

Tandem Queues with Hybrid Abandonment in Heavy Traffic

Cameron Curtis

`cameroncurtis2028@u.northwestern.edu`

Northwestern University

Izzy Grosof

Northwestern University

Seyed Iravani

Northwestern University

Research Article

Keywords: Queuing, Moment Generating Function, Markovian Arrival Process, Abandonment, Heavy Traffic, Asymptotics

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11309749/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Tandem Queues with Hybrid Abandonment in Heavy Traffic

Cameron Curtis^{1*}, Izzy Groszof¹ and Seyed Iravani¹

^{1*}Department of Industrial Engineering and Management Sciences,
Northwestern University, 2145 Sheridan Rd, Evanston, 60208, IL, US.

*Corresponding author(s). E-mail(s):

cameroncurtis2028@u.northwestern.edu;

Contributing authors: izzy.groszof@northwestern.edu;
s-iravani@northwestern.edu;

Abstract

Hybrid-abandonment systems, where jobs abandon from certain stages of service but not others, occur frequently in medical, service, and computing systems. Abandonment affects the heavy-traffic behavior of both the stages with and without abandonment. In contrast to existing heavy-traffic works that focus on cases where there is abandonment at either all stages or no stages, we focus on the heavy-traffic behavior for these hybrid-abandonment systems.

We provide the first heavy-traffic characterization of the second queue length in a two-stage tandem queue with abandonment only in the first stage. We characterize the scaled mean of the second stage queue length and prove that the scaled second queue length is asymptotically exponentially distributed and asymptotically independent of the first stage queue length.

We derive our results by analyzing Markovian arrival process queues with countably infinite phases. We do this by extending techniques used to analyze queues with Markov-modulated arrivals to the broader class of infinite Markovian arrival process queues. We prove asymptotic results for a more general infinite phase process condition, an infinite birth-death phase process, and the hybrid-abandonment tandem case.

We also empirically demonstrate that our heavy-traffic approximations are good approximations across various system settings.

Keywords: Queuing, Moment Generating Function, Markovian Arrival Process, Abandonment, Heavy Traffic, Asymptotics

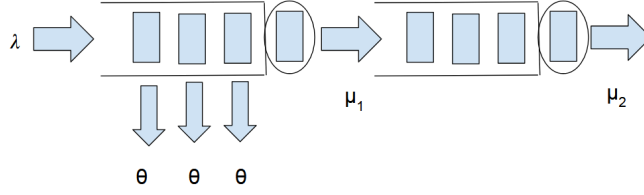


Fig. 1 A hybrid-abandonment tandem queue with abandonment in the first stage

1 Introduction

Existing literature on tandem queueing models, where jobs receive service at multiple stages in series, predominantly focus on systems where jobs can abandon from either no stages or all stages. These models are often a worthwhile proxy for certain real-world systems. This all-or-nothing approach to abandonment, though, leaves out systems where we observe abandonment in some stages but not in others.

In many real-world systems, the feasibility of abandoning may change significantly as a job moves through the system. We call systems like these with abandonment at some stages but not others *hybrid-abandonment* systems. We specifically focus on tandem systems with abandonment in the first stage but not the second stage as seen in Fig. 1. In a service context, an example of this structure is a coffee shop. While a customer waits in line, they can abandon but once their order is handed to the barista, abandonment is no longer possible. As another example, a prospective surgical patient on a waitlist might abandon to seek a different treatment. However, after surgery the patient cannot abandon from the waitlist for a follow-up appointment. In a manufacturing or other industrial example, this same model represents a raw material that can be repurposed until it undergoes some initial processing (cutting, drilling, etc.) and is then only fit for a particular use. In computing, a payment processing attempt may be abandoned before money moves between accounts, but once money leaves the payer's account, it must be placed into the recipient's account.

We study hybrid-abandonment queues in heavy traffic. The heavy-traffic regime considers a sequence of queues whose load (the fraction of time the second server is busy) approaches 1. This heavy-traffic limit is a worthwhile and valuable theoretical result on its own but also serves as a valuable approximation for the system's behavior in high-load settings (as we demonstrate in Section 6). Existing techniques to analyze tandem queues in heavy traffic are applicable to systems with either abandonment at no stages (see Jackson [1]) or abandonment at all stages (see Reed and Ward [2]) and don't easily generalize to hybrid-abandonment systems. We characterize the asymptotic distribution of both queue lengths in the heavy-traffic limit. We show that the first queue length and the scaled second queue length are asymptotically independent and that the scaled second queue length is asymptotically exponentially distributed and we compute its mean (Theorem 3).

We also generalize our results to provide sufficient conditions for analogous asymptotic independence and asymptotic exponentiality results for both Markovian arrival

processes with countably infinite phases (Theorem 1) and Markovian arrival processes with birth-death phase processes (Theorem 2).

1.1 Structure

In Section 2, we provide an overview of prior work relating to tandem queues, queues with Markovian-arrival processes, and heavy-traffic results for these systems and demonstrate the need for new methods in analyzing systems such as our hybrid-abandonment systems.

Next in Section 3 we define the necessary tools for our analysis: the infinite-phase Markovian arrival process (IMAP) queue and heavy-traffic families of IMAP queues. We also formally define the sufficient conditions we use in our analysis.

In Section 4, having defined all the necessary concepts in Section 3, we present our main theorems. We present our theorems from general to specific, beginning with our heavy-traffic analysis for general families of IMAP queues before narrowing to an analogous result for families of IMAP queues with birth-death phase processes. Finally, we present the narrowest result where we specifically consider hybrid-abandonment tandem systems.

In Section 5 we provide proofs of our main theorems and the necessary lemmas.

Lastly, in Section 6 we demonstrate via simulation that our heavy-traffic approximation is useful outside of the asymptotic regime.

2 Prior Work

The prior work we discuss can be broadly divided into three categories: works considering tandem queueing models with abandonment (Section 2.1), works considering systems with Markovian arrival processes (or MAPs) in heavy-traffic (Section 2.2), and works considering matrix-analytic methods (Section 2.3).

2.1 Tandem Queues with abandonment

While queueing networks with abandonment have been fairly well studied, far fewer works have considered hybrid-abandonment systems. This is especially true for systems with abandonment at earlier stages but not later stages.

Many of the works that have considered tandem queues with abandonment have done so in the dynamic control context. Zayas-Cabán and collaborators studied tandem queues in medical settings where patients proceed through two treatment stages. They considered dynamic control of servers in systems where customers can abandon in the second stage (as in Zayas-Caban et al. [3]) or in both stages (as in Zayas-Cabán and Cochran [4], Zayas-Cabán et al. [5]). Andradóttir and Ayhan [6] studied a similar system with abandonment only from the second stage and a finite second stage buffer and characterized the optimal dynamic server assignment in this system. While these works consider dynamic control of tandem systems, we focus on asymptotic behavior of static systems.

Reed and Ward [2] study a generalized Jackson network with abandonment at each station. They show that in heavy traffic such a system is well-approximated by a

reflected Ornstein-Uhlenbeck process. They consider a different heavy traffic regime, however. While they consider a regime where arrival rates and service rates jointly approach infinity and the abandonment rate stays fixed, we consider a regime where the arrival rate rises to a critical level while the service rates and abandonment rates stay fixed.

2.2 Markovian Arrival Processes in Heavy-Traffic

Queues with Markovian arrival processes (MAPs) and hybrid-abandonment tandem queues are conceptually similar and we draw on techniques designed for studying queues with Markovian arrival processes to study hybrid-abandonment tandem queues. In both cases, the arrivals to a single-server queue (the queue itself in the MAP case and the second stage queue in the tandem case) are determined by an independently evolving CTMC, the “phase process” in the MAP case and the first queue in the tandem case. The primary differences are two-fold: First, MAP queues allow arrivals to the queue to occur either while the process is in a state (“in-state arrivals”) or alongside a transition of the phase process (“between-state arrivals”). In contrast, arrivals to the second stage of the tandem system only occur alongside transitions in the first stage. Second, the tandem queues are usually allowed to have infinite buffers while MAP queues (with extremely limited exceptions) have only been considered in the case of finitely many phase states.

Choudhury and Whitt [7] study the tail behavior of the work in a queue with batch MAP arrivals, a generalization of MAPs that permits multiple arrivals at the same time, and general service times in heavy traffic. They show that in heavy traffic, the stationary distribution of the work in the queue has an exponential tail. We show that with an infinite-phase MAP and exponential service times, the scaled number of jobs in the queue is exponentially distributed rather than the work simply having an exponential tail.

Beyond Choudhury and Whitt [7], almost all of the work that considers MAP queues in heavy traffic specifically studies the special case of Markov-modulated arrival queues in heavy traffic. A Markov-modulated arrival queue is an MAP with only in-state arrivals, i.e., transitions of the phase process never cause arrivals. Yechiali and Naor [8] studied a system with two-level Markov-modulated arrival and service rates. In the two-level setting, they were able to derive the steady-state distribution and average queue length (up to a parameter being left as a root of a cubic equation). Their technique unfortunately doesn’t generalize to systems with more than two modulation states. Burman and Smith [9] provided the first heavy traffic result for queues with general finite-modulation-process Markov-modulated arrivals. They studied a system with Markov-modulated arrivals and general service times and provided a derivation of the mean queue length via a diffusion-scaling approach. They proved this result up to an interchange of limits assumption.

Falin and Falin [10] studied the same system and rigorously characterized the Laplace transform of the limiting heavy-traffic distribution. Dimitrov [11] extended the analysis of Falin and Falin [10] to a system with Markovian-service rates and arrival rates but still only considered the case of finite modulation processes. We extend Falin and Falin [10] by generalizing their methods to MAPs with infinitely many phases.

Mitrani [12] derives asymptotically exact approximations for the average queue length in Markov-modulated arrival queues. Rather than using a transform method like Falin and Falin [10] and Dimitrov [11], Mitrani [12] instead uses a spectral expansion-based technique to derive their approximations. As with Falin and Falin [10] and Dimitrov [11], Mitrani [12] only considers Markov-modulated arrival queues with finitely many phases while we consider the broader class of MAPs with countably infinite phases.

