbered phases. We present a procedure, which we call Clearing Analysis on Phases (CAP), for determining the limiting probabilities of such Markov chains exactly. The CAP method yields the limiting probability of each state in the repeating portion of the chain as a linear combination of *scalar bases* raised to a power corresponding to the level of the state. The weights in these linear combinations can be determined by solving a finite system of linear equations.

1. Introduction. Markov chains are frequently used to model and analyze queueing systems, especially those systems that operate at intermediate load. A particularly rich class of infinite repeating Markov chains that are used to model a variety of management problems arising in computing, service, and manufacturing systems are known as quasi-birth-death (QBD) processes. Unfortunately, despite the modeling power of QBD processes (hereafter, *QBD chains*), it is often difficult to find exact solutions for their limiting probability distributions.

In this paper, we introduce and analyze a large subclass of QBD chains, which we call *class* $\mathbb{M}$ (defined in Section 2). Chains in this class, which

Received April 2015.
[§]Partially supported by NSF-CMMI-1435261.
[¶]Funded by NSF-CMMI-1334194 and NSF-CMMI-1538204, by the Intel Science and Technology Center for Cloud Computing, and by a Google Faculty Research Award 2015/16.
MSC 2010 subject classifications: 60J22, 60J27.
Keywords and phrases: Markov chains, quasi-birth-death processes.

we call *class-*$\mathbb{M}$ *chains*, are most appropriate for modeling queueing systems where certain queueing parameters (such as arrival rates, service rates, the number of operating servers, etc.) change over time according to a restricted stochastic pattern. These changes can be a result of uncontrollable outside factors, policies implemented by the system's manager, or a combination of the two. For example, in [10, 11] a class-$\mathbb{M}$ chain is used to study power management in data centers. The class-$\mathbb{M}$ chains in these papers model data centers where servers are turned on as they are needed and put into a sleep state—and eventually turned off—when there are no more jobs waiting in the queue. As servers are turned on, the number of operating servers increases, while this number decreases as servers are put to sleep or turned off. We explore this example and two others in greater detail in Section 3.

Class-$\mathbb{M}$ Markov chains have recently appeared in the analytic study of waiting times in healthcare settings. In [8], class-$\mathbb{M}$ chains are used to study overwork effects, which arise in medical settings where human servers (e.g., healthcare providers) initially speed up when there is a lot of work to be done (e.g., when there is a long queue of patients), and then slow down if the queue has been long for an extended period of time. Meanwhile, [27] studies the "slowdown effect" impacting ICU patients—who experience delays upon arrival—by analyzing Markov chains that fall within class $\mathbb{M}$: see e.g., [5] for more on the slowdown effect phenomenon.

The primary contribution of this paper is our solution method, Clearing Analysis on Phases (CAP), which allows for determining the *exact* limiting probability distribution (i.e., the exact stationary distribution) of any class-$\mathbb{M}$ chain. While chains in *some* subsets of class $\mathbb{M}$ could be solved analytically, previously proposed methods for determining the limiting probability distribution of many class-$\mathbb{M}$ chains were restricted to *numerical* solutions. Moreover, the exact solution provided by the CAP method is in a linear combination form that is extremely convenient for the calculation of performance metrics such as the mean and variance of the queue length. Therefore, the CAP method is a novel tool that practitioners can employ to evaluate performance metrics for a variety of queueing systems. These evaluations can then be used to make informed managerial decisions with regards to admission control, staffing, and availability.

Our methodological contribution hinges on viewing class-$\mathbb{M}$ Markov chains as being composed of sequentially connected clearing models. Formally defined in Section 5.4, *clearing models* resemble ordinary M/M/1 queueing models, except that they allow for *clearing events* where the entire queue can be emptied at once, regardless of the current queue length. The CAP method

treats changes in the system parameters of a queueing system—which can occur at any queue length—as being roughly analogous to clearing events in clearing models. This approach complements existing frameworks used in the analysis of QBD chains by providing an alternative way of conceptualizing the stochastic evolution of these chains.

2. The model. We consider a class, $\mathbb{M}$, of ergodic continuous time Markov chains (ergodic CTMCs), which we call *class-*$\mathbb{M}$ *Markov chains* (or simply "class-$\mathbb{M}$ chains"), with a countably infinite state space $\mathcal{E}$ and a transition rate matrix $\mathbf{Q} \equiv [q(x,y)]_{x,y\in\mathcal{E}}$ (see Fig. 1 and Fig. 2). The infinite state space can be decomposed as $\mathcal{E} = \mathcal{R} \cup \mathcal{N}$, where $\mathcal{R}$ represents what we call the infinite *repeating portion* of the chain and $\mathcal{N}$ represents the finite *nonrepeating portion* of the chain.

The repeating portion of a class-$\mathbb{M}$ chain is given by

$$\mathcal{R} \equiv \{(m,j)\colon 0 \le m \le M, j \ge j_0\}$$

where both M and j_0 are finite nonnegative integers. We refer to a state $(m,j) \in \mathcal{R}$ as the state at phase m and level j. For each $j \ge j_0$, level j is given by

$$L_j \equiv \{(0,j),(1,j),\ldots,(M,j)\}.$$

The portion $\mathcal{R}$ is named the repeating portion because the transition rate structure within $\mathcal{R}$ is level-independent (but possibly phase-dependent). Throughout this paper, we index phases by i, k, m, and u, and we index levels by j and ℓ.

In a class-$\mathbb{M}$ Markov chain, transitions from a state in $\mathcal{N}$ to a state in $\mathcal{R}$ and vice versa are only allowed via the states in $L_{j_0} \subseteq \mathcal{R}$ (i.e., if $x \in \mathcal{N}$ and $(m,j) \in \mathcal{R}$, then the transition rate $q(x,(m,j)) = q((m,j),x) = 0$ unless $j = j_0$). Other than this restriction, *the structure of* $\mathcal{N}$ *is completely arbitrary.* In many contexts where the level, j, corresponds to the number of jobs present in a queueing system, it is natural to label the states in $\mathcal{N}$ by some subset of—if not all of the elements in—the set

$$\{(m,j)\colon 0 \le m \le M, 0 \le j \le j_0 - 1\},$$

where j_0 is the first level that the chain begins to exhibit a repeating structure. However, there exist contexts where the states within $\mathcal{N}$ (and the transitions between them) are completely arbitrary and need not be described by a phase and/or level (see the example given in Section 3.2). Throughout this paper, we index states in $\mathcal{N}$ in an arbitrary fashion, e.g., $\{x\}_{x\in\mathcal{N}}$ or $\{y\}_{y\in\mathcal{N}}$.

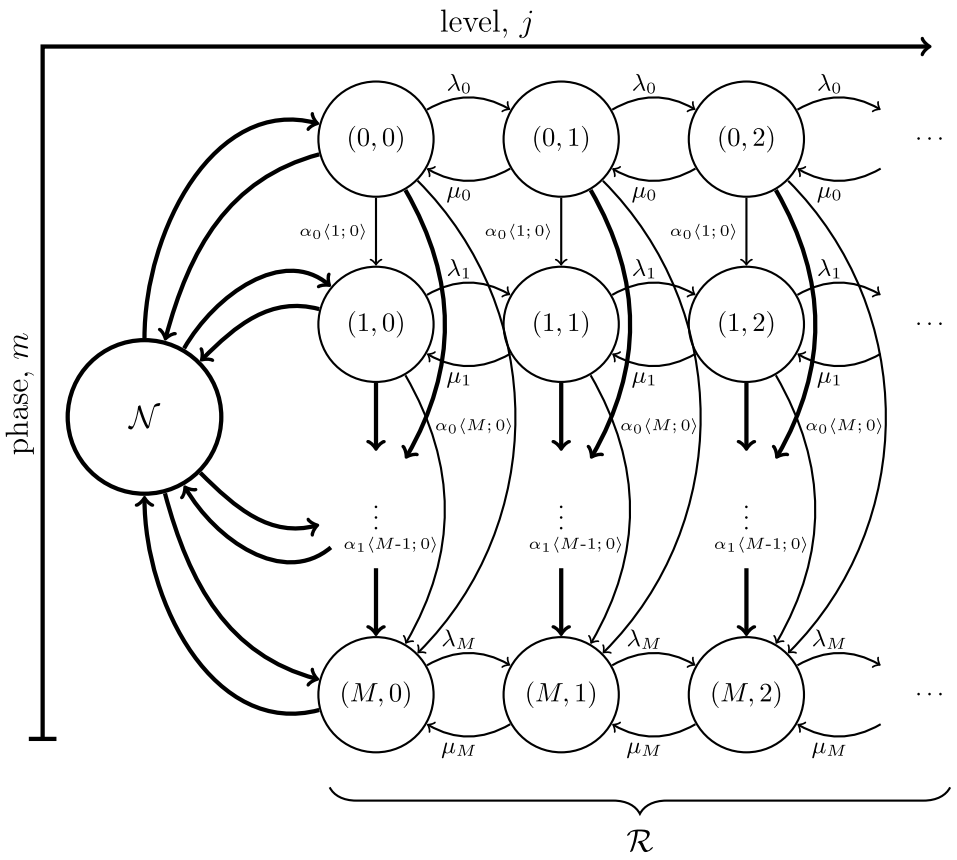

FIG 1. *The structure of class-*$\mathbb{M}$ *Markov chains. In this case* $j_0 = 0$ *and, for simplicity,* $\alpha_m\langle i - m; \pm 1\rangle = 0$. *The chain is made up of a non-repeating portion,* $\mathcal{N}$ *(shown here as an aggregation of states), and a repeating portion,* $\mathcal{R}$. *Within* $\mathcal{R}$, *each phase,* m, *corresponds to a "row" of states, and each level,* j, *corresponds to a "column" of states. Transitions between levels in each phase of the repeating portion,* $\mathcal{R}$, *are* skip-free*: all such transitions move only* one step *to the "left" or "right." Transitions between phases in each level of* $\mathcal{R}$ *are* unidirectional*: all such transitions move "downward." The thicker arrows denote* sets *of transitions (transitions rates for these sets are omitted from the figure).*

Transitions between two states in $\mathcal{R}$ are restricted to the following: (i) transitions between states in the same phase, m (e.g., the "horizontal" transitions in Fig. 1 and Fig. 2), which exist only between consecutive levels, with the rates

$$\begin{aligned} \lambda_m &\equiv q((m,j),(m,j+1)) && (0 \le m \le M, \quad j \ge j_0) \quad \text{and} \\ \mu_m &\equiv q((m,j),(m,j-1)) && (0 \le m \le M, \quad j \ge j_0 + 1), \end{aligned}$$

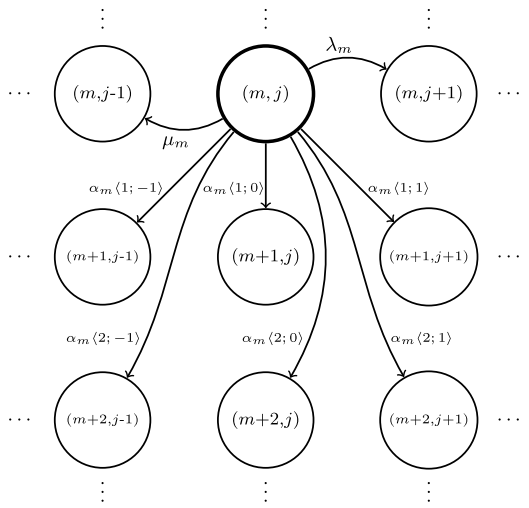

FIG 2. *Another more detailed look at the transition structure of class-*$\mathbb{M}$ *Markov chains. For simplicity, only the set of transitions that are possible* from *state* (m, j) *(where* $j \geq j_0 + 1$*) to states in phases* m*,* $m + 1$*, and* $m + 2$ *are shown. Note that all transitions from* (m, j) *are either to the left, to the right, or downward. Furthermore, all transitions can decrease or increase the level by at most one.*

and (ii) transitions across phases, from phase m to a state in another phase (e.g., the "vertical" transitions in Fig. 1 and the "vertical" and "diagonal" transitions in Fig. 2). For a transition of the latter type, say between state (m, j) and state $(m+\Delta_1, j+\Delta_2)$, the transition rate is given by $\alpha_m\langle\Delta_1; \Delta_2\rangle$, where $\Delta_1 \geq 1$ is the increase in phase from m to $m+\Delta_1$ (i.e., the "vertical" shift) and $\Delta_2 \in \{-1, 0, 1\}$ is the change in level, if any, from j to $j+\Delta_2$ (i.e., the "horizontal" shift). Note that $\Delta_1 \geq 1$ indicates that only transitions to higher-numbered phases are allowed, while $\Delta_2 \in \{-1, 0, 1\}$ indicates that each transition may change the level by at most 1 in either direction. More specifically, these transitions are described as follows:

$$\begin{aligned}
\alpha_m\langle i-m; -1\rangle &\equiv q((m,j),(i,j-1)) \qquad (0 \leq m < i \leq M, \quad j \geq j_0 + 1)\\
\alpha_m\langle i-m; 0\rangle &\equiv q((m,j),(i,j)) \qquad\quad (0 \leq m < i \leq M, \quad j \geq j_0)\\
\alpha_m\langle i-m; 1\rangle &\equiv q((m,j),(i,j+1)) \qquad (0 \leq m < i \leq M, \quad j \geq j_0).
\end{aligned}$$

We will also use the shorthand notation

$$\alpha_m = \sum_{i=m+1}^{M} \left(\alpha_m\langle i-m; -1\rangle + \alpha_m\langle i-m; 0\rangle + \alpha_m\langle i-m; 1\rangle\right)$$

throughout the paper to represent the *total outgoing transition rate to other phases* from states in phase m with level $j \geq j_0 + 1$.

Markov chains in class-$\mathbb{M}$ are examples of quasi-birth-death (QBD) processes, with increments and decrements in level corresponding to "births" and "deaths," respectively. We say that transitions in class-$\mathbb{M}$ chains are *skip-free in level*, in that the chain does not allow for the level to increase or decrease by more than 1 in a single transition. We also say that transitions in class-$\mathbb{M}$ chains are *unidirectional in phase*, in that transitions may only be made to states having either the same phase or a higher-numbered phase in the repeating portion. Note however that phases may be skipped: for example, transitions from a state in phase 2 to a state in phase 5 may exist with nonzero rate.