Thorsdottir and Verloop [13] consider the heavy-traffic behavior of M/G/1 queues with Markov-modulated arrivals and show that the work in the system is asymptotically exponentially distributed. We focus on the number of jobs in queue rather than the workload and consider models with MAPs and countably infinite phases.

Following the works by Falin and Falin [10] and Dimitrov [11], equivalent results were independently derived using the drift method as stepping stones to other results. Mou and Maguluri [14], for example, analyzed input-switched Markovian models in heavy traffic. They independently reproved via the drift method that the expected scaled queue length, $\mathbb{E}[(1 - \rho)Q]$, converges in heavy traffic and expressed the limiting value as an infinite summation.

Grosof et al. [15] derived the waiting time in first-come-first-served multiserver-job models. As part of their analysis, they independently reproved expected queue length results for systems with Markov-modulated service rates via the drift method. We consider the more general case of MAP queues instead and the case of infinitely many phase states.

In work currently under submission, Hurtado-Lange and Grosof [16] consider the heavy-traffic behavior of the join-the-shortest-queue load-balancing policy in queues with Markov-modulated arrivals, including arrival processes with countably infinite modulation states. While they assume sufficient conditions for their heavy-traffic results, we prove our heavy-traffic conditions for the class of hybrid-abandonment tandem systems.

2.3 Matrix Analytic Methods

Matrix analytic methods can be used to numerically compute performance metrics for a variety of queueing systems via matrix computations [17–20]. The methods were first used in Neuts [17] to compute queue length, busy period lengths, and other performance metrics. Regterschot and De Smit [21] used matrix analytic methods to find computationally tractable bounds and approximations for a number of performance metrics.

Ciucu and Poloczek [22] provided a more computationally efficient method to derive tight bounds on the transform of a system with Markov-modulated arrivals (among other systems).

Wang et al. [23] studied tandem queues with abandonment at both stations as level-dependent quasi-birth-death processes. This allowed them to numerically compute performance metrics of such a system efficiently, but required the second queue have a finite buffer. Our work considers the case where the second queue has an infinite buffer and no abandonment.

Matrix analytic methods provide numerically efficient methods to bound performance metrics in a wide array of stochastic systems, but we focus on exact characterization of these metrics and, particularly, asymptotic analysis in the heavy-traffic limit rather than performance in any single system. Matrix analytic methods also tend to focus on systems with one-dimensionally infinite state spaces, such as Markov-modulated queues with finitely many modulation states, rather than systems with doubly-infinite state spaces, such as arrival processes with countably infinite state spaces as in our hybrid-abandonment system.

The few matrix-analytic works that do consider arrival processes with countably infinite state spaces do not characterize the system’s heavy-traffic behavior. By contrast, we characterize heavy-traffic behavior, which allows us to derive simple and accurate approximations for the queueing behavior at all loads, as we show in Section 6. Ramaswami and Taylor [24] introduce an operator-theoretic method to characterize the rate matrix $\mathbf{R}$ and matrix $\mathbf{G}$ in level-dependent, infinite-phase quasi-birth-death processes, including MAP queues with countably infinite phases. Their techniques allow them to analyze particular infinite phase MAP queues but do not allow them to analyze heavy-traffic behavior or produce simple approximations. Other works that consider infinite phase MAP queues do so by building upon Ramaswami and Taylor [24]. Latouche et al. [25] and Kroese et al. [26], for instance, consider when the behavior of a sequence of truncated finite-phase systems converge to the behavior of an untruncated infinite-phase system. Lian and Zhao [27] uses a similar framework to create an algorithm to compute inter-departure times and busy period of different stages in a tandem queue. In contrast to works like Ramaswami and Taylor [24] that focus on numerically understanding the behavior of particular IMAP systems, we focus on the heavy-traffic behavior of tandem systems and IMAP queues.

3 Model and Definition

Our ultimate goal is to derive structural insights into the behavior of the hybrid-abandonment system across various loads. To do so, we seek to characterize the heavy-traffic behavior of the system, which we will use in Section 6 to derive such insights. To prove our heavy-traffic results, we begin by defining a generalization of the hybrid-abandonment queue, the infinite-phase Markovian arrival process queue, or IMAP queue, in Section 3.1. We show that the hybrid-abandonment queue is a special case of the infinite-phase Markovian arrival process queue in Section 3.2. Finally, we define the near-inverse matrix and M function for an IMAP queue (Section 3.3) and the families of IMAP queues we consider in our analysis (Section 3.4).

3.1 The Infinite-phase Markovian Arrival Queue (IMAP) Queue

We begin by defining a finite-phase Markovian arrival process (or “MAP”) queue before distinguishing our infinite-phase Markovian arrival process queue. The arrivals to an MAP queue are controlled by a finite-state CTMC Y , the “phase” of the arrival process, which evolves independently of the queue length process Q . While Y is in state i , arrivals to Q occur at rate λ_i . We will call these “in-state” arrivals. When Y

transitions from state i to state j , there is some probability p_{ij} that an arrival to Q is triggered. We will call these arrivals “between-state” arrivals. If $p_{ij} = 0$ for all i, j , the MAP queue is simply a Markov-modulated arrival queue.

The infinite-phase Markovian arrival process (or “IMAP”) queue generalizes the MAP queue to allow Y to have countably infinite states. Without loss of generality, we will assume that these states are $\mathbb{N}_0 = \{0, 1, 2, \dots\}$.

The phase process $Y(t)$ has a generator matrix $\mathbf{A} = [\alpha_{i,j}]_{i,j \in \mathbb{N}_0}$. Note that throughout this paper, we use bold font to indicate matrices and vectors. When $(Y(t), Q(t))$ has a unique steady-state distribution, we denote the steady-state probability of (Y, Q) as $\pi_{i,q} = \mathbb{P}(Y = i, Q = q)$ and the steady-state distribution of Y alone as $\boldsymbol{\pi}$ where $[\boldsymbol{\pi}]_i = \pi_i = \mathbb{P}(Y = i)$. When $Y(t) = i$, the instantaneous arrival rate is $\lambda_i + \sum_{j \neq i} p_{i,j} \alpha_{i,j}$ (the instantaneous rate of in-state arrivals plus the instantaneous rate of between-state arrivals). Thus, the long-run arrival rate to the queue is

$$\bar{\lambda} = \mathbb{E} \left[\lambda_Y + \sum_{j \neq Y} p_{Y,j} \alpha_{Y,j} \right] = \sum_i \pi_i \left(\lambda_i + \sum_{j \neq i} p_{i,j} \alpha_{i,j} \right).$$

We often collect the in-state and between-state arrival rates into the $\mathbf{F}$ matrix where the elements of $\mathbf{F}$ are given by:

$$[\mathbf{F}]_{i,j} = F_{i,j} = \begin{cases} \lambda_i, & i = j \\ p_{i,j} \alpha_{i,j}, & i \neq j \end{cases}$$

for $i, j \in \mathbb{N}_0$.

Most of our analysis of this system is done in terms of the system’s z -transform. The z -transform of the queue length random variable,

$$\bar{q}(z) = \mathbb{E}[z^Q],$$

captures the marginal distribution of our queue length random variable while the partial z -transforms,

$$\hat{q}(z; i) = \mathbb{E}[z^Q \mathbb{I}\{Y = i\}],$$

captures the dependence between the queue length random variable and the phase process.

Often in our analysis, it is useful to refer to $\boldsymbol{\pi}_{\text{idle}} = [P(Y = i, Q = 0)]_{i \in \mathbb{N}_0}$, the idleness vector, and $\varepsilon = P(Q = 0)$, the system’s total idleness. We also let $\mathbf{e}^T$ be the ones vector, i.e., $\mathbf{e}^T = [1, 1, \dots]^T$.

3.2 Hybrid-Abandonment Tandem Queues

We are particularly interested in the hybrid-abandonment systems we discussed in Section 1. These are tandem queues where jobs enter the first stage of queueing and either abandon the first queue or receive service and enter the second queue as depicted in Fig. 1. These are a subset of the IMAP queues we introduced above. To model a

hybrid-abandonment queue as an IMAP queue we will treat the first queue length as the phase process Y and the second queue length as the queue length process Q . When n jobs are in the first queue, the first queue transitions upward with rate $\alpha_{n,n+1} = \lambda$ and downward with rate $\alpha_{n,n-1} = (\mu_1 + (n-1)\theta) \mathbb{I}\{n > 0\}$. When the first queue transitions downward it triggers a between-state arrival to the second queue with probability $p_{n,n-1} = \frac{\mu_1}{\mu_1 + (n-1)\theta}$. These downward between-state arrivals are the only arrivals to the second queue. When discussing the hybrid-abandonment systems we refer to the service rate of the second queue as μ_2 rather than simply μ for clarity.

3.3 Near-Inverse Matrix $\mathbf{R}$ and $\mathbf{M}$ Function

Often in our analysis, a quantity we are interested in appears right-multiplied by the generator matrix $\mathbf{A}$. A way to “undo” the effects of this multiplication and recover the quantity of interest would therefore be extremely useful. However, because $\mathbf{A}$ is a generator matrix, it is not invertible. Lacking a true inverse, we often utilize a *near-inverse* of $\mathbf{A}$. Given a generator matrix $\mathbf{A}$ with steady state distribution $\boldsymbol{\pi}$ and state space $\mathbb{N}_0$, define the vector $\boldsymbol{\pi}_{\bar{0}} = [\pi_i]_{i>0}$. We do so without loss of generality: we may choose any mapping from the states of $\mathbf{A}$ to $\mathbb{N}_0$, and state 0 may be chosen arbitrarily. We then define a near-inverse matrix $\mathbf{R}$ to be a matrix that satisfies our near-inverse equation:

$$\mathbf{A}\mathbf{R} = \begin{bmatrix} 0 & -\frac{1}{\pi_0}\boldsymbol{\pi}_{\bar{0}} \\ 0 & \mathbf{I} \end{bmatrix} = \begin{bmatrix} 0 & -\frac{\pi_1}{\pi_0} & -\frac{\pi_2}{\pi_0} & \dots & -\frac{\pi_n}{\pi_0} & \dots \\ 0 & 1 & 0 & \dots & 0 & \dots \\ 0 & 0 & 1 & \dots & 0 & \dots \\ \vdots & & & & & \\ 0 & 0 & 0 & \dots & 1 & \dots \\ \vdots & & & & & \end{bmatrix}$$

or, equivalently,

$$\mathbf{A}\mathbf{R}\mathbf{x} = \mathbf{x} - \frac{x_0}{\pi_0}\boldsymbol{\pi}$$

for any vector $\mathbf{x}$.