Many common queueing systems arising in computing, service, and manufacturing systems can be modeled with CTMCs from class $\mathbb{M}$ (a few examples are given in Section 3). For such systems, one often needs to track both the number of jobs in the system and the state of the server(s), where each server may be in one of several states, e.g., working, fatigued, on vacation, etc. When modeling a system with a class-$\mathbb{M}$ Markov chain, we often use the level, j, of a state (m, j) to track the number of jobs in the system, and we use the phase, m, to track the state of the server(s) and/or the arrival process. For example, a change in phase could correspond to (i) a policy modification that results in admitting more customers, as captured by an increase in "arrival rate" from λ_m to λ_i, where $\lambda_i > \lambda_m$ or (ii) a change in the state of the servers leading to an increase or decrease in the service rate from μ_m to μ_i.

3. Examples of class-$\mathbb{M}$ Markov chains. In this section we provide several illustrative examples of queueing systems, and we model these systems as class-$\mathbb{M}$ Markov chains. In each example we will use the phase, $m \in \{0, 1, \ldots, M\}$, to track the "state" of the server(s) and/or the arrival process, and the level, j, to track the number of jobs in the system.

3.1. *Single server in different power states.* Consider a computer server that can be in one of three different power states: on, off, or sleep. In the **on** state, the server is fully powered and jobs are processed at rate μ. In the **off** state, the server consumes no power, but jobs cannot be processed. When the server is idle, it is desirable to switch to the off state in order to conserve power, however there is a long setup time, distributed Exponential(γ), needed to turn the server back on when work arrives. Because of this setup time, it is common to switch to a state called the **sleep** state, where the server consumes less power than the **on** state, but where

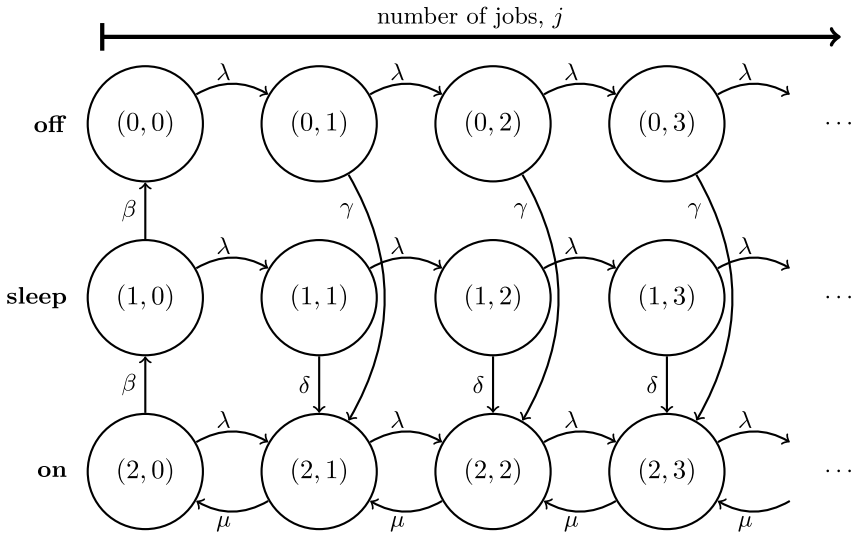

FIG 3. *The Markov chain for a single server in different power states. State* (m, j) *indicates that the server is in state* m *(0=off, 1=sleep, 2=on) with* j *jobs in the system.*

there is a shorter setup time, distributed Exponential(δ), for turning the server on. It is also common to purposefully impose a waiting period, distributed Exponential(β), in powering down a server (from on to sleep, and again from sleep to off) once it is idle, which is useful just in case new jobs arrive soon after the server becomes idle. See [12] for more details.

Fig. 3 depicts a class-$\mathbb{M}$ chain representing this setting. This is a class-$\mathbb{M}$ chain with $M+1=3$ phases: **off** ($m=0$), **sleep** ($m=1$), and **on** ($m=2$). For this chain, $j_0=1$ and the non-repeating portion of the state space is $\mathcal{N}=\{(0,0),(1,0),(2,0)\}$, while $\lambda_0=\lambda_1=\lambda_2=\lambda$, $\mu_0=\mu_1=0$, $\mu_2=\mu$, $\alpha_0\langle 2;0\rangle=\gamma$, and $\alpha_1\langle 1;0\rangle=\delta>\gamma$ (all other $\alpha_m\langle i-m;\Delta\rangle$ transition rates are zero).

The system becomes much more interesting when there are multiple servers, where each can be in one of the above 3 states. In the case of 2 servers, there will be 6 phases, corresponding to: (off,off), (off,sleep), (off,on), (sleep,sleep), (sleep,on), and (on,on). Note than in this case, phase transitions will include transitions with rates 2γ, $\gamma+\delta$, and 2δ, as both servers may be attempting to turn on at the same time. In general, a system with a servers and b server states will have $\binom{a+b-1}{a}$ phases.

3.2. *Server fatigue.* Consider a human server who starts her shift full of energy and works quickly (at rate μ_F). As time passes and fatigue sets in, she

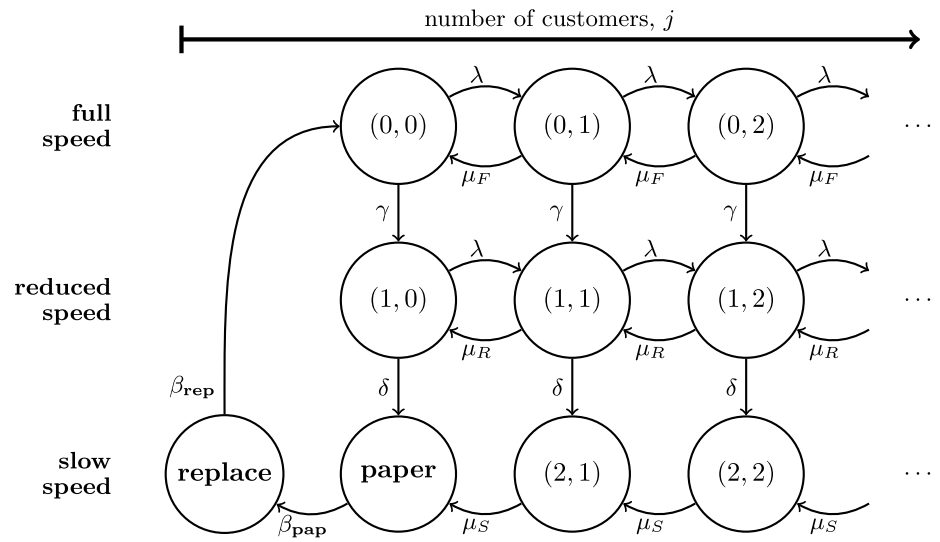

FIG 4. *The Markov chain for a server susceptible to fatigue. State (m, j) indicates server state m (0=full speed, 1=reduced speed, 2=slow speed) with j customers in the system. The states* **paper** *and* **replace** *indicate that the server is completing her paper work and that she is waiting for her replacement, respectively.*

becomes progressively slower (first she slows down to a reduced rate μ_R and eventually to a very slow rate μ_S, where $\mu_S < \mu_R < \mu_F$). At some point it makes sense to replace her with a fresh human server. However, before we can do that, she needs to finish serving her queue of existing customers, if any, while no longer accepting further arrivals and she needs to complete some paperwork. Once she has finished serving her queue (or her queue was empty at the time in which she slowed down to rate μ_S) she spends an amount of time that is distributed Exponential($\beta_{\mathbf{pap}}$) on paperwork (independent of anything that happened during her shift). Upon completing her paperwork, she calls in her replacement, who comes in and begins working after an amount of time that is distributed Exponential($\beta_{\mathbf{rep}}$).

Fig. 4 depicts a class-$\mathbb{M}$ chain representing this setting. This is a class-$\mathbb{M}$ chain with $M + 1 = 3$ phases: **full speed** ($m = 0$), **reduced speed** ($m = 1$), and **slow speed** ($m = 2$). For this chain, $j_0 = 1$ and the non-repeating portion of the state space is

$$\mathcal{N} = \{(0,0), (1,0), \mathbf{paper}, \mathbf{replace}\}$$

where the states **paper** and **replace** represent the times when the server is completing her paperwork and when she is waiting for her replacement,

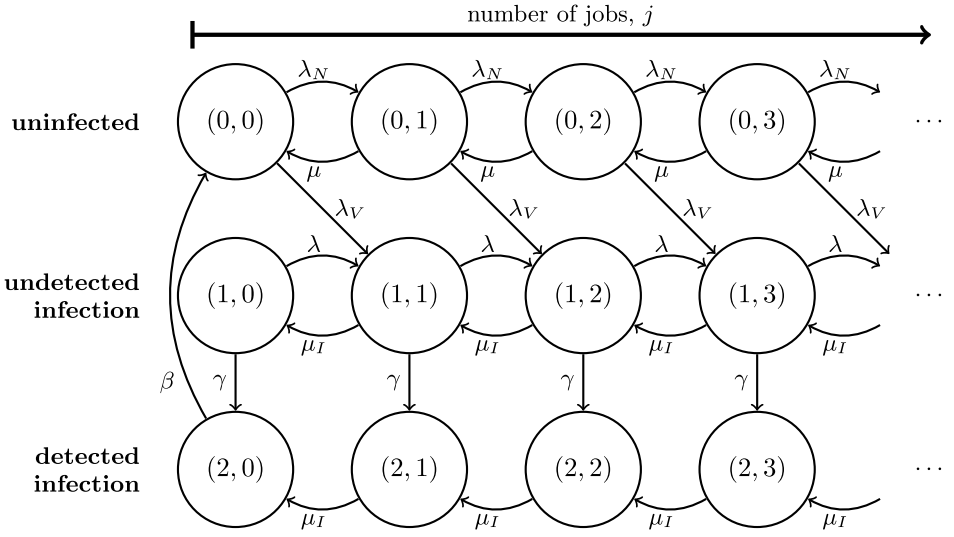

FIG 5. *The Markov chain for a server vulnerable to viruses. State (m, j) indicates server state m (0=uninfected, 1=undetected infection, 2=detected infection) with j jobs in the system.*

respectively. The rates in this chain are given by $\lambda_0 = \lambda_1 = \lambda$, $\lambda_2 = 0$, $\mu_0 = \mu_F$, $\mu_1 = \mu_R < \mu_F$, $\mu_2 = \mu_S < \mu_R$, $\alpha_0\langle 1; 0\rangle = \gamma$ and $\alpha_1\langle 1; 0\rangle = \delta$ (all other $\alpha_m\langle i - m; \Delta\rangle$ transition rates are zero).

While we could have easily called the state **paper** state $(2, 0)$, the state **replace** avoids such a natural phase-and-level classification. This is not a problem as the states in non-repeating portion of the chain need not have phases and levels.

Again, the system becomes much more interesting when there are multiple servers, where each can be in one of the 3 speed states.

3.3. *Server with virus infections.* Imagine a computer server that is vulnerable to viruses. We present a stylized model where normally, the server is **uninfected** and receives jobs with rate λ and processes them with rate μ. While most jobs are normal (i.e., not virus carriers), arriving at rate λ_N, every once in a while, one of the arriving jobs brings with it a virus, with rate $\lambda_V = \lambda - \lambda_N$. The virus causes the server to become **infected**, reducing the server's service rate from μ to μ_I. It takes a duration of time distributed Exponential(γ) for the server to detect that it is infected. Once the infection is **detected**, the server stops accepting new jobs, and once all remaining jobs are processed, the server is able to use antivirus software to remove the virus in a duration of time distributed Exponential(β). Once the

virus is removed, the server is again uninfected and will resume accepting jobs, processing them at a restored service rate of μ. We model a single server as being in one of 3 states, each of which will make up a *phase* of our Markov chain: **uninfected** ($m = 0$), **undetected infection** ($m = 1$), and **detected infection** ($m = 2$).

Fig. 5 depicts a class-$\mathbb{M}$ chain representing this setting. For this chain, $M = 2$, $j_0 = 1$, $\mathcal{N} = \{(0,0),(1,0),(2,0)\}$, $\lambda_0 = \lambda_N$, $\lambda_1 = \lambda = \lambda_N + \lambda_V$, $\lambda_2 = 0$, $\mu_0 = \mu$, $\mu_1 = \mu_2 = \mu_I$, $\alpha_0\langle 1;1\rangle = \lambda_V$, and $\alpha_1\langle 1;0\rangle = 0$ (all other $\alpha_m\langle i-m;\Delta\rangle$ transition rates are zero).

4. Literature review. In this section we review the literature on methods for solving QBD Markov chains.

4.1. *Matrix-geometric methods.* One of the most common methods used to study the stationary distribution of class-$\mathbb{M}$ Markov chains and other QBD chains is the matrix-geometric approach (and more broadly, the matrix-analytic approach). Given a QBD chain, this approach is primarily concerned with determining a matrix, $\mathbf{R} \in \mathbb{R}^{(M+1)\times(M+1)}$, that allows for the straightforward computation of the chain's limiting probability distribution. This matrix, $\mathbf{R}$, is referred to as the chain's *rate matrix* and will be discussed in greater detail in Sections 5.2 and 5.3 (for a formal definition of the rate matrix, see [22]).

For most QBD chains, one cannot derive an exact expression for each element of $\mathbf{R}$, but there are many ways to approximate $\mathbf{R}$ numerically: see e.g., [17, 4]. Arguably the most popular way of approximating $\mathbf{R}$ involves making use of an iterative scheme derived from the fact that $\mathbf{R}$ is the minimal nonnegative solution of a fixed-point equation. Another approach involves approximating $\mathbf{R}$ by instead using a similar iterative scheme to approximate a matrix $\mathbf{G}$: once $\mathbf{G}$ has been found, $\mathbf{R}$ can in principle be computed as well. Readers interested in further details on these approaches should consult the matrix-analytic texts of Neuts [22], Latouche and Ramaswami [18], and He [14]. Queueing textbooks of a broader scope that also discuss matrix-analytic methods include Asmussen [3] and Harchol-Balter [13].

There are many examples of QBD chains with a rate matrix, $\mathbf{R}$, that can be computed exactly through a finite number of operations. One class of QBD chains having closed-form rate matrices is presented in Ramaswami and Latouche [24], with an extension to Markov chains of GI/M/1-type given in Liu and Zhao [20]. Other classes of QBD chains having explicitly computable rate matrices are considered in the work of van Leeuwaarden and Winands [33] and van Leeuwaarden et al. [32], with both of these studies being much closer to our work, since most (but not all) of the types

of Markov chains studied in [33], and all of the chains discussed in [32] belong to class $\mathbb{M}$. In [33, 32] combinatorial techniques are used to derive expressions for each element of $\mathbf{R}$ that can be computed exactly after a finite number of operations. However, their methods are not directly applicable to all class-$\mathbb{M}$ Markov chains as they further assume that (i) for each $0 \leq m \leq M-1$, any transitions leaving phase m must next enter phase $m+1$ (i.e., both level *and phase* transitions must be *skip-free*) and that (ii) λ_m and μ_m are the same for $0 \leq m \leq M-1$. These assumptions preclude the analysis of many chains that arise in practice. For example, assumption (i) rules out the "single server in different power states" class-$\mathbb{M}$ chain from Section 3.1, while assumption (ii) rules out the "server fatigue" and "server with virus infections" class-$\mathbb{M}$ chains from Sections 3.2 and 3.3, respectively.

The CAP method avoids the task of finding $\mathbf{R}$, in that it provides a way of expressing all probabilities $\{\pi_x\}_{x\in\mathcal{R}}$ explicitly in terms of the probabilities $\{\pi_x\}_{x\in\mathcal{N}}$ and additional weighting terms $\{c_{m,k}\}_{m,k}$ that will be further discussed in Section 5. These remaining terms are then shown to satisfy a finite system of linear equations that can be solved either symbolically or numerically.

4.2. *The matrix-geometric method applied to tree-like QBD chains.* Even closer to our work is the work of Van Houdt and van Leeuwaarden [31], which presents an approach for the calculation of the rate matrix for a broad class of QBD chains including those in class $\mathbb{M}$. This approach involves solving higher order (scalar) polynomial equations, the solutions to which are expressed as infinite sums, which typically *cannot be computed in closed form.*

The same paper [31] gives an approach for calculating closed-form rate matrices for a class of Markov chains called tree-like QBD processes (hereafter, *tree-like chains*). Only some class-$\mathbb{M}$ Markov chains are also tree-like chains. While transitions between phases (within a level) in tree-like chains form a *directed tree*, transitions between phases in class-$\mathbb{M}$ chains form a *directed acyclic graph.* Specifically, unlike class-$\mathbb{M}$ chains, tree-like chains *do not* allow for a pair of phases $i \neq k$ to both have transitions to the same phase m. That is, tree-like chains do not allow for both $\alpha_i\langle m-i;\Delta\rangle > 0$ and $\alpha_k\langle m-k;\Delta'\rangle > 0$ when $i \neq k$ and $\Delta, \Delta' \in \{-1, 0, 1\}$).

Instances of class-$\mathbb{M}$ chains that are *not* tree-like chains—and hence *cannot* be solved in closed form by [31]—are frequently encountered in the study of multi-server systems. As an example, consider the "server fatigue" model presented in Section 3.2 extended to two servers. The phase transi-

tion structure associated with this two-server system is a non-tree directed acyclic graph. Observe that we can transition to the (reduced speed, slow speed) phase directly from one of two phases: (i) a transition from the (full speed, slow speed) phase to the (reduced speed, slow speed) occurs when the server operating at full speed experiences fatigue, and (ii) a transition from the (reduced speed, reduced speed) phase to the (reduced speed, slow speed) phase occurs when either server experiences fatigue. While the limiting probability distribution of this chain can be found in closed form via the CAP method, it *cannot* be found by the method presented in [31].

4.3. *Recursive renewal reward.* Gandhi et al. [10, 11] use renewal theory to determine exact mean values and z-transforms of various metrics for a subclass of class-$\mathbb{M}$ chains via the Recursive Renewal Reward (RRR) method. The class of chains they study do not allow for "diagonal" transitions (i.e., $\alpha_m\langle i-m;\pm 1\rangle = 0$). Unlike our method, RRR *cannot* be used to determine a formula for a chain's limiting probability distribution in finitely many operations. While there is overlapping intuition and flavor between CAP and RRR—both methods make use of renewal reward theory—CAP is *not* an extension of RRR and does not rely on the results from [10, 11].

4.4. *ETAQA.* The Efficient Technique for the Analysis of QBD-processes by Aggregation (ETAQA), first proposed by Ciardo and Simirni [7], combines ideas from matrix-analytic and state aggregation approaches in order to compute various exact values (e.g., mean queue length) for a wide class of Markov chains. By design, ETAQA yields the limiting probability of the states in the non-repeating portion, $\mathcal{N}$, along with the limiting probabilities of the states in the first level (or first few levels) of the repeating portion, $\mathcal{R}$. The limiting probabilities of the remaining states (i.e., higher level states) are aggregated, which allows for the speedy computation of exact mean values and higher moments of various metrics of interest. In particular, ETAQA involves solving a system of only $O(|\mathcal{N}|+M)$ linear equations. Although originally applicable to a narrow class of chains (see [7, 6] for details), ETAQA can be generalized so as to be applicable to M/G/1-type, GI/M/1-type, and QBD Markov chains, including those in class $\mathbb{M}$ (see the work of Riska and Smirni [25, 26]). Stathopoulos et al. [30] show that ETAQA is also well suited for numerical computations. Unlike the CAP method, ETAQA (like RRR) *cannot* be used to determine a formula for a chain's limiting probability distribution (across all states) in finitely many operations.

4.5. *Generating function techniques and the spectral expansion method.* For certain class-$\mathbb{M}$ Markov chains, one can also manipulate generating func-

tions to derive limiting probabilities, such as in the work of Levy and Yechiali [19] and the work of Phung-Duc [23], where this type of approach is used to solve multi-server vacation and setup models, respectively. This approach is covered in greater generality in a technical report by Adan and Resing [2], which is in turn extended in the work of Selen et al. [28] through the use of separation of variables. We note that although generating function approaches can yield solutions of a form similar to those found using the CAP method, the two approaches differ in methodology.

The spectral expansion method presented in the work of Mitrani and Chakka [21] also yields a solution that is similar to the form obtained by using generating function techniques or the CAP method. However, unlike the CAP method, this approach requires explicitly solving an eigenvalue problem, which could require solving higher order polynomials. Therefore, while this method is broadly applicable to QBD chains including those in class $\mathbb{M}$, the method is presented as a numerical technique, rather than one yielding explicit expressions for the limiting probability distributions of class-$\mathbb{M}$ chains.

5. Analysis and results. In this section we explain how the CAP method can be used to compute the steady-state distributions of ergodic (irreducible and positive recurrent) class-$\mathbb{M}$ Markov chains. Loosely speaking, the CAP method consists of computing the stationary distribution of a class-$\mathbb{M}$ Markov chain by applying a key result from [18] in an iterative manner that takes advantage of the unidirectional transition structure present within such chains. We then illustrate the method by deriving the steady-state distributions of two different types of class $\mathbb{M}$-chains. The calculations and techniques required to handle these two types of chains can be further combined to derive the stationary distribution of any other type of class-$\mathbb{M}$ Markov chain.

5.1. *A key idea.* Consider an ergodic CTMC having a countable state space S and transition rate matrix $\mathbf{Q} \equiv [q(x,y)]_{x,y\in S}$, where for each $x, y \in S$, $x \neq y$, $q(x,y)$ denotes the transition rate from state x to state y. These rates further define, for each state $x \in S$, the sojourn rate ν_x given by

$$\nu_x \equiv \sum_{z\in S\setminus\{x\}} q(x,z),$$

which represents the sum of all transition rates out of state x.

We are interested in developing techniques for computing the limiting distribution $\pi \equiv \{\pi_x\}_{x\in S}$ of various types of CTMCs, where for each state $x \in S$, π_x denotes the limiting probability of being in state x. Theorem 1

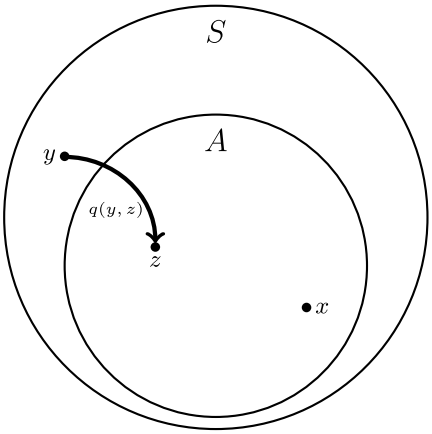

FIG 6. *For any $x \in A$, Theorem 1 gives π_x as a linear combination of quantities $\mathbb{E}_z\left[T_x^A\right]$ by conditioning on the states $y \in A^c$, by which one may transition to states $z \in A$. This figure shows one such (y, z) pair.*

(Theorem 5.5.1 of [18]) shows that given any nonempty proper subset A of S, each limiting probability π_x, $x \in A$, can be expressed in terms of the limiting probabilities $\{\pi_y\}_{y \in A^c}$, where $A^c \equiv S \backslash A$ denotes the complement of A.

THEOREM 1. *Suppose $A \subsetneq S$. Then for each state $x \in A$, π_x can be expressed as*

$$\pi_x = \sum_{y \in A^c} \sum_{z \in A} \pi_y q(y, z) \mathbb{E}_z\left[T_x^A\right],$$

where $\mathbb{E}_z\left[T_x^A\right]$ denotes the expected cumulative amount of time the chain spends in state x before leaving A, given the chain starts in state $z \in A$.

PROOF. See Theorem 5.5.1 of [18]. □

Intuitively, for each $x \in A$, Theorem 1 allows us to express π_x as a weighted average of the cumulative time spent in state x during an excursion of the CTMC in the set A, $\mathbb{E}_z\left[T_x^A\right]$, conditioned on the choice of state, $z \in A$, by which we enter A. The weights in this average represent the rate at which visits to A via z occur, which involves conditioning on the states $y \in A^c$ by which one may transition to $z \in A$. We illustrate S, A, y, z, and x in Fig. 6.

5.2. *A prior approach: The matrix-geometric method.* We provide some short background on the existing matrix-geometric method as it will be

useful to contrast CAP with this method in the following subsection (Section 5.3).

Theorem 1 is used in [18] to show that the limiting probability distribution of a QBD chain has a matrix-geometric structure on the repeating portion of its state space. The same argument can be used to establish the matrix-geometric structure satisfied by the stationary distribution of a class-$\mathbb{M}$ Markov chain on its repeating portion, $\mathcal{R}$.

To state this result, we define, for each level $\ell \geq j_0$, the vector

$$\vec{\pi}_\ell = \left(\pi_{(0,\ell)}, \pi_{(1,\ell)}, \ldots, \pi_{(M,\ell)}\right).$$

Our goal is to show, for each level $j \geq j_0$ that the vectors $\vec{\pi}_j$ and $\vec{\pi}_{j+1}$ are related via the formula

$$\vec{\pi}_{j+1} = \vec{\pi}_j \mathbf{R} \tag{5.1}$$

where $\mathbf{R} \equiv [R_{i,m}]_{0 \leq i,m \leq M}$ is referred to as the *rate matrix* associated with the chain. Each element of $\mathbf{R}$ has a nice probabilistic interpretation: $R_{i,m}$ represents $\nu_{(i,j_0)}$ times the expected amount of time the chain spends in state (m, j_0+1) before returning to the set $L_{j_0} \cup \mathcal{N}$, given the chain starts at state (i, j_0).

We are now ready to establish (5.1). Fix a level $j \geq j_0$, and define $A_j \equiv \bigcup_{\ell=j+1}^{\infty} L_\ell$. Then for each state $(m, j+1) \in L_{j+1}$, Theorem 1 yields

$$\begin{aligned}
\pi_{(m,j+1)} &= \sum_{i=0}^{M} \sum_{k=0}^{M} \pi_{(i,j)} q((i,j),(k,j+1)) \mathbb{E}_{(k,j+1)}\left[T^{A_j}_{(m,j+1)}\right] \\
&= \sum_{i=0}^{M} \pi_{(i,j)} \nu_{(i,j)} \sum_{k=0}^{M} \left(\frac{q((i,j),(k,j+1))}{\nu_{(i,j)}}\right) \mathbb{E}_{(k,j+1)}\left[T^{A_j}_{(m,j+1)}\right] \\
&= \sum_{i=0}^{M} \pi_{(i,j)} R_{i,m}.
\end{aligned}$$

For most QBD chains, one cannot derive an exact expression for each element of $\mathbf{R}$, but there are many ways to compute an approximation of $\mathbf{R}$ numerically: see for example [17]. Once $\mathbf{R}$—or a good approximation for $\mathbf{R}$—is found, each vector $\vec{\pi}_j$ for $j \geq j_0 + 1$ can be computed via $\vec{\pi}_j = \vec{\pi}_{j_0} \mathbf{R}^{j-j_0}$, and all remaining limiting probabilities, π_x, for $x \in \mathcal{N} \cup L_{j_0}$, can be found using the balance equations and the normalization constraint.

5.3. *Introducing CAP.* The CAP method consists of applying Theorem 1 in a different manner from that used in matrix-geometric methods, in order

to take advantage of the unidirectional transition structure present within class-$\mathbb{M}$ chains. For each integer $m \in \{0, 1, \ldots, M\}$, define the set P_m as

$$P_m \equiv \{(m, j_0 + 1), (m, j_0 + 2), (m, j_0 + 3), \ldots\}$$

which represents the set of states in phase m with level $j \geq j_0 + 1$ (i.e., the set of states in phase m of $\mathcal{R}$ excluding state (m, j_0)). We then repeatedly apply Theorem 1 using first the set P_0, then the set P_1, then P_2, and we stop after applying Theorem 1 with the set P_M. Applying Theorem 1 with the set P_0 yields, for each state $(0, j) \in P_0$,

$$\begin{aligned}\pi_{(0,j)} &= \sum_{y \in P_0^c} \sum_{z \in P_0} \pi_y q(y, z) \mathbb{E}_z \left[T^{P_0}_{(0,j)} \right] \\ &= \pi_{(0,j_0)} \lambda_0 \mathbb{E}_{(0,j_0+1)} \left[T^{P_0}_{(0,j)} \right]\end{aligned}$$

which shows all limiting probabilities $\pi_{(0,j)}$, $j \geq j_0 + 1$, can be expressed in terms of $\pi_{(0,j_0)}$.