In Lemma 1, we construct $\mathbf{R}$ as the time-in-state matrix by letting $[\mathbf{R}]_{i,j} = r_{i,j}$ be the expected time in state j between starting in state i and reaching state 0 for the first time and showing that it satisfies our near-inverse equation.

We use the notation $\mathbf{R}$ for our near-inverse to match the notation used by Falin and Falin [10]. However, it should be noted that our near-inverse matrix $\mathbf{R}$ is not the same as the rate matrix $\mathbf{R}$ that often appears in matrix-analytic works. While they share a symbol and have some superficial similarities (both are concerned with time spent in or visits to particular states before a certain event occurs), they are not meaningfully related. In the rate matrix $\mathbf{R}$, the element $[\mathbf{R}]_{i,j}$ is the expected number of visits to state $(m+1, j)$ between starting in state (m, i) and first returning to queue length m (which is independent of the particular queue length m) [28].

Once we define a near-inverse matrix $\mathbf{R}$ for an IMAP queue, we can also define the $\mathbf{M}$ function for the same system:

$$M(z) = 1 - \mu^{-1}\boldsymbol{\pi}(-z\mathbf{F} + \mu\mathbf{I})\mathbf{R}(-z\mathbf{F} + \mu\mathbf{I})e^T. \quad (1)$$

In families of queues with the necessary convergence conditions (including our original hybrid-abandonment queues as we demonstrate in Lemma 5), we prove that M 's asymptotic behavior is the same as the asymptotic scaled mean queue length. Thus, characterizing M and $\mathbf{R}$ are key points in our analysis.

3.4 Families of IMAP Queues

We frequently consider families of IMAP queues indexed by their idleness in our analysis. A family of IMAP queues indexed by idleness is a map from ε in some $\mathcal{E} = (0, \varepsilon_{\max})$ to IMAP queues such that the ε^{th} queue has idleness ε . The heavy-traffic limit in the system is a limit of the IMAP queues as $\varepsilon \rightarrow 0$. We are specifically interested in families of IMAP queues that, like our hybrid-abandonment queues, have a particular notion of “convergence” in the $\varepsilon \rightarrow 0$ limit, namely that the phase process and the queue process are well-behaved in this limit. Below we present the definition of our *heavy-traffic conditions* for an IMAP family that we use throughout our analysis. We defer proving that our original hybrid-abandonment queues have these features to Lemma 5.

Definition 1 (Heavy-Traffic Conditions) A family of IMAP queues parameterized by their idleness ε for $\varepsilon \in \mathcal{E}$ that satisfies our *heavy-traffic conditions* is a family of IMAP queues where there exist matrices $\mathbf{F}^{(0)}$ and $\mathbf{R}^{(0)}$, $\mu^{(0)} \in (0, \infty)$, and a distribution $\pi^{(0)}$ such that:

1. (a) For all $i \in \mathbb{N}_0$, $\lim_{\varepsilon \rightarrow 0} \pi_i^{(\varepsilon)} = \pi_i^{(0)}$.
 (b) For all $i, j \in \mathbb{N}_0$, $\lim_{\varepsilon \rightarrow 0} r_{i,j}^{(\varepsilon)} = r_{i,j}^{(0)}$.
 (c) For all $i, j \in \mathbb{N}_0$, $\lim_{\varepsilon \rightarrow 0} \mathbf{F}_{i,j}^{(\varepsilon)} = \mathbf{F}_{i,j}^{(0)}$.
 (d) $\lim_{\varepsilon \rightarrow 0} \mu^{(\varepsilon)} = \mu^{(0)}$.
2. Let $M^{(0)} = 1 - (\mu^{(0)})^{-1} \pi^{(0)} \left(-\mathbf{F}^{(0)} + \mu^{(0)} \mathbf{I} \right) \mathbf{R}^{(0)} \left(-\mathbf{F}^{(0)} + \mu^{(0)} \mathbf{I} \right) e^T$.
 For all $z \in (0, 1)$, $\lim_{\varepsilon \rightarrow 0} M(z^\varepsilon; \varepsilon) = M^{(0)}$.
3. $Y^{(\varepsilon)}$ is ergodic for all $\varepsilon \in \mathcal{E}$.

To aid in our analysis, we establish a few additional conditions for families of IMAP queues. These turn out to be useful sufficient conditions for families of IMAP queues to have well-defined heavy traffic asymptotic distributions.

Definition 2 (Rate Bounded) A *rate-bounded* family of IMAP queues is a family of IMAP queues where, there exist some $\hat{\lambda} > 0$, $T_{\max} > 0$, and $\hat{\mu} > 0$ such that for all modulation states and all possible idleness $\varepsilon \in \mathcal{E}$,

1. The instantaneous arrival rate to the queue is upper bounded by $\hat{\lambda}$.
2. The expected time until transition in any modulation state is upper bounded by $T_{\max}$.
3. The service rate is lower bounded by $\hat{\mu}$.

The *rate-bounded* condition essentially exists to rule out a few particularly ill-behaved families of IMAP queues. A family being rate bounded guarantees that the

phase process does not linger too long in any state or become explosive both for particular ε values and in the $\varepsilon \rightarrow 0$ limit.

Definition 3 (Birth-Death) A *birth-death* family of IMAP queues is a family of IMAP queues where $Y^{(\varepsilon)}$ is a birth-death process on $\mathbb{N}_0$ for all $\varepsilon \in \mathcal{E}$, i.e., for all $\varepsilon \in \mathcal{E}$, $Y^{(\varepsilon)}(t) \in \{0, 1, 2, \dots\}$ and from $Y^{(\varepsilon)} = i$, $Y^{(\varepsilon)}$ can only transition to $i - 1$ or $i + 1$.

A *birth-death* family of IMAP queues is simply an extension of the concept of a birth-death phase process to cover all $\varepsilon \in \mathcal{E}$.

Definition 4 (Uniformly Stable) A *uniformly stable* birth-death family of IMAP queues is a family of IMAP queues where there is some N , $a \in (0, 1)$, and $b \in (0, \frac{1}{2})$ such that:

1. For all $\varepsilon \in \mathcal{E}$ and all $n \geq N$, $\frac{\alpha_{n,n+1}^{(\varepsilon)}}{\alpha_{n+1,n}^{(\varepsilon)}} \leq a$
2. For all $\varepsilon \in \mathcal{E}$ and all $n \geq N$, $\frac{\alpha_{n,n+1}^{(\varepsilon)}}{\alpha_{n,n-1}^{(\varepsilon)} + \alpha_{n,n+1}^{(\varepsilon)}} \leq b$

Intuitively, a family of IMAP queues is *uniformly stable* if the phase process has a downward drift for all $\varepsilon \in \mathcal{E}$ when n is “big enough”.

Lastly, we have our desired characteristic: that a family of IMAP queues is asymptotically exponential and asymptotically independent.

Definition 5 (Asymptotically Exponential and Independent) An *asymptotically exponential and independent* family of IMAP queues with a heavy-traffic convergent phase process with mean M is a family of IMAP queues with heavy-traffic convergent phase process where

$$\lim_{\varepsilon \rightarrow 0} \bar{q}(z^\varepsilon; \varepsilon) = \frac{1}{1 + M \ln z} \quad (2)$$

and

$$\lim_{\varepsilon \rightarrow 0} \hat{q}(z^\varepsilon; i, \varepsilon) = \frac{\pi_i^{(0)}}{1 + M \ln z} \quad (3)$$

for all $z \in (0, 1)$ and all $i \in \mathbb{N}_0$.

Eq. (2) states that the z -transform of the random variable εQ approaches that of an exponential random variable with mean M . This is exactly equivalent to εQ converging in distribution to an exponential random variable with mean M .

Eq. (3) states that for any modulation state i , the queue length conditioned on the modulation state also is an exponential random variable with mean M . Thus, the queue length and phase process are independent.

The asymptotic exponential and independence properties are the focuses of our analysis. In each of Theorems 1 to 3 we will focus on showing that, under particular conditions, a family of IMAP queues has these asymptotic properties.

4 Main Results

We characterize the asymptotic behavior of hybrid abandonment queues as the second queue’s idleness approaches zero. We begin by analyzing a considerably broader class of models, infinite-phase Markovian arrival processes with “light-tailed phase processes” (Definition 6) and describing their behavior in the same heavy-traffic limit. This “light-tailed” condition can be roughly described as a weighted average of $\mathbf{R}$ values having finite first and second moments.

Here we present statements of each of Theorems 1, 2, and 3 before presenting a full proof of each in Sections 5.3, 5.6 and 5.8, respectively.

Theorem 1 *If a family of IMAP queues with a heavy-traffic convergent phase process has a light-tailed phase process (Definition 6, which will be defined in Section 5.3), then $\varepsilon Q^{(\varepsilon)}$ is asymptotically exponential with mean $M^{(0)}$ (where $M^{(0)}$ is defined in Definition 1) and asymptotically independent from $Y^{(\varepsilon)}$ in the $\varepsilon \rightarrow 0$ limit. Equivalently,*

$$\lim_{\varepsilon \rightarrow 0} \bar{q}(z; \varepsilon) = \frac{1}{1 + M^{(0)} \ln z}, \quad \lim_{\varepsilon \rightarrow 0} \hat{q}(z; i, \varepsilon) = \frac{\pi_i^{(0)}}{1 + M^{(0)} \ln z}.$$

Proof deferred to Section 5.3.

Next, we specialize our theorem to specifically focus on birth-death phase processes. We show that when the queue has a birth-death phase-process a set of easier-to-check conditions, rate-boundedness (see Definition 2) and uniform stability (see Definition 4), implies our light-tailed condition and thus implies the same asymptotic behavior.