Since phase transitions are unidirectional, we now proceed in an inductive manner: assume we have derived an expression for each $\pi_{(i,j)}$, $0 \leq i \leq m-1$, $j \geq j_0$. Applying Theorem 1 with the set P_m, $1 \leq m \leq M$ further yields, for each state $(m, j) \in P_m$,

$$\begin{aligned}\pi_{(m,j)} = \pi_{(m,j_0)} \lambda_m \mathbb{E}_{(m,j_0+1)} \left[T^{P_m}_{(m,j)} \right] \\ + \sum_{i=0}^{m-1} \sum_{\ell=j_0+1}^{\infty} \sum_{\Delta=-1}^{1} \pi_{(i,\ell-\Delta)} \alpha_i \langle m - i; \Delta \rangle \mathbb{E}_{(m,\ell)} \left[T^{P_m}_{(m,j)} \right].\end{aligned}$$

The unidirectionality of phase transitions in class-$\mathbb{M}$ Markov chains ensures that the limiting probabilities appearing in the right-hand side of the equation above are only those associated with phase m and lower-numbered phases.

A bit of thought shows that each limiting probability $\pi_{(m,j)}$, for $0 \leq m \leq M$, $j \geq j_0 + 1$, can be expressed in terms of the limiting probabilities $\{\pi_x\}_{x \in \mathcal{N}}$ and $\{\pi_{(m,j_0)}\}_{0 \leq m \leq M}$, but to make this observation useful we must show that each expectation $\mathbb{E}_{(m,\ell)} \left[T^{P_m}_{(m,j)} \right]$ can be computed analytically. These analytic computations are presented in Theorem 2.

THEOREM 2. *For any class-$\mathbb{M}$ Markov chain, if $\lambda_m, \mu_m > 0$ and $\ell, j \geq j_0 + 1$, we have*

$$\mathbb{E}_{(m,\ell)} \left[T^{P_m}_{(m,j)} \right] = \begin{cases} \Omega_m r_m^{j-\ell} \left(1 - (r_m \phi_m(\alpha_m))^{\ell - j_0}\right) & \text{if } \ell \leq j \\ \Omega_m \phi_m(\alpha_m)^{\ell - j} \left(1 - (r_m \phi_m(\alpha_m))^{j - j_0}\right) & \text{if } \ell \geq j, \end{cases} \tag{5.2}$$

where

$$\rho_m \equiv \lambda_m/\mu_m, \qquad r_m \equiv \rho_m \phi_m(\alpha_m), \qquad \Omega_m \equiv \frac{r_m}{\lambda_m(1 - r_m\phi_m(\alpha_m))},$$

and $\phi_m(\alpha_m)$ is the Laplace-Stieltjes transform of the length of the busy period of an M/M/1 queue having arrival rate λ_m and service rate μ_m, evaluated at α_m.

PROOF. The detailed proof of this result follows from the clearing model analysis that will be presented in Section 5.4 (see the statement and proof of Lemma 2 in particular). The key idea is to view the phase P_m as its own "smaller" Markov chain (in particular an M/M/1 clearing model), and to analyze the stochastic evolution of that Markov chain. This approach is valid because we are concerned only with the stochastic evolution of the class-$\mathbb{M}$ chain *within* P_m, and moreover, all quantities in the statement of this theorem depend only the properties of the chain at phase m (including the total outgoing transition rate, α_m). □

The $M + 1$ scalars $r_0, r_1, \ldots, r_M$ that are given by

$$r_m \equiv \rho_m \phi_m(\alpha_m) = \frac{\alpha_m + \lambda_m + \mu_m - \sqrt{(\alpha_m + \lambda_m + \mu_m)^2 - 4\lambda_m\mu_m}}{2\mu_m} \tag{5.3}$$

when $\mu_m > 0$ and by

$$r_m \equiv \frac{\lambda_m}{\lambda_m + \alpha_m} \tag{5.4}$$

otherwise, are referred to throughout as the *base terms* (or *bases*) associated with a class-$\mathbb{M}$ Markov chain. These base terms are actually the diagonal elements of the rate matrix $\mathbf{R}$. To see why this is the case, define $C_{j_0+1} = \bigcup_{\ell=j_0+1}^{\infty} L_\ell$ and observe that for $0 \le m \le M$, it follows from both the unidirectional transition structure of class-$\mathbb{M}$ Markov chains, and the probabilistic interpretation of each element of $\mathbf{R}$ that

$$R_{m,m} = \lambda_m \mathbb{E}_{(m,j_0+1)}\left[T^{C_{j_0+1}}_{(m,j_0+1)}\right] = \lambda_m \mathbb{E}_{(m,j_0+1)}\left[T^{P_m}_{(m,j_0+1)}\right] = \lambda_m\left(\frac{r_m}{\lambda_m}\right) = r_m.$$

We can also use this probabilistic interpretation of the elements of $\mathbf{R}$ to show that $\mathbf{R}$ is a lower-triangular matrix: given two phases i, m satisfying $0 \le m < i \le M$, unidirectional phase transitions guarantee that one cannot spend any time in P_m before reaching $\mathcal{N}$, given that one starts in P_i, establishing that $R_{i,m} = 0$.

What sort of expressions for the limiting probabilities $\{\pi_x\}_{x\in\mathcal{R}}$ does the CAP method yield? When the base terms $r_0, r_1, \ldots, r_M$ are all distinct, we have for $0 \le m \le M$, $j \ge j_0$ that

$$\pi_{(m,j)} = \sum_{k=0}^{m} c_{m,k} r_k^{j-j_0} \tag{5.5}$$

where $\{c_{m,k}\}_{0\le k\le m\le M}$ are a doubly-indexed set of coefficients that will be discussed in further detail later. Furthermore, when the base terms are all equal, we instead have

$$\pi_{(m,j)} = \sum_{k=0}^{m} c_{m,k} \binom{j-(j_0+1)+k}{k} r_0^{j-j_0} \tag{5.6}$$

where again, the doubly-indexed coefficients $\{c_{m,k}\}_{0\le m\le M, 0\le k\le m}$ will be defined in more detail later.

It is not surprising that $\pi_{(m,j)}$ can be expressed as a linear combination of scalars, each raised to the power of $j-j_0$, as in Equations (5.5) and (5.6). Since $\mathbf{R}$ is a lower-triangular matrix, its eigenvalues are simply its diagonal elements, which are also the diagonal elements of the Jordan normal form of $\mathbf{R}$—see e.g., Chapter 3 of Horn and Johnson [15]—from which we know that $\pi_{(m,j)}$ can be expressed as a linear combination of scalars, each raised to the power of $j-j_0$. Although in theory, our solution form could be recovered by first computing $\mathbf{R}$ and then numerically determining $\mathbf{R}$ in Jordan normal form, such a procedure is often inadvisable. The structure of the Jordan normal form of a matrix can be extremely sensitive to small changes in one or more of its elements, particularly when some of its eigenvalues have algebraic multiplicity larger than one. Fortunately, the CAP method can handle these cases as well with little additional difficulty: the CAP method reduces the problem of determining the eigenvectors of $\mathbf{R}$ to the problem of computing the $c_{m,k}$ coefficients, and later we will see that these coefficients always appear as part of the solution to a well-defined finite linear system of equations.

5.4. *Clearing analysis.* In this subsection we define and analyze M/M/1 clearing models in order to prove Theorem 2.

Like the ordinary M/M/1 queueing model, the M/M/1 *clearing* model (see Fig. 7) is a CTMC having state space $\{0, 1, 2, 3, \ldots\}$, where for each state $j \ge 0$, transitions from state j to state $j+1$ occur with rate λ, while transitions from state j to state $j-1$ occur with rate μ for each $j \ge 2$ (transitions from state 1 to state 0 will be addressed shortly). What distinguishes this chain

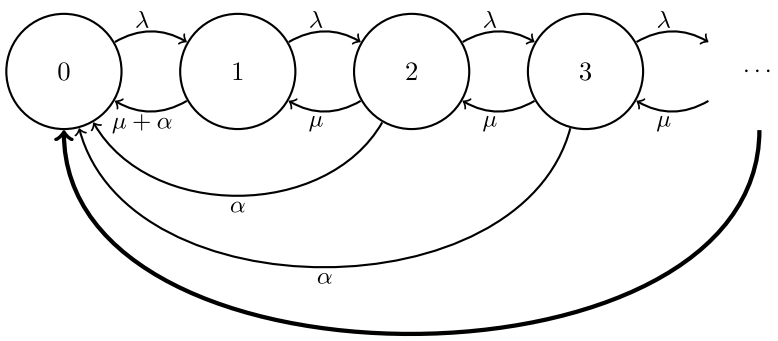

FIG 7. *Markov chain for the M/M/1 clearing model. For any state $j \geq 0$, clearings occur with rate α. Note that the transition rate from state 1 to state 0 is $\mu+\alpha$ as either a departure or a clearing can cause this transition. The thicker arrow denotes a* set *of transitions.*

from the ordinary M/M/1 chain is that for each state $j \geq 1$, it is possible to move directly from state j to state 0 due to the occurrence of a *clearing* (or catastrophe or disaster). All clearing transitions occur in accordance to a Poisson process having rate α, which we call the *clearing rate*. Note that this chain can move from state 1 to state 0 through either a departure with rate μ, or a clearing with rate α, so the transition rate from state 1 to state 0 is simply $\mu + \alpha$.

Why is an understanding of M/M/1 clearing models key to understanding the behavior of a class-$\mathbb{M}$ Markov chain? Whenever a class-$\mathbb{M}$ Markov chain is in the set P_m, $0 \leq m \leq M$, it behaves like an M/M/1 clearing model having arrival rate λ_m, service rate μ_m, and clearing rate

$$\alpha_m \equiv \sum_{i=m+1}^{M} \sum_{\Delta=-1}^{1} \alpha_m \langle i - m; \Delta \rangle,$$

except now "clearings" correspond to transitions to a state having a higher-numbered phase. This connection allows us to reformulate the problem of deriving an expression for $\mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right]$ in a class-$\mathbb{M}$ Markov chain—and thus the problem of proving Theorem 2—into a problem about clearing models. Namely, we seek to determine expressions for quantities of the form $\mathbb{E}_\ell\left[T^A_j\right]$ for M/M/1 clearing models, where $A = \{1, 2, 3, \ldots\}$ is the set of nonzero states. In order to derive such expressions, we make use of the following result adapted from [16].[1]

[1]See problems 22 and 23 from Chapter 7 of [16]: these problems actually pertain to a Brownian motion, but the same technique works when studying the difference of two homogeneous Poisson processes.

LEMMA 1. *In an M/M/1 clearing model with arrival, departure, and clearing rates λ, μ, and α, respectively, the probability that one reaches state $j > 0$ before state 0, given that one starts in state $\ell > 0$, is given by*

$$p_{\ell \to j} = \begin{cases} \dfrac{(\rho\phi(\alpha))^{j-\ell}(1-(\rho\phi(\alpha)^2)^\ell)}{1-(\rho\phi(\alpha)^2)^j} & \text{if } \ell \le j \\ \phi(\alpha)^{\ell-j} & \text{if } \ell \ge j, \end{cases}$$

where $\rho = \lambda/\mu$ and $\phi(\cdot)$ is the Laplace-Stieltjes transform of the length of the busy period of an M/M/1 system: for $\alpha > 0$,

$$\phi(\alpha) = \frac{\alpha + \lambda + \mu - \sqrt{(\alpha+\lambda+\mu)^2 - 4\lambda\mu}}{2\lambda}.$$

We now use Lemma 1 to compute $\mathbb{E}_\ell\left[T_j^A\right]$ in an M/M/1 clearing model, where A is the set of nonzero states. This result is presented in Lemma 2. Moreover, by recasting this lemma in the context of class-$\mathbb{M}$ Markov chains, one readily obtains $\mathbb{E}_{(m,\ell)}\left[T_{(m,j)}^{P_m}\right]$; that is, Theorem 2 (from Section 5.3) follows immediately from Lemma 2.

LEMMA 2. *In an M/M/1 clearing model with arrival, departure, and clearing rates λ, μ, and α, respectively, if $A = \{1, 2, 3, \ldots\}$ denotes the set of nonzero states of the state space of the underlying Markov chain, then*

$$\mathbb{E}_\ell\left[T_j^A\right] = \begin{cases} \Omega(\alpha)(\rho\phi(\alpha))^{j-\ell}\left(1-(\rho\phi(\alpha)^2)^\ell\right) & \text{if } \ell \le j \\ \Omega(\alpha)\phi(\alpha)^{\ell-j}\left(1-(\rho\phi(\alpha)^2)^j\right) & \text{if } \ell \ge j. \end{cases}$$

where

$$\Omega(\alpha) \equiv \frac{\rho\phi(\alpha)}{\lambda(1-\rho\phi(\alpha)^2)}.$$

PROOF. We first consider the case where $\ell \le j$. We claim that

$$\mathbb{E}_1\left[T_j^A\right] = (p_{1\to\ell})\mathbb{E}_\ell\left[T_j^A\right], \tag{5.7}$$

recalling that $p_{1\to\ell}$ is the probability that one reaches state ℓ before state 0 given initial state 1. Equivalently, in our setting, we may interpret $p_{1\to\ell}$ to be the probability that one reaches state ℓ before leaving A, given initial state 1, as 0 is the only state not in A. The claim in Equation (5.7) follows from conditional expectation and the fact that given we start in state 1, we either (i) reach state ℓ before leaving A, in which case the expected cumulative time

spent in state j before leaving A is $\mathbb{E}_\ell\left[T_j^A\right]$—note that no time is spent in j before reaching ℓ, as $\ell \leq j$—or (ii) we do not reach state ℓ before leaving A, in which case we also do not reach state j, and hence we spend 0 time in state j before leaving A.

From Lemma 1, we know that for $\ell \leq j$, we have

$$p_{\ell \to j} = \frac{(\rho\phi(\alpha))^{j-\ell}\left(1-(\rho\phi(\alpha)^2)^\ell\right)}{1-(\rho\phi(\alpha)^2)^j}. \tag{5.8}$$

Hence, in order to determine $\mathbb{E}_\ell\left[T_j^A\right]$ from Equation (5.7), we need only determine $\mathbb{E}_1\left[T_j^A\right]$. We compute this quantity via the renewal reward theorem. Let us earn reward in state j at rate 1, and consider a cycle from state 0 until one returns to 0 again (after leaving 0). We also use the known fact that the limiting probability of being in state j in an M/M/1 clearing model is given by $(1-\rho\phi(\alpha))(\rho\phi(\alpha))^j$ (see e.g., Corollary 4.2.2 of [1], as well as Exercise 10.7 of [13]). Hence, by the renewal reward theorem, we have

$$\frac{\mathbb{E}_1\left[T_j^A\right]}{\mathbb{E}[B_C]+1/\lambda} = (1-\rho\phi(\alpha))(\rho\phi(\alpha))^j, \tag{5.9}$$

where B_C denotes the length of the busy period of an M/M/1 clearing model. To determine $\mathbb{E}[B_C]$, observe that $B_C = \min\{B, \zeta_\alpha\}$, where B is an independent random variable distributed like the length of the busy period of an M/M/1 model *without* clearing, and $\zeta_\alpha \sim \text{Exponential}(\alpha)$ is an exponentially distributed clearing time. Taking the expectation, we have

$$\begin{aligned}\mathbb{E}[B_C] = \mathbb{E}[\min(B,\zeta_\alpha)] &= \int_0^\infty \mathbb{P}(B>t)\mathbb{P}(\zeta_\alpha > t)\,dt = \int_0^\infty \mathbb{P}(B>t)e^{-\alpha t}\,dt \\ &= \frac{1}{\alpha}\int_0^\infty \mathbb{P}(B>t)\left(\alpha e^{-\alpha t}\right)\,dt = \frac{\mathbb{P}(B>\zeta_\alpha)}{\alpha} = \frac{1-\mathbb{P}(B\leq\zeta_\alpha)}{\alpha} \\ &= \frac{1-\phi(\alpha)}{\alpha},\end{aligned}$$

where the final step follows from an alternate interpretation of the Laplace-Stieltjes transform (see Appendix A for details), noting that $\phi(\cdot)$ is the Laplace-Stieltjes transform of B.

Returning to Equation (5.9), we find that

$$\mathbb{E}_1\left[T_j^A\right] = \left(\frac{1-\phi(\alpha)}{\alpha}+\frac{1}{\lambda}\right)(1-\rho\phi(\alpha))(\rho\phi(\alpha))^j = \frac{(\rho\phi(\alpha))^j}{\lambda}, \tag{5.10}$$

where we make use of the identity

$$\left(\frac{1-\phi(\alpha)}{\alpha}+\frac{1}{\lambda}\right)(1-\rho\phi(\alpha))=\frac{1}{\lambda}$$

in our simplification. This identity can be verified algebraically by using the explicit form of $\phi(s)$. Alternatively, let $\mathbb{E}_0[T_0]$ be the expected duration of time spent in state 0 in a cycle starting from state 0, and ending with a return to state 0 from a nonzero state. Then by the renewal reward theorem,

$$\mathbb{E}_0[T_0]=\left(\mathbb{E}[B_C]+\frac{1}{\lambda}\right)(1-\rho\phi(\alpha))=\left(\frac{1-\phi(\alpha)}{\alpha}+\frac{1}{\lambda}\right)(1-\rho\phi(\alpha)).$$

We can also observe that during such a cycle, the only time spent in state 0 is during the initial residence, as a revisit to state 0 ends the cycle, so $\mathbb{E}_0[T_0]=1/\lambda$. Setting these quantities equal to one another yields the claimed identity directly.

We proceed to use Equation (5.7) in determining $\mathbb{E}_\ell\left[T_j^A\right]$ (in the case where $1\le\ell\le j$). By substituting in values from Equations (5.8) and (5.10), while recalling that

$$\Omega(\alpha)\equiv\frac{\rho\phi(\alpha)}{\lambda(1-\rho\phi(\alpha)^2)},$$

we have:

$$\begin{aligned}\mathbb{E}_\ell\left[T_j^A\right]&=\frac{\mathbb{E}_1\left[T_j^A\right]}{p_{1\to\ell}}=\left(\frac{(\rho\phi(\alpha))^j}{\lambda}\right)\left(\frac{1-(\rho\phi(\alpha)^2)^\ell}{(\rho\phi(\alpha))^{\ell-1}(1-\rho\phi(\alpha)^2)}\right)\\&=\frac{(\rho\phi(\alpha))^{j-\ell+1}\left(1-(\rho\phi(\alpha)^2)^\ell\right)}{\lambda(1-\rho\phi(\alpha)^2)}\\&=\Omega(\alpha)(\rho\phi(\alpha))^{j-\ell}\left(1-(\rho\phi(\alpha)^2)^\ell\right)\end{aligned}$$

Next, we consider the case where $\ell\ge j$ (note that the two branches in the claimed expression coincide when $\ell=j$). We again use conditional expectation, this time obtaining

$$\begin{aligned}\mathbb{E}_\ell\left[T_j^A\right]&=(p_{\ell\to j})\,\mathbb{E}_j\left[T_j^A\right]=\left(\phi(\alpha)^{\ell-j}\right)\left(\frac{\rho\phi(\alpha)(1-(\rho\phi(\alpha)^2)^j)}{\lambda(1-\rho\phi(\alpha)^2)}\right)\\&=\Omega(\alpha)\phi(\alpha)^{\ell-j}\left(1-(\rho\phi(\alpha)^2)^j\right),\end{aligned}$$

which completes the proof of the claim. Note that we have obtained $\mathbb{E}_j\left[T_j^A\right]$ by substituting $\ell=j$ into the expression for $\mathbb{E}_\ell\left[T_j^A\right]$, which we found for

$\ell \leq j$, and we have also used the fact from Lemma 1 that $p_{\ell \to j} = \phi(\alpha)^{\ell - j}$ whenever $\ell \geq j$. □

With Lemma 2—and hence Theorem 2—proven, we now turn our attention to using the expression we derived for $\mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right]$ in order to determine the exact limiting probability distributions of class-$\mathbb{M}$ chains.

5.5. *Overview of main results.* Our next goal is to derive the limiting probabilities $\{\pi_x\}_{x \in \mathcal{R}}$ associated with a class-$\mathbb{M}$ Markov chain. In Section 5.6, we consider the case where all nonzero bases are distinct. Distinct bases arise in many models where there is no structure connecting the transition rates associated with each phase: for example, the class-$\mathbb{M}$ Markov chain representing the "server in different power states" model presented in Section 3.1 has distinct bases. In Section 5.7 we consider the case where all bases are the same (i.e., $r_0 = r_1 = \cdots = r_M$), while requiring that $\lambda_m, \mu_m > 0$ for ease of exposition. We study this setting because it is the simplest case featuring repeated nonzero bases. In principle, the CAP method can be used to determine the limiting probabilities of *any* class-$\mathbb{M}$ Markov chain, but to make the paper readable we do not cover other possible cases, since (i) the expressions become more cumbersome when other possible relationships between base terms are considered, but fortunately (ii) it will be clear to readers how the CAP approach can be adjusted to handle any other type of class-$\mathbb{M}$ Markov chain.

5.6. *The case where all nonzero bases are distinct.* We are now ready to present our main result for the case where all nonzero bases are distinct. Theorem 3 expresses the stationary distribution of such class-$\mathbb{M}$ Markov chains as the solution to a finite system of linear equations.

THEOREM 3. *For any class-$\mathbb{M}$ Markov chain such that all nonzero bases $r_1, r_2, \ldots, r_M$—given in Equations (5.3) and (5.4)—are distinct (i.e., $r_m \neq r_i$ implies either $m \neq i$ or $r_m = \lambda_m = 0$), for all $j \geq j_0$, we have a limiting probability distribution of the form*

$$\pi_{(m,j)} = \sum_{k=0}^{m} c_{m,k} r_k^{j - j_0},$$

where $\{c_{m,k}\}_{0 \leq k \leq m \leq M}$ are constants with respect to j. Moreover, the $\{c_{m,k}\}_{0 \leq k \leq m \leq M}$ values, together with $\{\pi_{(m,j_0)}\}_{0 \leq m \leq M}$ and $\{\pi_x\}_{x \in \mathcal{N}}$, constitute $M(M+5)/2 + |\mathcal{N}| + 2$ "unknown variables" satisfying the following system of $M(M+5)/2 + |\mathcal{N}| + 3$ linear equations:

$$
\begin{cases}
c_{m,k} = \dfrac{r_k r_m \left(\sum\limits_{i=k}^{m-1} \sum\limits_{\Delta=-1}^{1} c_{i,k} \alpha_i \langle m-i;\Delta\rangle r_k^{-\Delta} \right)}{\lambda_m (r_k - r_m)(1-\phi_m(\alpha_m) r_k)} & (0 \le k < m \le M : r_m, r_k > 0) \\
c_{m,k} = \dfrac{\sum\limits_{i=k}^{m-1} \sum\limits_{\Delta=-1}^{1} c_{i,k} \alpha_i \langle m-i;\Delta\rangle r_k^{-\Delta}}{\mu_m (1-r_k) + \alpha_m} & (0 \le k < m \le M : r_k > r_m = 0) \\
c_{m,k} = 0 & (0 \le k < m \le M : r_k = 0) \\
c_{m,m} = \pi_{(m,j_0)} - \sum\limits_{k=0}^{m-1} c_{m,k} & (0 \le m \le M) \\
\pi_{(m,j_0)} = \dfrac{\mu_m \sum\limits_{k=0}^{m} c_{m,k} r_k + \sum\limits_{x \in \mathcal{N}} q(x,(m,j_0)) \pi_x + \sum\limits_{i=0}^{m-1} \sum\limits_{\Delta=-1}^{0} \alpha_i \langle m-i;\Delta\rangle \pi_{(i,j_0-\Delta)}}{\lambda_m + \sum\limits_{i=m+1}^{M} \sum\limits_{\Delta=0}^{1} \alpha_m \langle i-m;\Delta\rangle + \sum\limits_{x \in \mathcal{N}} q((m,j_0),x)} & (0 \le m \le M) \\
\pi_x = \dfrac{\sum\limits_{m=0}^{M} q((m,j_0),x) \pi_{(m,j_0)} + \sum\limits_{y \in \mathcal{N}} q(y,x) \pi_y}{\sum\limits_{m=0}^{M} q(x,(m,j_0)) + \sum\limits_{y \in \mathcal{N}} q(x,y)} & (x \in \mathcal{N}) \\
1 = \sum\limits_{x \in \mathcal{N}} \pi_x + \sum\limits_{m=0}^{M} \sum\limits_{k=0}^{m} \dfrac{c_{m,k}}{1-r_k},
\end{cases}
$$

where $q(x,y)$ denotes the transition rate from state x to state y.

We note before proving Theorem 3 that solving this system of equations *symbolically* will yield closed-form solutions for the limiting probabilities. Alternatively, if all parameter values are fixed and known, an exact numerical solution can be found by solving the system numerically using exact methods. Note that there is one more equation than there are unknowns, as is often the case in representations of limiting equations through balance equations. Although one equation can be omitted from the system, the normalization equation must be used in order to guarantee a unique solution.

It is also worth observing that once the values $\{\pi_x\}_{x \in \mathcal{N}}$ and $\{\pi_{(m,j_0)}\}_{0 \le m \le M}$ are known, all other $c_{m,k}$ terms can be computed recursively, without having to apply Gaussian elimination to the entire linear system given in Theorem 3.

This recursion may also simplify further for some types of class-$\mathbb{M}$ Markov chains. For example, if $\alpha_m \langle \Delta_1; \Delta_2 \rangle = 0$ for all $\Delta_1 \ge 2$, $\Delta_2 \in \{-1,0,1\}$, and $0 \le m \le M$, then when all bases are positive, for any $k < m$, we have

$$
c_{m,k} = c_{m-1,k} \frac{r_k r_m}{\lambda_m (r_k - r_m)(1-\phi_m(\alpha_m) r_k)} \sum_{\Delta=-1}^{1} \alpha_{m-1} \langle 1;\Delta\rangle r_k^{-\Delta}
$$

which further implies, for $k < m$,

$$
c_{m,k} = c_{k,k} \prod_{\ell=1}^{m-k} \frac{r_k r_{k+\ell}}{\lambda_{k+\ell} (r_k - r_{k+\ell})(1-\phi_{k+\ell}(\alpha_{k+\ell}) r_k)} \sum_{\Delta=-1}^{1} \alpha_{k+\ell-1} \langle 1;\Delta\rangle r_k^{-\Delta}
$$

meaning that only the $\{c_{k,k}\}_{0\le k\le M}$ terms need to be computed recursively.

PROOF OF THEOREM 3. For simplicity, we present the proof for the case where $\lambda_m, \mu_m > 0$ for all phases $m \in \{0, 1, 2, \ldots, M\}$. The complete proof that includes the cases where one or both of λ_m and μ_m may be 0 for some phases, m, is given in Appendix B.

We prove the theorem via strong induction on the phase, m. Specifically, for each phase m, we will show that $\pi_{(m,j)}$ takes the form $\pi_{(m,j)} = \sum_{k=0}^{m} c_{m,k} r_k^{j-j_0}$ for all $j \ge j_0 + 1$ (and also for the special case $j = j_0$), and show that $\{c_{m,k}\}_{0\le k\le m-1}$ satisfies

$$c_{m,k} = \frac{r_k r_m}{\lambda_m(r_k - r_m)(1 - \phi_m(\alpha_m)r_k)} \left(\sum_{i=k}^{m-1} \sum_{\Delta=-1}^{1} c_{i,k}\alpha_i \langle m - i; \Delta\rangle r_k^{-\Delta} \right)$$

while $c_{m,m} = \pi_0 - \sum_{k=0}^{m-1} c_{m,k}$. Finally, after completing the inductive proof, we justify that the remaining linear equations in the proposed system are ordinary balance equations together with the normalization constraint.

Base case:

We begin our strong induction by verifying that the claim holds for the base case (i.e., for $m = 0$). In this case, Equation (5.2) yields

$$\mathbb{E}_{(0,j_0+1)}\left[T^{P_0}_{(0,j)}\right] = \Omega_0 r_0^{j-j_0-1}(1 - r_0\phi_0(\alpha_0)) = \frac{r_0^{j-j_0}}{\lambda_0}.$$

We can now apply Theorem 1, yielding

$$\pi_{(0,j)} = \pi_{(0,j_0)}\lambda_0 \mathbb{E}_{(0,j_0+1)}\left[T^{P_0}_{(0,j)}\right] = \pi_{(0,j_0)}\lambda_0 \left(\frac{r_0^{j-j_0}}{\lambda_0}\right) = c_{0,0} r_0^{j-j_0},$$

where $c_{0,0} = \pi_{(0,j_0)}$. Hence, $\pi_{(0,j)}$ takes the claimed form. Moreover, $c_{0,0}$ satisfies the claimed constraint as $c_{0,0} = \pi_{(0,j_0)} - \sum_{k=0}^{m-1} c_{m,k} = \pi_{(0,j_0)} - 0 = \pi_{(0,j_0)}$, because the sum is empty when $m = 0$. Note that when $m = 0$, $\{c_{m,k}\}_{0\le k<m\le M}$ is empty, and hence, there are no constraints on these values that require verification.