Theorem 2 *In a family of IMAP queues with birth-death phase processes that meets our heavy traffic conditions, if the family is uniformly stable and rate bounded then it also has a light-tailed phase process (Definition 6). Consequently $\varepsilon Q^{(\varepsilon)}$ is asymptotically exponential with mean $M^{(0)}$ (where $M^{(0)}$ is defined in Section 3.3) and asymptotically independent from $Y^{(\varepsilon)}$ in the $\varepsilon \rightarrow 0$ limit. Equivalently,*

$$\lim_{\varepsilon \rightarrow 0} \bar{q}(z; \varepsilon) = \frac{1}{1 + M^{(0)} \ln z}, \quad \lim_{\varepsilon \rightarrow 0} \hat{q}(z; i, \varepsilon) = \frac{\pi_i^{(0)}}{1 + M^{(0)} \ln z}.$$

Proof deferred to Section 5.6.

Intuitively, rate-boundedness and uniform stability ensure that the queue isn’t ill-behaved at any ε or in the $\varepsilon \rightarrow 0$ limit. Rate-boundedness guarantees that transitions don’t occur arbitrarily quickly (to prevent explosivity) or arbitrarily slowly (to prevent any states becoming isolated). Uniform stability guarantees that for “large enough” states the queue always has a downward transition bias across all ε values.

Finally, we return to our original system of interest, the tandem queue with hybrid abandonment. We show that a tandem queue with hybrid abandonment meets the birth-death conditions of Theorem 2 and satisfies our heavy-traffic conditions and therefore has the desired asymptotic behavior.

Theorem 3 *In a tandem queue with hybrid abandonment, as the arrival rate rises to a critical threshold and the second queue’s idleness approaches 0, $\varepsilon Q^{(\varepsilon)}$ is asymptotically exponential with mean $M^{(0)}$ (where $M^{(0)}$ is defined in Definition 1) and asymptotically independent from $Y^{(\varepsilon)}$. Equivalently,*

$$\lim_{\varepsilon \rightarrow 0} \bar{q}(z; \varepsilon) = \frac{1}{1 + M^{(0)} \ln z}, \quad \lim_{\varepsilon \rightarrow 0} \hat{q}(z; i, \varepsilon) = \frac{\pi_i^{(0)}}{1 + M^{(0)} \ln z}.$$

Proof deferred to Section 5.8.

5 Proofs

We present the proofs for Theorems 1 to 3 as well as the necessary lemmas for these proofs.

Lemma 1 and Lemma 2 derive near-inverse matrix and the near-independence and near-exponential equations, respectively, that serve as the bedrock of our asymptotic analyses.

Theorem 1 analyzes these near-independence and near-exponential equations to show that if the phase processes of a heavy-traffic family are “light-tailed” (which we will define in Definition 6 in Section 5.3), i.e., necessary quantities are negligible, then the near-independence and near-exponential equations imply the asymptotic-independence and asymptotic-exponential equations of Definition 5.

Lemma 3 provides three fundamental bounds on the quantities $\boldsymbol{\pi}$ and $\boldsymbol{R}$ when considering birth-death heavy-traffic families that are rate-bounded (Definition 2) and uniformly stable (Definition 4). Lemma 4 uses these fundamental bounds to verify that these birth-death heavy-traffic families satisfy the light-tailed phase process conditions necessary for Theorem 1. Theorem 2 applies Theorem 1 to these birth-death heavy-traffic families to show the desired asymptotic exponential and asymptotic-independence results.

Lemma 5 proves that hybrid-abandonment families are birth-death heavy-traffic families that are uniformly stable and rate-bounded. Theorem 3 then applies Theorem 2 to these hybrid-abandonment families to finally reach the asymptotic exponential and asymptotic independence equations for our original hybrid-abandonment system.

5.1 Existence of the Near-Inverse Matrix

We begin by showing that a near-inverse matrix $\boldsymbol{R}$ exists for all the systems we consider in Theorems 1 to 3.

Lemma 1 *For any ergodic continuous time Markov chain Y with generator matrix $\mathbf{A}$, there exists a near-inverse matrix $\mathbf{R} = [r_{i,j}]_{i,j \in \mathbb{N}_0}$ satisfying*

$$\mathbf{A}\mathbf{R} = \begin{bmatrix} 0 & -\frac{1}{\pi_0}\pi_0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} 0 & -\frac{\pi_1}{\pi_0} & -\frac{\pi_2}{\pi_0} & \dots & -\frac{\pi_n}{\pi_0} & \dots \\ 0 & 1 & 0 & \dots & 0 & \dots \\ 0 & 0 & 1 & \dots & 0 & \dots \\ \vdots & & & & & \\ 0 & 0 & 0 & \dots & 1 & \dots \\ \vdots & & & & & \end{bmatrix}$$

or, equivalently,

$$\mathbf{A}\mathbf{R}x = x - \frac{x_0}{\pi_0}\pi$$

for any vector x .

Note that π_0 is defined in Section 3.3.

We give a full proof in Appendix B.1. A sketch of our proof is the following. We begin by constructing the matrix $\mathbf{R} = [r_{i,j}]_{i,j \in \mathbb{N}_0}$ by making $r_{i,j}$ be the expected time the chain spends in state j between starting in state i and reaching state 0 for the first time. We show that when $\mathbf{A}$ is ergodic each of these elements exist and then that $\mathbf{R}$ satisfies our desired *near-inverse* property.

Intuitively, this near-inverse matrix allows us to mostly cancel out the matrix $\mathbf{A}$. This allows us to convert the balance equations of Lemma 2 into the more useful near-independence equation Eq. (4) and near-exponential equation Eq. (5).

In the interest of clarity, we do not rule out the possibility that there exists another matrix $\mathbf{R}' \neq \mathbf{R}$ that satisfies the near-inverse property. Whenever we invoke $\mathbf{R}$ or the near-inverse matrix we consider this formulation of $\mathbf{R}$.

5.2 Near-Independence and Near-Exponential Equations

With our near-inverse matrix guaranteed to exist by Lemma 1, we can establish analogous equations to those in Falin and Falin [10]. These equations are the basis of our heavy traffic analysis. In particular the most general of the theorems, Theorem 1, requires only that we establish the limiting behavior of each of the terms in the near-independent and near-exponential equations.

Lemma 2 *For any infinite-phase Markovian arrival process, if there exists a near-inverse matrix $\mathbf{R}$ then the system has the following near-independence and near-exponential equations for any $M > 0$:*

$$\hat{q}(z) = \hat{q}(z; 0) \frac{\pi}{\pi_0} + \hat{q}(z; 0) \frac{\pi}{\pi_0} \left(\mathbf{F}(1-z) + \mu \left(1 - \frac{1}{z} \right) \mathbf{I} \right) \mathbf{R} + \bar{\mathbf{Y}}(z) \quad (4)$$

$$\hat{q}(z; 0) = \pi_0 \frac{1 + U(z)}{1 + M \ln(z)/\varepsilon + V(z)} \quad (5)$$

where

$$\bar{\mathbf{Y}}(z) = \left(1 - \frac{1}{z} \right) \hat{q}(z) \left((-z\mathbf{F} + \mu\mathbf{I}) \mathbf{R} \right)^2 - \mu \left(1 - \frac{1}{z} \right) \pi_{idle} \mathbf{R} (-z\mathbf{F} + \mu\mathbf{I}) \mathbf{R} - \mu \pi_{idle} \mathbf{R}$$

$$\begin{aligned}
U(z) &= -\mu^{-1}\varepsilon^{-1} \left(1 - \frac{1}{z}\right) \bar{\mathbf{Y}}(z) (-z\mathbf{F} + \mu\mathbf{I}) e^T \\
V(z) &= \varepsilon^{-1} \left(1 - \frac{1}{z}\right) \left(-z\frac{\bar{\lambda}}{\mu} + \mu^{-1}\boldsymbol{\pi}(-z\mathbf{F} + \mu\mathbf{I})\mathbf{R}(-z\mathbf{F} + \mu\mathbf{I})e^T\right) - M \ln(z)/\varepsilon.
\end{aligned}$$

We give a full proof in Appendix B.2, but here we present a brief proof sketch. By careful manipulation and summation, the balance equation for the Markov chain (Y, Q) can be converted into a balance equation for the partial z -transform vector $\hat{\mathbf{q}}(z)$, Eq. (B1), providing $\hat{\mathbf{q}}(z)$ in terms of $\hat{\mathbf{q}}(z)$ and $\hat{\mathbf{q}}(z; 0)$. By substituting Eq. (B1) for $\hat{\mathbf{q}}(z)$ in Eq. (B1) itself, we now have a new formula for $\hat{\mathbf{q}}(z)$ which we use to derive our near-independence and near-exponential equations.

We call these equations *near-independence* and *near-exponential* to specifically draw parallels with “true-independence” and “true-exponential” equations,

$$\hat{\mathbf{q}}(z) = \hat{\mathbf{q}}(z; 0) \frac{\boldsymbol{\pi}}{\pi_0} \quad (6)$$

$$\hat{\mathbf{q}}(z; 0) = \pi_0 \frac{1}{1 + M \ln z}. \quad (7)$$

Eqs. (6) and (7) exactly state that the variable Q is independent of Y and Q has an exponential distribution with mean M , respectively. In Theorem 1 we show that, under specified conditions, Eqs. (2) and (3) asymptotically become Eqs. (6) and (7), respectively.

5.3 Asymptotic Independence and Exponentiality for Light-tailed Modulation Processes

With our lemmas finished, we are now ready to begin considering the heavy traffic limit. We no longer consider a single system, but a family of systems parameterized by idleness ε . In our proofs for Theorems 1 to 3, we let $x^{(\varepsilon)}$ refer to the quantity x in the system with idleness ε . For example, $\mu^{(\varepsilon)}$ refers to the service rate of the queue with idleness ε and $\boldsymbol{\pi}^{(\varepsilon)}$ refers to the stationary distribution of the phase process in the same system.

We start with the most general case: considering any convergent heavy-traffic family which satisfies the light-tailed phase process conditions described below. We show that these light-tailed conditions are sufficient to characterize the heavy-traffic limit of the family even when the phase process has a countably infinite state space.