Helpful computations:

Before proceeding to the inductive step, we compute two useful expressions: First, we have $\lambda_m \mathbb{E}_{(m,j_0+1)}\left[T^{P_m}_{(m,j)}\right] = r_m^{j-j_0}$, which follows from applying Equation (5.2). Next, we have

$$\sum_{\ell=j_0+1}^{\infty} r_k^{\ell-j_0} \mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] = \frac{r_k r_m (r_k^{j-j_0} - r_m^{j-j_0})}{\lambda_m(r_k - r_m)(1 - \phi_m(\alpha_m)r_k)},$$

which follows from well-known geometric sum identities. Note that this expression is well-defined because $r_k \neq r_m$ by assumption and $\phi_m(\alpha_m) r_k < 1$.

Inductive step:

Next, we proceed to the inductive step and assume the induction hypothesis holds for all phases $i \in \{0, 1, \ldots, m-1\}$. In particular, we assume that $\pi_{(i,j)} = \sum_{k=0}^{i} c_{i,k} r_k^{j-j_0}$ for all $i < m$. Applying Theorem 1, the induction hypothesis, and our computations above, we have[2]

$$
\begin{aligned}
&\pi_{(m,j)} \\
&\quad = \pi_{(m,j_0)} \lambda_m \mathbb{E}_{(m,j_0+1)} \left[T^{P_m}_{(m,j)} \right] + \sum_{i=0}^{m-1} \sum_{\ell=j_0+1}^{\infty} \sum_{\Delta=-1}^{1} \pi_{(i,\ell-\Delta)} \alpha_i \langle m-i; \Delta \rangle \mathbb{E}_{(m,\ell)} \left[T^{P_m}_{(m,j)} \right] \\
&\quad = \pi_{(m,j_0)} r_m^{j-j_0} + \sum_{i=0}^{m-1} \sum_{\ell=j_0+1}^{\infty} \sum_{\Delta=-1}^{1} \alpha_i \langle m-i; \Delta \rangle \left(\sum_{k=0}^{i} c_{i,k} r_k^{\ell-j_0-\Delta} \mathbb{E}_{(m,\ell)} \left[T^{P_m}_{(m,j)} \right] \right) \\
&\quad = \pi_{(m,j_0)} r_m^{j-j_0} + \sum_{k=0}^{m-1} \sum_{i=k}^{m-1} \left(c_{i,k} \sum_{\Delta=-1}^{1} \alpha_i \langle m-i; \Delta \rangle r_k^{-\Delta} \right) \left(\sum_{\ell=j_0+1}^{\infty} r_k^{\ell-j_0} \mathbb{E}_{(m,\ell)} \left[T^{P_m}_{(m,j)} \right] \right) \\
&\quad = \pi_{(m,j_0)} r_m^{j-j_0} + \sum_{k=0}^{m-1} \sum_{i=k}^{m-1} \left(c_{i,k} \sum_{\Delta=-1}^{1} \alpha_i \langle m-i; \Delta \rangle r_k^{-\Delta} \right) \left(\frac{r_k r_m (r_k^{j-j_0} - r_m^{j-j_0})}{\lambda_m (r_k - r_m)(1 - \phi_m(\alpha_m) r_k)} \right) \\
&\quad = \sum_{k=0}^{m} c_{m,k} r_k^{j-j_0},
\end{aligned}
$$

where we have collected terms with

$$
c_{m,k} = \frac{r_k r_m \left(\sum_{i=k}^{m-1} \sum_{\Delta=-1}^{1} c_{i,k} \alpha_i \langle m-i; \Delta \rangle r_k^{-\Delta} \right)}{\lambda_m (r_k - r_m)(1 - \phi_m(\alpha_m) r_k)} \qquad (0 \leq k < m \leq M)
$$

and $c_{m,m} = \pi_{(m,j_0)} - \sum_{k=0}^{m-1} c_{m,k}$, as claimed. This completes the inductive step and the proof by induction.

The balance equations and normalization constraint:

The equations with $\pi_{(m,j_0)}$ and π_x in their left-hand sides in our proposed system are ordinary balance equations (that have been normalized so that there are no coefficients on the left-hand side).

It remains to verify the normalization constraint:

$$
1 = \sum_{x \in \mathcal{N}} \pi_x + \sum_{m=0}^{M} \pi_{(m,j_0)} + \sum_{m=0}^{M} \sum_{j=j_0+1}^{\infty} \pi_{(m,j)}
$$

[2] Note that we have also used the fact that $\pi_{(i,j_0)}$ also satisfies the claimed form for all $i < m$, which is true as $c_{i,i} = \pi_{(i,j_0)} - \sum_{k=0}^{i-1} c_{i,k}$ (from the inductive hypothesis) implies that $\pi_{(i,j_0)} = \sum_{k=0}^{i} c_{i,k} = \sum_{k=0}^{i} c_{i,k} r_k^0$.

$$= \sum_{x \in \mathcal{N}} \pi_x + \sum_{m=0}^{M} \sum_{k=0}^{M} c_{m,k} + \sum_{m=0}^{M} \sum_{k=0}^{m-1} \sum_{j=j_0+1}^{\infty} c_{m,k} r_k^{j-j_0}$$

$$= \sum_{x \in \mathcal{N}} \pi_x + \sum_{m=0}^{M} \sum_{k=0}^{m} \frac{c_{m,k} r_k}{1 - r_k}. \qquad \square$$

5.7. *The case where all bases agree.* The CAP method can also be used in cases where some of the base terms coincide. We assume, for the sake of readability, that λ_m and μ_m are both positive for each phase m, but analogous results can still be derived when this is no longer the case.

THEOREM 4. *For a class-$\mathbb{M}$ Markov chain satisfying $\lambda_m, \mu_m > 0$ for $0 \le m \le M$, and $r_0 = r_1 = \cdots = r_M$, we have for each level $j \ge j_0$ and each phase m that*

$$\pi_{(m,j)} = \sum_{k=0}^{m} c_{m,k} \binom{j - (j_0 + 1) + k}{k} r_0^{j-j_0}$$

where the $\{c_{m,k}\}_{0 \le k \le m \le M}$ values satisfy the system of linear equations

$$\begin{cases} c_{m,0} = \pi_{(m,j_0)}, & (0 \le m \le M) \\ c_{m,k} = \Omega_m r_0 \displaystyle\sum_{u=k}^{m-1} \sum_{i=u}^{m-1} c_{i,u} \left[\sum_{\Delta=-1}^{1} \frac{\alpha_i \langle m-i; \Delta \rangle \phi_m(\alpha_m)^{\Delta+1}}{(1 - r_0 \phi_m(\alpha_m))^{u+1-k}} \right] \\ \qquad - \dfrac{1}{\lambda_m} \displaystyle\sum_{i=k}^{m-1} c_{i,k} \alpha_i \langle m-i; 1 \rangle \\ \qquad + \Omega_m \displaystyle\sum_{i=k-1}^{m-1} c_{i,k-1} \left[\sum_{\Delta=-1}^{1} \alpha_i \langle m-i; \Delta \rangle r_0^{-\Delta} \right] & (1 \le k \le m-1) \\ c_{m,m} = c_{m-1,m-1} \Omega_m \displaystyle\sum_{\Delta=-1}^{1} \alpha_{m-1} \langle 1; \Delta \rangle r_0^{-\Delta} & (1 \le m \le M), \end{cases}$$

together with the usual balance equations and normalization constraint.

Theorem 4 can be established with an induction argument, while making use of the following three identities which hold for class-$\mathbb{M}$ Markov chains where $\lambda_m, \mu_m > 0$, $r_0 = r_1 = \cdots = r_M$, $u \ge 0$, and $j \ge j_0 + 1$:

- $\displaystyle\sum_{\ell=j_0+2}^{\infty} \binom{\ell - (j_0+1) + u}{u} r_0^{\ell-j_0} \mathbb{E}_{(m,\ell-1)} \left[T_{(m,j)}^{P_m} \right]$

$$= \sum_{k=1}^{u+1} \frac{\Omega_m r_0}{(1 - r_0\phi_m(\alpha_m))^{u+1-k}} \binom{j-(j_0+1)+k}{k} r_0^{j-j_0},$$

- $$\sum_{\ell=j_0+1}^{\infty} \binom{\ell-(j_0+1)+u}{u} r_0^{\ell-j_0} \mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right]$$
$$= \Omega_m \binom{j-(j_0+1)+u+1}{u+1} r_0^{j-j_0} + \sum_{k=1}^{u} \frac{\Omega_m r_0 \phi_m(\alpha_m)}{(1 - r_0\phi_m(\alpha_m))^{u+1-k}} \binom{j-(j_0+1)+k}{k} r_0^{j-j_0},$$

- $$\sum_{\ell=j_0+1}^{\infty} \binom{\ell-(j_0+1)+u}{u} r_0^{\ell-j_0} \mathbb{E}_{(m,\ell+1)}\left[T^{P_m}_{(m,j)}\right]$$
$$= \frac{\Omega_m}{r_0} \binom{j-(j_0+1)+u+1}{u+1} r_0^{j-j_0} - \binom{j-(j_0+1)+u}{u} \frac{r_0^{j-j_0}}{\lambda_m} + \sum_{k=1}^{u} \frac{\Omega_m r_0 \phi_m(\alpha_m)^2}{(1 - r_0\phi_m(\alpha_m))^{u+1-k}} \binom{j-(j_0+1)+k}{k} r_0^{j-j_0}.$$

Each of these identities can be derived by using the negative binomial lemmas presented in Appendix C of [9].

6. Extending the scope of the CAP Method. In this section we briefly touch upon ways in which the CAP method can be extended beyond class-$\mathbb{M}$ Markov chains.

6.1. *Chains with "catastrophes".* Recall that the M/M/1 clearing model is used to model a system where there can be a catastrophe from any nonzero state causing an immediate transition to state 0. Similarly, we can consider a modification of a class-$\mathbb{M}$ Markov chain where from any state (m, j) with $j \geq j_0 + 1$, a catastrophe can occur taking one to state $x \in \mathcal{N}$ with rate $\alpha_m\langle x\rangle \equiv q((m, j), x)$.[3] That is, each phase can have several catastrophe rates, one for each state in the non-repeating portion. In this case, it will be useful to redefine α_m as follows:

$$\alpha_m \equiv \sum_{x\in\mathcal{N}} \alpha_m\langle x\rangle + \sum_{i=m+1}^{M} \sum_{\Delta=-1}^{1} \alpha_m\langle i-m;\Delta\rangle.$$

[3]Whether or not catastrophes can also occur in states (m, j_0) will not change the analysis as arbitrary transitions from states (m, j_0) to states $x \in \mathcal{N}$ are already allowed in class-$\mathbb{M}$ Markov chains.

The CAP method can easily be modified to give limiting probabilities for these types of Markov chains.

6.2. *Skipping levels when transitioning between phases.* Although the assumption that transitions from state (m, j) to state (m, ℓ) can only occur only if $\ell = j \pm 1$ is essential to the CAP method, the assumption that transitions from state (m, j) to state (i, ℓ) (where $i > m$) can only occur if $\ell = j \pm 1$ is much less important. That is, the CAP method may be extended to allow for nonzero transition rates of the form $\alpha_m\langle\Delta_1; \Delta_2\rangle$ with $d \leq \Delta_2 \leq D$ for some $d, D \in \mathbb{Z}$. However, it is advisable to treat the levels $L_{j_0}, L_{j_0+1}, \ldots, L_{j_0+\max\{|d|,|D|\}-1}$ as special cases, just as L_{j_0} was treated as a special case in the analysis presented throughout this paper.

6.3. *Chains with an infinite number of phases.* Consider a chain with the structure of a class-$\mathbb{M}$ chain, except with infinitely many *phases* (i.e., $m \in \{0, 1, 2, \ldots\}$), and a possibly infinite non-repeating portion, $\mathcal{N}$. The CAP method may be used to determine the $\{c_{m,k}\}_{0\leq k\leq m}$ values in terms of $\{\pi_x\}_{x\in\mathcal{N}}$ for the first K phases by solving a system of at most $O(K^2)$ equations. This is because the CAP method provides recurrences such that each $\{c_{m,k}\}_{0\leq k\leq m}$ value can be expressed in terms of $\{c_{i,k}\}_{0\leq k\leq i\leq m-1}$ values; that is, only information about lower-numbered phases (and the non-repeating portion) is needed to compute each $c_{m,k}$. We can first express such values for phase $m = 0$, then phase $m = 1$, and so on. Once these values—along with the easily determined corresponding base terms—have been obtained, we can use the CAP method to find the limiting probabilities for all states in the first K phases as long as we know the $\{\pi_x\}_{x\in\mathcal{N}}$ values.

Such a procedure is typically not useful, as the $\{\pi_x\}_{x\in\mathcal{N}}$ values are usually determined via the normalization constraint, which requires expressing limiting probabilities, $\pi_{(m,j)}$, in terms of $\{\pi_x\}_{x\in\mathcal{N}}$ for *all* phases, rather than for only the first K phases. However, there are settings where sufficient information about the structure of $\{\pi_x\}_{x\in\mathcal{N}}$ may be obtained via other analytic approaches, allowing for the CAP method to compute the limiting probability of the first K phases (where K can be as high as desired, subject to computational constraints). For example, a two-class priority queue can be modeled by an infinite phase variant of a class-$\mathbb{M}$ Markov chain. In that setting, queueing-theoretic analysis provides sufficient information about the structure of the limiting probabilities in the non-repeating portion (see [29]), making the CAP method an appropriate tool for that problem.

7. Conclusion. This paper presents a study of the stationary distribution of quasi-birth-death (QBD) continuous time Markov chains in class $\mathbb{M}$. class-$\mathbb{M}$ Markov chains are ergodic chains consisting of a finite nonrepeating portion and an infinite repeating portion. The repeating portion of a class-$\mathbb{M}$ chain consists of an infinite number of levels and a finite number of phases. Moreover, transitions in such chains are *skip-free in level*, in that one can only transition between consecutive levels, and *unidirectional in phase*, in that one can only transition from lower-numbered phases to higher-numbered phases. Despite these restrictions, class-$\mathbb{M}$ Markov chains are used extensively in modeling computing, service, and manufacturing systems, as they allow for keeping track of both the number of jobs in a system (via levels), and the state of the server(s) and/or the arrival process to the system (via phases).

This paper develops and introduces a novel technique, Clearing Analysis on Phases (CAP), for determining the limiting probabilities of class-$\mathbb{M}$ chains exactly. This method proceeds iteratively among the phases, by first determining the form of the limiting probabilities of the states in phase 0, then proceeding to do the same for the states in phase 1, and so on. As suggested by its name, the CAP method uses clearing model analysis to determine the structure of the limiting probabilities in each phase.

Unlike most existing techniques for solving for the limiting probability distribution of QBD chains, which rely upon the matrix-geometric approach, the CAP method avoids the task of finding the complete rate matrix, $\mathbf{R}$, entirely. Instead, the CAP method yields the limiting probabilities of each state, (m, j), in the repeating portion of the Markov chain as a linear combination of scalar *base terms* (with weights dependent on the phase, m), each raised to a power corresponding to the level, j. These *base terms* turn out to be the diagonal elements of the rate matrix, $\mathbf{R}$. The weights of these linear combinations can be determined by solving a finite system of linear equations. We also observe that the structure of the weights of these linear combinations can depend on the multiplicity structure of the base terms.

The CAP method can be applied to Markov chains beyond those in class $\mathbb{M}$, as discussed in Section 6. For example, the CAP method can be used to determine limiting probabilities in chains where one or more phases allow for immediate "catastrophe" transitions to states in the non-repeating portion. As another example, the CAP method can also be applied to Markov chains where transitions between phases can be accompanied with a change in level exceeding 1. The CAP method can also be used to study some chains with

an infinite number of phases. There is ample room for future work to extend the CAP method in a variety of directions.

The CAP method and the solution form it provides offer several impactful advantages. First, while many existing methods for determining the limiting probabilities of QBD chains exploit the relationship between successive *levels*, the CAP method exploits the relationship between successive *phases*, thereby offering complementary probabilistic intuition on the structure and steady-state behavior of class-$\mathbb{M}$ Markov chains. This method also provides an additional tool for practitioners who are studying systems that can be modeled by class-$\mathbb{M}$ chains. Depending on the application domain, the scalar solution form of the CAP method may have advantages over other solution forms for computing certain metrics of interest (e.g., mean values, higher moments, tail probabilities, etc.). While this paper does not cover using the solution of the CAP method to derive metrics of interest, as such metrics are often application specific, we hope that future work can find novel uses for the CAP method in a variety of settings.

APPENDIX A: AN ALTERNATIVE INTERPRETATION OF THE LAPLACE-STIELTJES TRANSFORM

Let X be a nonnegative random variable, with well-defined Laplace-Stieltjes transform $\psi(\cdot)$ (i.e., ψ is defined on all positive reals), cumulative distribution function, $F_X(\cdot)$, and probability density function, $f_X(\cdot)$; note that X may have nonzero probability mass at $+\infty$, in which case $\int_0^\infty f_X(t)\,dt < 1$ (where we interpret the integral as being evaluated on $\{t \in \mathbb{R}\colon 0 \le t < \infty\}$). Then for any constant $w > 0$, we have the following interpretation of ψ:

$$\begin{aligned}\psi(w) &= \int_0^\infty e^{-wt} f_X(t)\,dt \\ &= e^{-wt} F_X(t)\big|_0^\infty + \int_0^\infty F_X(t)\left(we^{-wt}\right)\,dt \\ &= \mathbb{P}\{X \le \zeta_w\},\end{aligned}$$

where $\zeta_w \sim \text{Exponential}(w)$ is a random variable independent of X.

APPENDIX B: THE COMPLETE PROOF OF THEOREM 3

Proof. We prove the theorem via strong induction on the phase, m. Specifically, for each phase m, we will show that $\pi_{(m,j)}$ takes the form $\pi_{(m,j)} = \sum_{k=0}^{m} c_{m,k} r_k^{j-j_0}$ for all $j \ge j_0 + 1$ (and also for the special case $j = j_0$), and show that $\{c_{m,k}\}_{0 \le k \le m-1}$ satisfies

$$c_{m,k} = \begin{cases} \dfrac{r_k r_m}{\lambda_m(r_k - r_m)(1-\phi_m(\alpha_m)r_k)} \left(\displaystyle\sum_{i=k}^{m-1} \sum_{\Delta=-1}^{1} c_{i,k}\alpha_i \langle m-i;\Delta\rangle r_k^{-\Delta} \right) & \text{if } r_m, r_k > 0 \\ \dfrac{\displaystyle\sum_{i=k}^{m-1} \sum_{\Delta=-1}^{1} c_{i,k}\alpha_i \langle m-i;\Delta\rangle r_k^{-\Delta}}{\mu_m(1-r_k)+\alpha_m} & \text{if } r_k > r_m = 0 \\ 0 & \text{if } r_k = 0, \end{cases}$$

while $c_{m,m} = \pi_0 - \sum_{k=0}^{m-1} c_{m,k}$. Finally, after completing the inductive proof, we justify that the remaining linear equations in the proposed system are ordinary balance equations together with the normalization constraint.

Base case:

We begin our strong induction by verifying that the claim holds for the base case (i.e., for $m = 0$). By the ergodicity requirement on class-$\mathbb{M}$ Markov chains, $\lambda_0 > 0$, leaving two sub-cases when $m = 0$: the case where $\mu_0 > 0$, and the case where $\mu_0 = 0$. In the first case, where $\mu_0 > 0$, Equation (5.2) yields

$$\mathbb{E}_{(0,j_0+1)}\left[T^{P_0}_{(0,j)}\right] = \Omega_0 r_0^{j-j_0-1}(1 - r_0\phi_0(\alpha_0)) = \frac{r_0^{j-j_0}}{\lambda_0}.$$

Now consider the other sub-case, where $\mu_0 = 0$, recalling that in this case, we have $r_0 = \lambda_0/(\lambda_0 + \alpha_0)$. We calculate $\mathbb{E}_{(0,j_0+1)}\left[T^{P_0}_{(0,j)}\right]$ for this case, by noting that transitions within states in P_0 cannot decrease the level, as follows: starting at state $(0, j_0 + 1)$, we either never visit state $(0, j)$ before leaving P_0, or we visit state $(0, j)$ *exactly once* before leaving P_0. The latter occurs with probability

$$\left(\frac{\lambda_0}{\lambda_0+\alpha_0}\right)^{j-j_0-1} = r_0^{j-j_0-1},$$

in which case, we spend an average of $1/(\lambda_0 + \alpha_0) = r_0/\lambda_0$ units of time in state $(0, j)$. Hence, we find that

$$\mathbb{E}_{(0,j_0+1)}\left[T^{P_0}_{(0,j)}\right] = r_0^{j-j_0-1}\left(\frac{r_0}{\lambda_0}\right) = \frac{r_0^j}{\lambda_0},$$

which coincides with our finding for the case where $\mu_0 > 0$.

In both cases, applying Theorem 1 yields

$$\begin{aligned} \pi_{(0,j)} &= \pi_{(0,j_0)}\lambda_0 \mathbb{E}_{(0,j_0+1)}\left[T^{P_0}_{(0,j)}\right] = \pi_{(0,j_0)}\lambda_0\left(\frac{r_0^{j-j_0}}{\lambda_0}\right) = \pi_{(0,j_0)} r_0^{j-j_0} \\ &= c_{0,0} r_0^{j-j_0}, \end{aligned}$$

where $c_{0,0} = \pi_{(0,j_0)}$. Hence, $\pi_{(0,j)}$ takes the claimed form. Moreover, $c_{0,0}$ satisfies the claimed constraint as $c_{0,0} = \pi_{(0,j_0)} - \sum_{k=0}^{m-1} c_{m,k} = \pi_{(0,j_0)} - 0 = \pi_{(0,j_0)}$, because the sum is empty when $m = 0$. Note that when $m = 0$, $\{c_{m,k}\}_{0 \le k < m \le M}$ is empty, and hence, there are no constraints on these values that require verification.

Inductive step:

Next, we proceed to the inductive step and assume the induction hypothesis holds for all phases $i \in \{0, 1, \ldots, m-1\}$. In particular, we assume that $\pi_{(i,j)} = \sum_{k=0}^{i} c_{i,k} r_k^{j-j_0}$ for all $i < m$. For convenience, we introduce the notation

$$\Upsilon_{m,j} \equiv \lambda_m \mathbb{E}_{(m,j_0+1)}\left[T^{P_m}_{(m,j)}\right] \quad \text{and} \quad \Psi_{m,k,j} \equiv \sum_{\ell=j_0+1}^{\infty} r_k^{\ell-j_0} \mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right].$$

Using this notation, we apply Theorem 1 and the induction hypothesis, which yields[4]

$$\begin{aligned}
\pi_{(m,j)} &= \pi_{(m,j_0)} \lambda_m \mathbb{E}_{(m,j_0+1)}\left[T^{P_m}_{(m,j)}\right] + \sum_{i=0}^{m-1} \sum_{\ell=j_0+1}^{\infty} \sum_{\Delta=-1}^{1} \pi_{(i,\ell-\Delta)} \alpha_i \langle m-i;\Delta\rangle \mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] \\
&= \pi_{(m,j_0)} \Upsilon_{m,j} + \sum_{i=0}^{m-1} \sum_{\ell=j_0+1}^{\infty} \sum_{\Delta=-1}^{1} \alpha_i \langle m-i;\Delta\rangle \left(\sum_{k=0}^{i} c_{i,k} r_k^{\ell-j_0-\Delta} \mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] \right) \\
&= \pi_{(m,j_0)} \Upsilon_{m,j} + \sum_{k=0}^{m-1} \sum_{i=k}^{m-1} \left(c_{i,k} \sum_{\Delta=-1}^{1} \alpha_i \langle m-i;\Delta\rangle r_k^{-\Delta} \right) \left(\sum_{\ell=j_0+1}^{\infty} r_k^{\ell-j_0} \mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] \right) \\
&= \pi_{(m,j_0)} \Upsilon_{m,j} + \sum_{k=0}^{m-1} \sum_{i=k}^{m-1} \left(c_{i,k} \sum_{\Delta=-1}^{1} \alpha_i \langle m-i;\Delta\rangle r_k^{-\Delta} \right) \Psi_{m,k,j}.
\end{aligned} \tag{B.1}$$

We proceed to compute $\Upsilon_{m,j}$ and $\Psi_{m,k,j}$ separately in the following cases:

- **Case 1**: $\lambda_m, \mu_m > 0$
- **Case 2**: $\lambda_m > \mu_m = 0$

[4] Note that $\sum_{\Delta=-1}^{1} \alpha_i \langle m-i;\Delta\rangle r_k^{-\Delta}$ is not well-defined when $r_k = 0$, as 0^{-1} and 0^0 are not well-defined. However, this is just a convenient formal manipulation which will remain true if we assign any real value to $\sum_{\Delta=-1}^{1} \alpha_i \langle m-i;\Delta\rangle r_k^{-\Delta}$ as $\Psi_{m,k,j} = 0$ in the $r_k = 0$ case, and the "contribution" to the sum by an index k such that $r_k = 0$ is also 0. One can verify that this is "harmless" by examining such k indices in isolation. Note further that we have additionally used the fact that $\pi_{(i,j_0)}$ also satisfies the claimed form for all $i < m$. This fact is true because $c_{i,i} = \pi_{(i,j_0)} - \sum_{k=0}^{i-1} c_{i,k}$ (from the inductive hypothesis) implies that $\pi_{(i,j_0)} = \sum_{k=0}^{i} c_{i,k} = \sum_{k=0}^{i} c_{i,k} r_k^0$, except that values of $r_k = 0$ yield undefined quantities of the form 0^0. Once again, this is a convenient formal manipulation that will not affect our results if we simply assign $0^0 = 1$ in this context.

- **Case 3**: $\mu_m > \lambda_m = 0$
- **Case 4**: $\mu_m = \lambda_m = 0$

Computations for Case 1 ($\lambda_m, \mu_m > 0$):

When $\lambda_m, \mu_m > 0$, Equation (5.2) yields $\Upsilon_{m,j} = \lambda_m \mathbb{E}_{(m,j_0+1)}\left[T^{P_m}_{(m,j)}\right] = r_m^{j-j_0}$. We also find that

$$\begin{aligned}
\Psi_{m,k,j} &= \sum_{\ell=j_0+1}^{\infty} r_k^{\ell-j_0} \mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] \\
&= \sum_{\ell=j_0+1}^{j} r_k^{\ell-j_0} \mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] + \sum_{\ell=j+1}^{\infty} r_k^{\ell-j_0} \mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] \\
&= \Omega_m \left(\sum_{\ell=j_0+1}^{j} r_k^{\ell-j_0} r_m^{j-\ell} \left(1 - (r_m \phi_m(\alpha_m))^{\ell-j_0}\right) \right. \\
&\qquad \left. + \sum_{\ell=j+1}^{\infty} r_k^{\ell-j_0} \phi_m(\alpha_m)^{\ell-j} \left(1 - (r_m \phi_m(\alpha_m))^{j-j_0}\right) \right) \\
&= \frac{r_k r_m (r_k^{j-j_0} - r_m^{j-j_0})}{\lambda_m (r_k - r_m)(1 - \phi_m(\alpha_m) r_k)},
\end{aligned}$$

where the last equality follows from well known geometric sum identities. Note that this expression is well-defined because $r_k \neq r_m$ by assumption and $\phi_m(\alpha_m) r_k < 1$.

Computations for Case 2 ($\lambda_m > \mu_m = 0$):

When $\lambda_m > \mu_m = 0$, we recall that $r_m = \lambda_m/(\lambda_m + \alpha_m)$ and compute $\mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right]$ as follows: starting at state (m,ℓ), we either never visit state (m,j) before leaving P_m, or we visit state (m,j) *exactly once* before leaving P_m. If $\ell > j$, we never visit state (m,j) before leaving P_m (and so $\mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] = 0$), but if $\ell \leq j$, we visit state (m,j) *exactly once* before leaving P_m with probability $r_m^{j-\ell}$, and this visit will last an average time of $1/(\lambda_m + \alpha_m) = r_m/\lambda_m$, yielding

$$\mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] = r_m^{j-\ell}\left(\frac{r_m}{\lambda_m}\right) = \frac{r_m^{j-\ell+1}}{\lambda_m}.$$

In particular, $\Upsilon_{m,j} = \lambda_m \mathbb{E}_{(m,j_0+1)}\left[T^{P_m}_{(m,j)}\right] = r_m^{j-j_0}$, coinciding with the expression for $\Upsilon_{m,j}$ from Case 1, and furthermore, we have

$$\begin{aligned}
\Psi_{m,k,j} &= \sum_{\ell=j_0+1}^{\infty} r_k^{\ell-j_0} \mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] \\
&= \sum_{\ell=j_0+1}^{j} r_k^{\ell-j_0} \mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] + \sum_{\ell=j+1}^{\infty} r_k^{\ell-j_0} \mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] \\
&= \sum_{\ell=j_0+1}^{j} \frac{r_k^{\ell-j_0} r_m^{j-\ell+1}}{\lambda_m} \\
&= \frac{r_k r_m (r_k^{j-j_0} - r_m^{j-j_0})}{\lambda_m (r_k - r_m)} = \frac{r_k r_m (r_k^{j-j_0} - r_m^{j-j_0})}{\lambda_m (r_k - r_m)(1 - \phi_m(\alpha_m) r_k)}.
\end{aligned}$$

which coincides with the expression for $\Psi_{m,k,j}$ that we found in Case 1. The last equality follows by noting that in this case we have $\phi_m(s) = 0$ for all s, and hence $1 - \phi_m(\alpha_m) r_k = 1$.