Definition 6 (Light-Tailed Phase Process) A family of IMAP queues that meets our heavy-traffic conditions (Definition 1) with a *light-tailed phase process* is a family of IMAP queues that meets our heavy-traffic conditions where:

1. For all $z \in (0, 1)$, as $\varepsilon \rightarrow 0$:

- (a) $(1 - z^{-\varepsilon}) \hat{\mathbf{q}}(z^\varepsilon; \varepsilon) \left((-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I}) \mathbf{R}^{(\varepsilon)} \right)^2 (-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu \mathbf{I}) e^T \rightarrow 0$
- (b) $\mu^{(\varepsilon)} (1 - z^\varepsilon) \pi_0^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} \left((-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I}) \mathbf{R}^{(\varepsilon)} \right) (-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I}) e^T \rightarrow 0$

- (c) $\pi_0^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T \rightarrow 0$
- (d) $(1 - z^{-\varepsilon}) \pi^{(\varepsilon)} \left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right) \mathbf{R}^{(\varepsilon)} e^T \rightarrow 0$
- (e) $(1 - z^{-\varepsilon}) \hat{q}(z^\varepsilon; \varepsilon) \left(\left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right) \mathbf{R}^{(\varepsilon)} \right)^2 e^T \rightarrow 0$
- (f) $\mu^{(\varepsilon)} (1 - z^{-\varepsilon}) \pi_{\text{idle}}^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} \left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right) \mathbf{R}^{(\varepsilon)} e^T \rightarrow 0$
- (g) $\mu^{(\varepsilon)} \pi_{\text{idle}}^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} e^T \rightarrow 0$
- 2. For all $z \in (0, 1)$ and all $i \in \mathbb{N}_0$ as $\varepsilon \rightarrow 0$:
 - (a) $[(1 - z^{-\varepsilon}) \pi^{(\varepsilon)} \left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right) \mathbf{R}^{(\varepsilon)}]_i \rightarrow 0$
 - (b) $[\hat{q}(z^\varepsilon; \varepsilon) \left(\left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) \mathbf{R} \right)^2]_i \rightarrow 0$
 - (c) $[\mu^{(\varepsilon)} (1 - z^{-\varepsilon}) \pi_0^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) \mathbf{R}^{(\varepsilon)}]_i \rightarrow 0$
 - (d) $[\mu^{(\varepsilon)} \pi_0^{(\varepsilon)} \mathbf{R}^{(\varepsilon)}]_i \rightarrow 0$

Each of these terms appear in at least one of $U(z^\varepsilon)$, $V(z^\varepsilon)$, $\bar{Y}(z^\varepsilon)$, or $\bar{Y}_i(z^\varepsilon)$. With all of these terms being negligible, we can conclude (as we do in Theorem 1) that we have our desired heavy traffic results.

Theorem 1 *If a family of IMAP queues with a heavy-traffic convergent phase process has a light-tailed phase process, then $\varepsilon Q^{(\varepsilon)}$ is asymptotically exponential with mean $M^{(0)}$ (where $M^{(0)}$ is defined in Definition 1) and asymptotically independent from $Y^{(\varepsilon)}$ in the $\varepsilon \rightarrow 0$ limit. Equivalently,*

$$\lim_{\varepsilon \rightarrow 0} \bar{q}(z^\varepsilon; \varepsilon) = \frac{1}{1 + M^{(0)} \ln z} \quad (8)$$

$$\lim_{\varepsilon \rightarrow 0} \hat{q}(z^\varepsilon; i, \varepsilon) = \frac{\pi_i^{(0)}}{1 + M^{(0)} \ln z} \quad (9)$$

Proof We now shift our focus to the scaled system $\varepsilon Q^{(\varepsilon)}$ rather than the original system $Q^{(\varepsilon)}$. Notice that

$$\mathbb{E}[z^{\varepsilon Q^{(\varepsilon)}}] = \mathbb{E}[(z^\varepsilon)^{Q^{(\varepsilon)}}] = \bar{q}(z^\varepsilon; \varepsilon)$$

and

$$\mathbb{E} \left[z^{\varepsilon Q^{(\varepsilon)}} \mathbb{I}\{Y^{(\varepsilon)} = i\} \right] = \mathbb{E} \left[(z^\varepsilon)^{Q^{(\varepsilon)}} \mathbb{I}\{Y^{(\varepsilon)} = i\} \right] = \hat{q}(z^\varepsilon; i, \varepsilon).$$

Therefore, our analysis of $\varepsilon Q^{(\varepsilon)}$ can rely on our existing near-independence and near-exponential equations Eqs. (4) and (5) for $Q^{(\varepsilon)}$.

We break our result into three parts. In part 1, we show that $[\varepsilon Q^{(\varepsilon)} | Y^{(\varepsilon)} = 0]$ is asymptotically exponential or

$$\lim_{\varepsilon \rightarrow 0} \hat{q}(z^\varepsilon; 0, \varepsilon) = \frac{\pi_0^{(0)}}{1 + M^{(0)} \ln z}. \quad (10)$$

Next, in part 2, we show this for any $i \in \mathbb{N}_0$, $[\varepsilon Q^{(\varepsilon)} | Y^{(\varepsilon)} = i]$ is asymptotically exponential or

$$\lim_{\varepsilon \rightarrow 0} \hat{q}(z^\varepsilon; i, \varepsilon) = \frac{\pi_i^{(0)}}{1 + M^{(0)} \ln z}. \quad (9)$$

Finally, in part 3, we show that $\varepsilon Q^{(\varepsilon)}$ (absent any conditioning on $Y^{(\varepsilon)}$) is asymptotically exponential or

$$\lim_{\varepsilon \rightarrow 0} \bar{q}(z^\varepsilon; \varepsilon) = \frac{1}{1 + M^{(0)} \ln z}. \quad (8)$$

5.3.1 Part 1: Asymptotic Exponentiality Conditioned on State 0

First, we show that $[\varepsilon Q^{(\varepsilon)} | Y^{(0)} = 0]$ is asymptotically exponential by proving Eq. (10).

By Lemma 2, for any member of the parameterized family we have

$$\begin{aligned} \hat{q}(z; \varepsilon) &= \hat{q}(z; 0, \varepsilon) \frac{\pi^{(\varepsilon)}}{\pi_0^{(\varepsilon)}} + \hat{q}(z; 0, \varepsilon) \frac{\pi^{(\varepsilon)}}{\pi_0^{(\varepsilon)}} \left(\mathbf{F}^{(\varepsilon)} (1 - z) + \mu^{(\varepsilon)} \left(1 - \frac{1}{z} \right) \mathbf{I} \right) \mathbf{R}^{(\varepsilon)} + \bar{\mathbf{Y}}(z; \varepsilon) \\ \hat{q}(z; 0, \varepsilon) &= \pi_0^{(\varepsilon)} \frac{1 + U(z; \varepsilon)}{1 + M^{(0)} \ln(z) / \varepsilon + V(z; \varepsilon)} \end{aligned}$$

We begin by showing $U(z^\varepsilon; \varepsilon) \rightarrow 0$. We can consider $U(z^\varepsilon; \varepsilon)$ in two parts:

$$U(z^\varepsilon; \varepsilon) = \left(-\frac{1 - z^{-\varepsilon}}{\varepsilon \mu^{(\varepsilon)}} \right) \cdot \left(\bar{\mathbf{Y}}(z^\varepsilon; \varepsilon) \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T \right)$$

We analyze the two multiplicands separately. Because $-\frac{1 - z^{-\varepsilon}}{\varepsilon \mu^{(\varepsilon)}} \rightarrow -\frac{\ln(z)}{\mu^{(\varepsilon)}}$ and $|\frac{\ln(z)}{\mu^{(\varepsilon)}}| < \infty$ for any $z > 0$, if $\bar{\mathbf{Y}}(z^\varepsilon; \varepsilon) \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T \rightarrow 0$ then $U(z^\varepsilon; \varepsilon) \rightarrow 0$ as well.

Expanding $\bar{\mathbf{Y}}(z^\varepsilon; \varepsilon) \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T$, we have 3 distinct terms:

$$\begin{aligned} &\bar{\mathbf{Y}}(z^\varepsilon; \varepsilon) \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T \\ &= \left(1 - z^{-\varepsilon} \right) \hat{\mathbf{q}}(z^\varepsilon; \varepsilon) \left(\left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) \mathbf{R} \right)^2 \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T \quad (11) \end{aligned}$$

$$- \mu^{(\varepsilon)} (1 - z^\varepsilon) \pi_0^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T \quad (12)$$

$$- \mu^{(\varepsilon)} \pi_0^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T \quad (13)$$

Because the family of IMAP queues has a light-tailed phase process, Properties 1a, 1b, and 1c of Definition 6 guarantee that each of Eqs. (11) to (13), respectively, are negligible as $\varepsilon \rightarrow 0$.

Thus, $U(z^\varepsilon; \varepsilon) \rightarrow 0$ as $\varepsilon \rightarrow 0$.

Next, let us show $V(z^\varepsilon; \varepsilon) \rightarrow 0$. It is useful to reformulate $V(z^\varepsilon; \varepsilon)$ by adding and subtracting $\varepsilon^{-1} (1 - z^{-\varepsilon})$ and then carefully rearranging terms.

$$\begin{aligned} V(z^\varepsilon; \varepsilon) &= \varepsilon^{-1} (1 - z^{-\varepsilon}) \left(-z^\varepsilon \frac{\bar{\lambda}^{(\varepsilon)}}{\mu^{(\varepsilon)}} \right) \\ &+ \varepsilon^{-1} (1 - z^{-\varepsilon}) \left(\left(\mu^{(\varepsilon)} \right)^{-1} \pi^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I}^{(\varepsilon)} \right) \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T \right) \\ &- M^{(0)} \ln(z) \\ &= \varepsilon^{-1} (1 - z^{-\varepsilon}) \left(1 - 1 - z^\varepsilon \frac{\bar{\lambda}^{(\varepsilon)}}{\mu^{(\varepsilon)}} \right) \end{aligned}$$

$$\begin{aligned}
& + \varepsilon^{-1} (1 - z^{-\varepsilon}) \left(-1 + \left(\mu^{(\varepsilon)} \right)^{-1} \boldsymbol{\pi}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T \right) \\
& - M^{(0)} \ln(z) \\
& = \varepsilon^{-1} (1 - z^{-\varepsilon}) \left(1 - z^\varepsilon \frac{\bar{\lambda}^{(\varepsilon)}}{\mu^{(\varepsilon)}} \right) \tag{14}
\end{aligned}$$

$$\begin{aligned}
& + \varepsilon^{-1} (1 - z^{-\varepsilon}) \left(-1 + \left(\mu^{(\varepsilon)} \right)^{-1} \boldsymbol{\pi}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T \right) \\
& - M^{(0)} \ln(z) \tag{15}
\end{aligned}$$

$$- M^{(0)} \ln(z) \tag{16}$$

Let us consider Eq. (14) separately from the other terms of $V(z^\varepsilon; \varepsilon)$, Eqs. (15) and (16).