Computations for Case 3 ($\mu_m > \lambda_m = 0$):

When $\mu_m > \lambda_m = 0$, we have $\Upsilon_{m,j} = \lambda_m \mathbb{E}_{(m,j_0+1)}\left[T^{P_m}_{(m,j)}\right] = 0$. Next, we compute $\mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right]$ as follows: starting at state (m,ℓ), if $\ell < j$, we never visit j before leaving P_m, while if $\ell \geq j$ we will visit j exactly once with probability $\mu_m^{\ell-j}/(\mu_m+\alpha_m)^{\ell-j}$ and this visit will last an average duration of $1/(\mu_m + \alpha_m)$ units of time. Consequently, $\mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] = 0$ in the former case and

$$\mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] = \frac{\mu_m^{\ell-j}}{(\mu_m + \alpha_m)^{\ell-j+1}}$$

in the latter case. Finally, we have

$$\begin{aligned}
\Psi_{m,k,j} &= \sum_{\ell=j_0+1}^{\infty} r_k^{\ell-j_0} \mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] \\
&= \sum_{\ell=j_0+1}^{j-1} r_k^{\ell-j_0} \mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] + \sum_{\ell=j}^{\infty} r_k^{\ell-j_0} \mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] \\
&= \sum_{\ell=j}^{\infty} \frac{r_k^{\ell-j_0} \mu_m^{\ell-j}}{(\mu_m + \alpha_m)^{\ell-j+1}} = \frac{r_k^{j-j_0}}{\mu_m(1-r_k) + \alpha_m}.
\end{aligned}$$

Computations for Case 4 ($\mu_m = \lambda_m = 0$):

When $\mu_m = \lambda_m = 0$, we again have $\Upsilon_{m,j} = \lambda_m \mathbb{E}_{(m,j_0+1)}\left[T^{P_m}_{(m,j)}\right] = 0$, as in Case 3. Next, we compute $\mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right]$ as follows: in this case any visit to

P_m will consist entirely of one visit to the initial state in P_m, as there are no transitions to other states in the same phase. Hence, $\mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] = 1/\alpha_m$ if $\ell = j$, and $\mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] = 0$ otherwise. Consequently,

$$\begin{aligned}\Psi_{m,k,j} &= \sum_{\ell=j_0+1}^{\infty} r_k^{\ell-j_0}\mathbb{E}_{(m,\ell)}\left[T^{P_m}_{(m,j)}\right] = r_k^{j-j_0}\mathbb{E}_{(m,j)}\left[T^{P_m}_{(m,j)}\right] = \frac{r_k^{j-j_0}}{\alpha_m} \\ &= \frac{r_k^{j-j_0}}{\mu_m(1-r_k)+\alpha_m},\end{aligned}$$

which coincides with the expression for $\Psi_{m,k,j}$ that we found in Case 3. The last equality follows by noting that $\mu_m = 0$, and hence $\mu_m(1-r_k) = 0$.

Completing the inductive step:

We now proceed to substitute the results of our computations into Equation (B.1). As $\Upsilon_{m,j}$ can be given by the same expression for both Case 1 and 2, and the same holds for $\Psi_{m,k,j}$, we consider these two cases together, and note that they jointly make up the case where $r_m > 0$. For $j \geq j_0 + 1$,

$$\begin{aligned}&\pi_{(m,j)} \\ &= \pi_{(m,j_0)}\Upsilon_{m,j} + \sum_{k=0}^{m-1}\sum_{i=k}^{m-1}\left(c_{i,k}\sum_{\Delta=-1}^{1}\alpha_i\langle m-i;\Delta\rangle r_k^{-\Delta}\right)\Psi_{m,k,j} \\ &= \pi_{(m,j_0)}r_m^{j-j_0} + \sum_{k=0}^{m-1}\sum_{i=k}^{m-1}\left(c_{i,k}\sum_{\Delta=-1}^{1}\alpha_i\langle m-i;\Delta\rangle r_k^{-\Delta}\right)\left(\frac{r_k r_m(r_k^{j-j_0}-r_m^{j-j_0})}{\lambda_m(r_k-r_m)(1-\phi_m(\alpha_m)r_k)}\right) \\ &= \sum_{k=0}^{m} c_{m,k}r_k^{j-j_0},\end{aligned}$$

where we have collected terms with

$$c_{m,k} = \frac{r_k r_m\left(\sum_{i=k}^{m-1}\sum_{\Delta=-1}^{1}c_{i,k}\alpha_i\langle m-i;\Delta\rangle r_k^{-\Delta}\right)}{\lambda_m(r_k-r_m)(1-\phi_m(\alpha_m)r_k)} \quad (0 \leq k < m \leq M : r_m, r_k > 0)$$

and $c_{m,k} = 0$ when $r_m > r_k = 0$ and $c_{m,m} = \pi_{(m,j_0)} - \sum_{k=0}^{m-1} c_{m,k}$.

The expressions for $\Upsilon_{m,j}$ and $\Psi_{m,k,j}$ also coincide across Cases 3 and 4 (although they are distinct from their Case 1 and 2 counterparts), so we similarly consider these two cases together, noting that they jointly make up the case where $\lambda_m = r_m = 0$:

$$\pi_{(m,j)} = \pi_{(m,j_0)}\Upsilon_{m,j} + \sum_{k=0}^{m-1}\sum_{i=k}^{m-1}\left(c_{i,k}\sum_{\Delta=-1}^{1}\alpha_i\langle m-i;\Delta\rangle r_k^{-\Delta}\right)\Psi_{m,k,j}$$

$$= 0 + \sum_{k=0}^{m-1} \sum_{i=k}^{m-1} \left(c_{i,k} \sum_{\Delta=-1}^{1} \alpha_i \langle m-i; \Delta \rangle r_k^{-\Delta} \right) \left(\frac{r_k^{j-j_0}}{\mu_m(1-r_k)+\alpha_m} \right)$$
$$= \sum_{k=0}^{m} c_{m,k} r_k^{j-j_0},$$

where we have collected terms with

$$c_{m,k} = \frac{\sum_{i=k}^{m-1} \sum_{\Delta=-1}^{1} c_{i,k} \alpha_i \langle m-i; \Delta \rangle r_k^{-\Delta}}{\mu_m(1-r_k)+\alpha_m} \qquad (0 \le k < m \le M : r_m, r_k > 0)$$

and $c_{m,k} = 0$ when $r_m = r_k = 0$. Observe that since $r_m = 0$, it appears that we can allow $c_{m,m}$ to take any real value, so in order to satisfy the induction hypothesis, we set $c_{m,m} = \pi_{(m,j_0)} - \sum_{k=0}^{m-1} c_{m,k}$ in the $r_m = 0$ case as well. Also note that we have set $c_{m,k} = 0$ when $r_k = 0$ in both the $r_m > 0$ and $r_m = 0$ cases. This completes the inductive step and the proof by induction.

The balance equations and normalization constraint:

The equations with $\pi_{(m,j_0)}$ and π_x in their left-hand sides in our proposed system are ordinary balance equations (that have been normalized so that there are no coefficients on the left-hand side).

It remains to verify the normalization constraint:

$$1 = \sum_{x \in \mathcal{N}} \pi_x + \sum_{m=0}^{M} \pi_{(m,j_0)} + \sum_{m=0}^{M} \sum_{j=j_0+1}^{\infty} \pi_{(m,j)}$$
$$= \sum_{x \in \mathcal{N}} \pi_x + \sum_{m=0}^{M} \sum_{k=0}^{M} c_{m,k} + \sum_{m=0}^{M} \sum_{k=0}^{m-1} \sum_{j=j_0+1}^{\infty} c_{m,k} r_k^{j-j_0}$$
$$= \sum_{x \in \mathcal{N}} \pi_x + \sum_{m=0}^{M} \sum_{k=0}^{m} \frac{c_{m,k} r_k}{1-r_k}.$$

□

REFERENCES

[1] J. Abate and W. Whitt. Transient behavior of the M/M/1 queue via Laplace transforms. *Advances in Applied Probability*, pages 145–178, 1988. MR0932538

[2] I. Adan and J. Resing. A class of Markov processes on a semi-infinite strip. Technical Report 99-03, Eindhoven University of Technology, Department of Mathematics and Computing Sciences, 1999.

[3] S. Asmussen. *Applied Probability and Queues*, volume 51. Springer Science & Business Media, 2003. MR1978607

[4] L. Bright and P. G. Taylor. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Stochastic Models*, 11(3):497–525, 1995. MR1340970

[5] C. W. Chan, V. F. Farias, and G. J. Escobar. The impact of delays on service times in the intensive care unit. *Management Science*, 2016.
[6] G. Ciardo, W. Mao, A. Riska, and E. Smirni. ETAQA-MG1: an efficient technique for the analysis of a class of M/G/1-type processes by aggregation. *Performance Evaluation*, 57(3):235–260, 2004.
[7] G. Ciardo and E. Smirni. ETAQA: an efficient technique for the analysis of QBD-processes by aggregation. *Performance Evaluation*, 36:71–93, 1999.
[8] M. Delasay, A. Ingolfsson, and B. Kolfal. Modeling load and overwork effects in queueing systems with adaptive service rates. *Operations Research*, 2016. MR3532859
[9] S. Doroudi, B. Fralix, and M. Harchol-Balter. Clearing analysis on phases: Exact limiting probabilities for skip-free, unidirectional, quasi-birth-death processes. arXiv preprint arXiv:1503.05899v3, 2015.
[10] A. Gandhi, S. Doroudi, M. Harchol-Balter, and A. Scheller-Wolf. Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward. In *Proceedings of the ACM SIGMETRICS/international conference on Measurement and modeling of computer systems*, pages 153–166. ACM, 2013.
[11] A. Gandhi, S. Doroudi, M. Harchol-Balter, and A. Scheller-Wolf. Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward. *Queueing Systems*, 77(2):177–209, 2014. MR3206189
[12] A. Gandhi, M. Harchol-Balter, R. Raghunathan, and M. A. Kozuch. Autoscale: Dynamic, robust capacity management for multi-tier data centers. *ACM Transactions on Computer Systems (TOCS)*, 30(4):14, 2012.
[13] M. Harchol-Balter. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press, 2013.
[14] Q.-M. He. *Fundamentals of Matrix-Analytic Methods*. Springer, 2014. MR3112230
[15] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2012.
[16] S. Karlin and H. Taylor. *A First Course in Stochastic Processes*. Acadmic Press, New York, 1975.
[17] G. Latouche and V. Ramaswami. A logarithmic reduction algorithm for quasi-birth-death processes. *Journal of Applied Probability*, pages 650–674, 1993.
[18] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM, Philadelphia, 1999.
[19] Y. Levy and U. Yechiali. An M/M/s queue with servers' vacations. *INFOR*, 14:153–163, 1976.
[20] D. Liu and Y. Zhao. Determination of explicit solution for a general class of Markov processes. *Matrix-Analytic Methods in Stochastic Models*, page 343, 1996. MR1427280
[21] I. Mitrani and R. Chakka. Spectral expansion solution for a class of Markov models: Application and comparison with the matrix-geometric method. *Performance Evaluation*, 23(3):241–260, 1995.
[22] M. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. John Hopkins University Press, Baltimore, Maryland, 1981.
[23] T. Phung-Duc. Exact solutions for M/M/c/setup queues. *Telecommunication Systems*, pages 1–16, 2014.
[24] V. Ramaswami and G. Latouche. A general class of Markov processes with explicit matrix-geometric solutions. *Operations-Research-Spektrum*, 8(4):209–218, 1986.
[25] A. Riska and E. Smirni. Exact aggregate solutions for M/G/1-type Markov processes. In *ACM SIGMETRICS Performance Evaluation Review*, volume 30, pages 86–96. ACM, 2002.
[26] A. Riska and E. Smirni. ETAQA solutions for infinite Markov processes with repetitive structure. *INFORMS Journal on Computing*, 19(2):215–228, 2007.

[27] J. Selen, I. J. Adan, V. G. Kulkarni, and J. van Leeuwaarden. The snowball effect of customer slowdown in critical many-server systems. *Stochastic Models*, 32:366–391, 2016. MR3505449

[28] J. Selen, I. J. Adan, and J. S. Van Leeuwaarden. Product-form solutions for a class of structured multidimensional Markov processes. *SIAM Journal on Applied Mathematics*, 74(3):844–863, 2014.

[29] A. Sleptchenko, J. Selen, I. Adan, and G.-J. van Houtum. Joint queue length distribution of multi-class, single-server queues with preemptive priorities. *Queueing Systems*, 81(4):379–395, 2015.

[30] A. Stathopoulos, A. Riska, Z. Hua, and E. Smirni. Bridging ETAQA and ramaswami's formula for the solution of M/G/1-type processes. *Performance Evaluation*, 62(1):331–348, 2005.

[31] B. Van Houdt and J. van Leeuwaarden. Triangular M/G/1-Type and Tree-Like Quasi-Birth-Death Markov Chains. *INFORMS Journal on Computing*, 23(1):165–171, 2011.

[32] J. van Leeuwaarden, M. Squillante, and E. Winands. Quasi-birth-and-death processes, lattice path counting, and hypergeometric functions. *Journal of Applied Probability*, 46(2):507–520, 2009.

[33] J. van Leeuwaarden and E. Winands. Quasi-birth-and-death processes with an explicit rate matrix. *Stochastic Models*, 22(1):77–98, 2006.

Sherwin Doroudi
University of Minnesota
E-mail: sdoroudi@umn.edu

Brian Fralix
Clemson University
E-mail: bfralix@clemson.edu

Mor Harchol-Balter
Carnegie Mellon University
E-mail: harchol@cs.cmu.edu