Beginning with Eq. (14), we note that $\varepsilon^{-1} (1 - z^{-\varepsilon}) \rightarrow -\ln z$ (where $|\ln z| < \infty$ for all $z \in (0, 1)$) as $\varepsilon \rightarrow 0$. Next, using the fact that $\frac{\bar{\lambda}^{(\varepsilon)}}{\mu^{(\varepsilon)}} = (1 - \varepsilon)$ we can see that $\left(1 - z^\varepsilon \frac{\bar{\lambda}^{(\varepsilon)}}{\mu^{(\varepsilon)}} \right) = (1 - z^\varepsilon (1 - \varepsilon)) \rightarrow 0$ as $\varepsilon \rightarrow 0$. Thus, their product, Eq. (14), is negligible as $\varepsilon \rightarrow 0$.

Next, we move onto Eqs. (15) and (16). We again note that $\varepsilon^{-1} (1 - z^{-\varepsilon}) \rightarrow -\ln z$. By property 2 of the heavy-traffic family conditions (Definition 1), we can see that

$$-1 + \left(\mu^{(\varepsilon)} \right)^{-1} \boldsymbol{\pi}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T \rightarrow -M^{(0)}$$

Thus, Eq. (15) approaches $M^{(0)} \ln z$ in the $\varepsilon \rightarrow 0$ limit and Eqs. (15) and (16) cancel in this limit.

We therefore conclude that $V(z^\varepsilon; \varepsilon) \rightarrow 0$.

Having shown that $U(z^\varepsilon; \varepsilon), V(z^\varepsilon; \varepsilon) \rightarrow 0$, we have our desired result:

$$\hat{q}(z^\varepsilon; 0, \varepsilon) = \pi_0^{(\varepsilon)} \frac{1 + U(z^\varepsilon; \varepsilon)}{1 + M^{(0)} \ln z - V(z^\varepsilon; \varepsilon)} \rightarrow \frac{\pi_0^{(0)}}{1 + M^{(0)} \ln z} \tag{10}$$

We have thus shown that the scaled queue length process is asymptotically exponential when conditioned on $Y^{(\varepsilon)} = 0$.

5.3.2 Part 2: Asymptotic Exponentiality Conditioned on State i

Next, we show that the scaled queue length process is asymptotically exponential when conditioned on $Y^{(\varepsilon)} = i$ any $i \in \mathbb{N}_0$.

Using Eq. (4), for any $i \in \mathbb{N}_0$, we can express the i^{th} partial z -transform as

$$\hat{q}(z^\varepsilon; i, \varepsilon) = \hat{q}(z^\varepsilon; 0, \varepsilon) \frac{\pi_i^{(\varepsilon)}}{\pi_0^{(\varepsilon)}} \tag{17}$$

$$+ \left[\hat{q}(z^\varepsilon; 0, \varepsilon) \frac{\boldsymbol{\pi}^{(\varepsilon)}}{\pi_0^{(\varepsilon)}} \left(\mathbf{F}^{(\varepsilon)} (1 - z^\varepsilon) + \mu^{(\varepsilon)} (1 - z^{-\varepsilon}) \mathbf{I} \right) \mathbf{R}^{(\varepsilon)} \right]_i \tag{18}$$

$$+ \bar{Y}^i(z^\varepsilon; \varepsilon). \tag{19}$$

If we can show that Eqs. (18) and (19) in our expression for $\hat{q}(z^\varepsilon; i, \varepsilon)$ are asymptotically negligible then we have our desired asymptotic independence result. Let us focus on the terms separately. We begin with Eq. (18)

$$\left[\hat{q}(z^\varepsilon; 0, \varepsilon) \frac{\boldsymbol{\pi}^{(\varepsilon)}}{\pi_0^{(\varepsilon)}} \left(\mathbf{F}^{(\varepsilon)} (1 - z^\varepsilon) + \mu^{(\varepsilon)} (1 - z^{-\varepsilon}) \mathbf{I} \right) \mathbf{R}^{(\varepsilon)} \right]_i$$

$$= \frac{\hat{q}(z^\varepsilon; 0, \varepsilon)}{\pi_0^{(\varepsilon)}} \cdot \left[(1 - z^{-\varepsilon}) \pi^{(\varepsilon)} \left(\mu^{(\varepsilon)} I - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right) \mathbf{R}^{(\varepsilon)} \right]_i$$

Because $\frac{\hat{q}(z^\varepsilon; 0, \varepsilon)}{\pi_0^{(\varepsilon)}} \rightarrow \frac{1}{1 - M^{(0)} \ln z}$ and, by property 2a of our light-tailed phase process conditions, $\left[(1 - z^{-\varepsilon}) \pi^{(\varepsilon)} \left(\mu^{(\varepsilon)} I - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right) \mathbf{R}^{(\varepsilon)} \right]_i \rightarrow 0$, the second summand approaches 0.

Moving onto the third summand, we can expand $\bar{Y}^i(z^\varepsilon; \varepsilon)$ to

$$\bar{Y}^i(z^\varepsilon; \varepsilon) = \left[(1 - z^{-\varepsilon}) \hat{q}(z^\varepsilon; \varepsilon) \left((-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} I) \mathbf{R}^{(\varepsilon)} \right)^2 \right]_i \quad (20)$$

$$- \left[\mu^{(\varepsilon)} (1 - z^{-\varepsilon}) \pi_{\text{idle}}^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} I \right) \mathbf{R} \right]_i \quad (21)$$

$$- \left[\mu^{(\varepsilon)} \pi_{\text{idle}}^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} \right]_i \quad (22)$$

The light-tailed phase process conditions, specifically properties 2a, 2b, and 2c, provide that each of the three parts above go to 0 as $\varepsilon \rightarrow 0$ and therefore $\bar{Y}^i(z^\varepsilon; \varepsilon) \rightarrow 0$ as well.

With these two terms shown to be negligible, we have our desired asymptotic independence result: For each i ,

$$\lim_{\varepsilon \rightarrow 0} \hat{q}(z^\varepsilon; i, \varepsilon) = \frac{\pi_i^{(0)}}{1 + M^{(0)} \ln z} \quad (8)$$

This expression exactly states that $\varepsilon Q^{(\varepsilon)}$ and $Y^{(\varepsilon)}$ are asymptotically independent and $[\varepsilon Q^{(\varepsilon)} | Y^{(0)} = i]$ is asymptotically exponential with mean $M^{(0)}$.

5.3.3 Part 3: Asymptotic Exponentiality Without Conditioning

Because we have infinitely many modulation states i , our prior results do not immediately prove that $\varepsilon Q^{(\varepsilon)}$ is asymptotically exponential. In particular, if $\hat{q}(z^\varepsilon; i, \varepsilon)$ converges slower and slower for larger i , $\varepsilon Q^{(\varepsilon)}$ could fail to converge to the exponential distribution we'd otherwise expect it to. Thus, we show explicitly that the convergence to the exponential still occurs absent any conditioning.

We begin by expressing $\bar{q}(z^\varepsilon; \varepsilon)$ in terms of $\hat{q}(z^\varepsilon; i, \varepsilon)$ and applying our near-independence equation (Eq. (4)).

$$\begin{aligned} \bar{q}(z^\varepsilon; \varepsilon) &= \sum_{i \in \mathbb{N}_0} \hat{q}(z^\varepsilon; \varepsilon, i) = \hat{q}(z^\varepsilon; \varepsilon) e^T \\ &= \hat{q}(z^\varepsilon; 0, \varepsilon) \frac{\pi^{(\varepsilon)}}{\pi_0^{(\varepsilon)}} e^T \end{aligned} \quad (23)$$

$$+ \hat{q}(z^\varepsilon; 0, \varepsilon) \frac{\pi^{(\varepsilon)}}{\pi_0^{(\varepsilon)}} \left(\mathbf{F}^{(\varepsilon)} (1 - z^\varepsilon) + \mu^{(\varepsilon)} (1 - z^{-\varepsilon}) I \right) \mathbf{R}^{(\varepsilon)} e^T \quad (24)$$

$$+ \bar{Y}(z^\varepsilon; \varepsilon) e^T \quad (25)$$

Let us begin with Eq. (24).

$$\frac{\hat{q}(z^\varepsilon; 0, \varepsilon)}{\pi_0^{(\varepsilon)}} \cdot \pi^{(\varepsilon)} \left(\mathbf{F}^{(\varepsilon)} (1 - z^\varepsilon) + \mu^{(\varepsilon)} (1 - z^{-\varepsilon}) I \right) \mathbf{R}^{(\varepsilon)} e^T$$

We can notice that $\frac{\hat{q}(z^\varepsilon; 0, \varepsilon)}{\pi_0^{(\varepsilon)}} \rightarrow \frac{1}{1 - M^{(0)} \ln z}$ (with $|\frac{1}{1 - M^{(0)} \ln z}| < \infty$) by Eq. (8) and that the Property 1d of the light-tailed phase process (Definition 6) conditions guarantee

$$\pi^{(\varepsilon)} \left(\mathbf{F}^{(\varepsilon)} (1 - z^\varepsilon) + \mu^{(\varepsilon)} (1 - z^{-\varepsilon}) I \right) \mathbf{R}^{(\varepsilon)} e^T \rightarrow 0. \quad (26)$$

Together, these imply that this second term is asymptotically negligible.

Focusing on the Eq. (25),

$$\bar{\mathbf{Y}}(z^\varepsilon; \varepsilon) e^T = (1 - z^{-\varepsilon}) \hat{\mathbf{q}}(z^\varepsilon; \varepsilon) \left((-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I}) \mathbf{R}^{(\varepsilon)} \right)^2 e^T \quad (27)$$

$$- \mu^{(\varepsilon)} (1 - z^{-\varepsilon}) \pi_{\text{idle}}^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} (-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I}) \mathbf{R}^{(\varepsilon)} e^T \quad (28)$$

$$- \mu^{(\varepsilon)} \pi_{\text{idle}}^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} e^T \quad (29)$$

Each of Eqs. (27) to (29) approach to 0 as $\varepsilon \rightarrow 0$ by properties 1e, 1f, and 1g of Definition 6. Thus, Eq. (24), like Eq. (25), is asymptotically negligible and we need only consider the asymptotic behavior of Eq. (23).

In particular, because $\pi^{(\varepsilon)} e^T = 1$ for all ε ,

$$\hat{\mathbf{q}}(z^\varepsilon; 0, \varepsilon) \frac{\pi^{(\varepsilon)} e^T}{\pi_0^{(\varepsilon)}} = \frac{\hat{\mathbf{q}}(z^\varepsilon; 0, \varepsilon)}{\pi_0^{(\varepsilon)}}$$

and therefore

$$\lim_{\varepsilon \rightarrow 0} \bar{q}(z^\varepsilon; \varepsilon) = \lim_{\varepsilon \rightarrow 0} \frac{\hat{\mathbf{q}}(z^\varepsilon; 0, \varepsilon)}{\pi_0^{(\varepsilon)}} = \frac{1}{1 + M^{(0)} \ln z} \quad (8)$$

With Eq. (8) proven, we now have completed our proof for Theorem 1: that εQ is asymptotically exponentially distributed and is asymptotically independent of Y . $\square$

5.4 Birth-Death Modulation Process Fundamental Bounds

Before we explicitly consider the asymptotic behavior of any birth-death families, we establish a few fundamental bounds on the elements in $\mathbf{R}^{(\varepsilon)}$ and $\pi^{(\varepsilon)}$ in Lemma 3. These form the backbone of Lemma 4 and, through it, Theorem 2.

Lemma 3 *A birth-death family that meets our heavy traffic conditions and is uniformly stable and rate-bounded has the following bounds for some C_1, C_2, C_3 :*

1. $\pi_n^{(\varepsilon)} \leq C_1 a^n$ for all $n \in \mathbb{N}, \varepsilon \in \mathcal{E}$
2. $r_{i,j}^{(\varepsilon)} \leq C_2 \tilde{b}^{(j-i)^+}$ where $\tilde{b} := \frac{b}{1-b} < 1$ for all $i, j \in \mathbb{N}_0, \varepsilon \in \mathcal{E}$
3. $\left| [\mathbf{R}^{(\varepsilon)} e^T]_i \right| \leq C_3 i$ for all $i \in \mathbb{N}_0, \varepsilon \in \mathcal{E}$

where a, b are defined in the uniform stability definition (Definition 4).

Proof Let us prove the three bounds in order.

1. $\pi_n^{(\varepsilon)} \leq C_1 a^n$.

The bounded rate conditions (Definition 2) guarantee that there is some $a \in (0, 1)$ and N such that for all $n \geq N$, $\frac{\alpha_{n,n+1}^{(\varepsilon)}}{\alpha_{n+1,n}^{(\varepsilon)}} \leq a$ and therefore

$\pi_n^{(\varepsilon)} = \pi_N^{(\varepsilon)} \prod_{m=N}^n \frac{\alpha_{m,n+1}^{(\varepsilon)}}{\alpha_{m+1,m}^{(\varepsilon)}} \leq \pi_N^{(\varepsilon)} a^{n-N}$. Letting $C_1 = a^{-N}$, this bound trivially holds for all $n \leq N$ and we have our desired bound for all n .

This bound also holds for $\pi_{n,0}^{(\varepsilon)}$ and $\hat{q}(z; n)$ as $\pi_{n,0}^{(\varepsilon)} \leq \pi_n^{(\varepsilon)}$ and $\hat{q}(z; n) \leq \pi_n^{(\varepsilon)}$.

$$2. \ r_{i,j}^{(\varepsilon)} \leq C_2 \tilde{b}^{(j-i)^+}.$$

It is useful to consider this bound in three cases: when $i = j$, when $i \geq j$, and when $i < j$.

To simplify notation we let $\mathbb{P}(x_1 \rightarrow x_2 \rightarrow \dots \rightarrow x_n / y_1, y_2, \dots, y_m)$ be the probability that the phase process begins in state x_1 and transitions (over any number of steps) to state $x_2 \dots$ and then finally to state x_n without entering states $y_1, y_2, \dots, y_m$.

(a) Case 1: $i = j$.

Starting in state i there is some probability, p_i , that the process reaches state 0 before returning to state i . The number of visits to state i is distributed as a shifted geometric variable, $K_{i,i} \sim \text{Geom}(p_i)$ (where $K_{i,i} \in \{1, 2, \dots\}$). Thus if we can uniformly lower bound p_i for all i and ε then we can uniformly upper bound $r_{i,i}^{(\varepsilon)}$.

Let us start with the case when $i \leq N$. In this case we can naively bound the probability of moving from state i to state 0 without returning to i by the probability of transitioning downward i times in a row: $p_{i \downarrow 0} := \prod_{m=1}^i \frac{\alpha_{m,m-1}}{\alpha_{m,m-1} + \alpha_{m,m+1}}$.

When $i > N$ we can break the process of moving from i to 0 into two components: moving from i to N without returning to i and then moving from N to 0 without returning to i . We already have a lower bound on the probability of transitioning from N to 0 without returning to state i , $p_{N \downarrow 0} = \prod_{m=1}^N \frac{\alpha_{m,m-1}^{(\varepsilon)}}{\alpha_{m,m-1}^{(\varepsilon)} + \alpha_{m,m+1}^{(\varepsilon)}}$, so we only need to bound the probability $\mathbb{P}(i \rightarrow N / i)$.

To move from i to N the process has to first transition downward from i to $i - 1$, which occurs with probability at least $1 - b$ by our uniform stability conditions. From there we can think of the process as akin to a gambler's ruin which either reaches state i or state N first.

Consider a biased gambler's ruin on $\{N, N + 1, \dots, i\}$ with “success” probability b starting in state $i - 1$. The probability the gambler runs out of money (i.e., reaches state N before state i) is $\frac{1 - (\frac{b}{1-b})^1}{1 - (\frac{b}{1-b})^{i-N}}$. We can uniformly upper bound this probability for any i by $1 - \frac{b}{1-b} > 0$.

Because the probability of transitioning upward for any state $i \geq N$ is bounded above by b , a coupling argument shows that the probability of the process reaching state i before state N from state $i - 1$ has the same upper bound of $1 - \frac{b}{1-b}$.

Letting

$$\underline{p} := (1 - b) \left(1 - \frac{b}{1 - b}\right) \prod_{m=1}^N \left(\frac{\alpha_{m,m-1}}{\alpha_{m,m-1} + \alpha_{m,m+1}}\right)$$

we have for any state i , $p_i \geq \underline{p} > 0$ and therefore

$$r_{i,i} \leq T_{\max} \frac{1 - \underline{p}}{\underline{p}}$$

where $T_{\max}$ is the maximum expected time in state for any state as defined in Definition 2.

(b) Case 2: $i > j$.

Next we consider the case when $i > j$. Because the phase process is Markovian, if the process enters state j after starting in state i then the expected time in state j before reaching state 0 is the same as if the process had started in state j . With probability 1, the phase process started in state i eventually reaches state 0 so, as a birth-death process, it must pass through state j at least once on the way to 0. Therefore,

$$r_{i,j} = r_{j,j} \leq T_{\max} \frac{1 - \underline{p}}{\underline{p}}.$$

(c) Case 3: $i < j$.

Lastly we consider the case when $i < j$. In this case, the phase process might reach 0 without ever going through state j . If we can bound the probability that the process started in i visits j before 0 then we can bound $r_{i,j}$.

Consider the probability in two cases: from state i to state j the process either transitions through state N or not:

$$\mathbb{P}(i \rightarrow j/0) = \mathbb{P}(i \rightarrow j/0, N) + \mathbb{P}(i \rightarrow N \rightarrow j/0)$$

We can analyze these two possibilities separately. The probability the process moves from i to j without reaching state N is bounded by $(\frac{b}{1-b})^{j-i}$ by coupling with a biased gambler's ruin.

Next consider the probability of touching state N but not state 0 on the way to state j from state i . We can immediately upper bound this by the probability of simply moving from state N to state j from state N .

$$\mathbb{P}(i \rightarrow N \rightarrow j/0) \leq \mathbb{P}(N \rightarrow j/0)$$

We can also notice that the probability that the system reaches j before 0 after starting in N with some number of returns is the same as the probability that, given this is the last time the system is in N , the system reaches state j before 0. Thus

$$\mathbb{P}(N \rightarrow j/0) = \frac{\mathbb{P}(N \rightarrow j/0, N)}{\mathbb{P}(N \rightarrow j/0, N) + \mathbb{P}(N \rightarrow 0/j, N)} \leq \frac{\mathbb{P}(N \rightarrow j/0, N)}{\mathbb{P}(N \rightarrow 0/j, N)}$$

We thus have a bound in terms of two easier to analyze quantities: $\mathbb{P}(N \rightarrow j/0, N)$ and $\mathbb{P}(N \rightarrow 0/j, N)$. We have an existing lower bound on $\mathbb{P}(N \rightarrow$

$0/j, N)$ in $\underline{p} = \prod_{i=1}^N \frac{\alpha_{i,i-1}}{\alpha_{i,i-1} + \alpha_{i,i+1}}$. We can also use our gambler's ruin technique again to upper bound $\mathbb{P}(N \rightarrow j/0, N)$. Once the process moves from state N to $N+1$ (with probability no more than b), we can couple the process with a gambler's ruin on $\{N, N+1, N+2, \dots, j\}$. The probability the gambler's ruin reaches state j can be upper bounded by $(\frac{b}{1-b})^{j-N-1}$.

Taken together we have a bound on the probability of reaching state j ,

$$\mathbb{P}(i \rightarrow j/0) \leq \left(\frac{b}{1-b}\right)^{j-i} + \frac{1}{\underline{p}} \left(\frac{b}{1-b}\right)^{j-N-1},$$

and can find our final bound

$$\begin{aligned} r_{i,j}^{(\varepsilon)} &\leq T_{max} \frac{1-\underline{p}}{\underline{p}} \left(\left(\frac{b}{1-b}\right)^{j-i} + \frac{1}{\underline{p}} \left(\frac{b}{1-b}\right)^{j-N-1} \right) \\ &= T_{max} \frac{1-\underline{p}}{\underline{p}} \left(1 + \frac{1}{\underline{p}} \left(\frac{b}{1-b}\right)^{i-N-1} \right) \left(\frac{b}{1-b}\right)^{j-i} \\ &\leq T_{max} \frac{1-\underline{p}}{\underline{p}} \left(1 + \frac{1}{\underline{p}} \left(\frac{b}{1-b}\right)^{-N-1} \right) \left(\frac{b}{1-b}\right)^{j-i} \end{aligned}$$

Taking these three cases together we have our desired bound that for any i, j ,

$$r_{i,j}^{(\varepsilon)} \leq C_2 \left(\frac{p}{1-p}\right)^{(j-i)^+},$$

where $C_2 := T_{max} \frac{1-\underline{p}}{\underline{p}} \left(1 + \frac{1}{\underline{p}} \left(\frac{b}{1-b}\right)^{-N-1} \right) \geq T_{max} \left(\frac{b}{1-b}\right)$.

$$3. \left| [\mathbf{R}^{(\varepsilon)} e^T]_i \right| \leq C_3 i.$$

$$\left| [\mathbf{R}^{(\varepsilon)} e^T]_i \right| = \left| \sum_{j=0}^{\infty} r_{i,j}^{(\varepsilon)} \right| \leq \left| \sum_{j=0}^{i-1} r_{i,j}^{(\varepsilon)} \right| + \left| \sum_{j=i}^{\infty} r_{i,j}^{(\varepsilon)} \right|$$

Using property 2 we can bound $r_{i,j}^{(\varepsilon)}$ to find that

$$\left| [\mathbf{R}^{(\varepsilon)} e^T]_i \right| \leq C_2 i + C_2 \sum_{j=i}^{\infty} a^{j-i} = C_2 i + C_2 \frac{1}{1-a} \leq C_3 i,$$

where $C_3 := C_2 \left(1 + \frac{1}{1-a} \right)$.

We thus have our desired bounds on $\pi_n^{(\varepsilon)}$, $r_{i,j}^{(\varepsilon)}$, and $\left| [\mathbf{R}^{(\varepsilon)} e^T]_i \right|$. □

5.5 Birth-Death and Light-Tailed Modulation Processes

With the bounds in Lemma 3 proven, we are ready to prove in Lemma 4 that the uniform stability and rate bounded conditions imply the light-tailed phase process conditions when we restrict our attention to birth-death families meeting our heavy traffic conditions.

Lemma 4 *A birth-death heavy-traffic family which is uniformly stable and rate bounded satisfies the light-tailed phase process conditions.*

We provide a full proof in Appendix B.3. We first prove extensions of the inequalities in Lemma 3 to the quantities $\pi^{(\varepsilon)} \left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right)$ and $\mathbf{R}^{(\varepsilon)} \left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right)$. In particular we show that just like we have the bounds

$$\begin{aligned} \pi_n^{(\varepsilon)} &\leq C_1 a^n \\ r_{i,j}^{(\varepsilon)} &\leq C_2 \tilde{b}^{(j-i)^+} \\ \left| [\mathbf{R}^{(\varepsilon)} e^T]_i \right| &\leq C_3 i \end{aligned}$$

for appropriate constants C_1, C_2, C_3 defined in Lemma 3, we will also have bounds

$$\begin{aligned} \left| [\pi^{(\varepsilon)} \left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right)]_j \right| &\leq \tilde{C}_1 a^j \\ \left| [\mathbf{R}^{(\varepsilon)} \left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right)]_{i,j} \right| &\leq \tilde{C}_2 \tilde{b}^{(j-i)^+} \\ \left| [\mathbf{R}^{(\varepsilon)} \left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right) e^T]_i \right| &\leq \tilde{C}_3 i \end{aligned}$$

for appropriate $\tilde{C}_1, \tilde{C}_2, \tilde{C}_3$ defined in Appendix B.3.

We then use the fundamental bounds and our new derived bounds to show that each of the necessary quantities in the light-tailed phase process definition are asymptotically negligible.

5.6 Asymptotic Independence and Exponentiality for Birth-death Modulation Processes

With Lemma 4 proven, we can directly apply Theorem 1 to our birth-death heavy-traffic families.

Theorem 2 *In a family of IMAP queues with birth-death phase processes that meets our heavy traffic conditions, if the family is uniformly stable and rate bounded then it is also light-tailed. Consequently $\varepsilon Q^{(\varepsilon)}$ is asymptotically exponential with mean $M^{(0)}$ (where $M^{(0)}$ is defined in Section 3.3) and asymptotically independent from $Y^{(\varepsilon)}$ in the $\varepsilon \rightarrow 0$ limit.*

Equivalently,

$$\lim_{\varepsilon \rightarrow 0} \bar{q}(z; \varepsilon) = \frac{1}{1 + M^{(0)} \ln z} \quad (8)$$

$$\lim_{\varepsilon \rightarrow 0} \hat{q}(z; i, \varepsilon) = \frac{\pi_i^{(0)}}{1 + M^{(0)} \ln z} \quad (9)$$

Proof Lemma 4 proves that any birth-death convergent heavy-traffic family that is uniformly stable and rate bounded satisfies the light-tailed phase process conditions. Theorem 1 provides any heavy-traffic family which satisfies the light-tailed phase process conditions has the desired asymptotic independence and asymptotic exponential distribution. Thus, the uniform stability and rate bounded conditions provide our desired heavy-traffic results. $\square$

5.7 Hybrid-Abandonment Systems

Finally, we are able to return to our original motivating system, a tandem queue with hybrid-abandonment. Consider a sequence of hybrid-abandonment systems with fixed service rates μ_1, μ_2 and abandonment rate θ , but changing arrival rate λ . In particular, we consider the sequence of systems as λ increases toward some $\lambda^{(0)}$ such that as $\lambda \rightarrow \lambda^{(0)}, \varepsilon \rightarrow 0$.

Lemma 5 *A family of hybrid-abandonment queues with changing $\lambda \in (\lambda_{\min}, \lambda^{(0)})$, where $\varepsilon \rightarrow 0$ as $\lambda \rightarrow \lambda^{(0)}$, is a family of IMAP queues that meets our heavy-traffic condition and is birth-death, rate bounded, and uniformly stable.*

Proof We prove this lemma in two parts. In Section 5.7.1, we show that the family of hybrid-abandonment queues meets our heavy-traffic conditions. In Section 5.7.2, we show that it is also birth-death, rate bounded, and uniformly stable.

5.7.1 Part 1: Heavy-Traffic Conditions

We first prove that the sequence of hybrid-abandonment queues is a family of IMAP queues that meets our heavy traffic conditions.

Because μ_1, μ_2 , and θ are fixed, $\mathbf{F}^{(\varepsilon)}$ and $\mu^{(\varepsilon)}$ are fixed. Thus checking our heavy traffic conditions simplifies to verifying that $\mathbf{R}^{(\varepsilon)}$ and $\boldsymbol{\pi}^{(\varepsilon)}$ converge and that $M(z; \varepsilon) \rightarrow M^{(0)}$ for all $z \in (0, 1)$.

We begin by establishing bounds on $\pi_n^{(0)} - \pi_n^{(\varepsilon)}$ and $r_{i,j}^{(0)} - r_{i,j}^{(\varepsilon)}$ necessary to prove the heavy-traffic conditions.

1. Bounding $|\pi_n^{(0)} - \pi_n^{(\varepsilon)}|$.

We first show that

$$|\pi_n^{(0)} - \pi_n^{(\varepsilon)}| \leq \hat{C}_1 \varepsilon e^{-n}. \quad (30)$$

We begin by establishing a relationship between ε and $\pi_0^{(\varepsilon)}$, the long-run probability the second queue and first queue are empty, respectively. Notice that in any stable system the arrival rate at which jobs arrive to and leave the second queue are equal, i.e.,

$$\mu_1(1 - \pi_0^{(\varepsilon)}) = \mu_2(1 - \varepsilon) \implies 1 - \varepsilon = (1 - \pi_0^{(\varepsilon)}) \mu_1 / \mu_2 \implies \pi_0^{(\varepsilon)} = 1 - \frac{\mu_2}{\mu_1}(1 - \varepsilon).$$

Next, because the phase process is a birth-death process, for any n, ε

$$\pi_n^{(\varepsilon)} = \pi_0^{(\varepsilon)} \prod_{i=1}^n \frac{(\lambda^{(\varepsilon)})}{(\mu_1 + (i-1)\theta)}.$$

Thus

$$\begin{aligned} |\pi_n^{(0)} - \pi_n^{(\varepsilon)}| &= \left| \pi_0^{(0)} \frac{(\lambda^{(0)})^n}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)} - \pi_0^{(\varepsilon)} \frac{(\lambda^{(\varepsilon)})^n}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)} \right| \\ &= \left| \pi_0^{(0)} \frac{(\lambda^{(0)})^n}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)} - \pi_0^{(\varepsilon)} \frac{(\lambda^{(\varepsilon)})^n + (\lambda^{(0)})^n - (\lambda^{(0)})^n}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)} \right| \\ &= \left| (\pi_0^{(0)} - \pi_0^{(\varepsilon)}) \frac{(\lambda^{(0)})^n}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)} - \pi_0^{(\varepsilon)} \frac{(\lambda^{(\varepsilon)})^n - (\lambda^{(0)})^n}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)} \right| \\ &\leq \left| (\pi_0^{(0)} - \pi_0^{(\varepsilon)}) \frac{(\lambda^{(0)})^n}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)} \right| \end{aligned} \quad (31)$$

$$+ \left| \pi_0^{(\varepsilon)} \frac{(\lambda^{(\varepsilon)})^n - (\lambda^{(0)})^n}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)} \right| \quad (32)$$

We can consider Eq. (31) and Eq. (32) separately. Let's begin with Eq. (31).

$$\begin{aligned} \left| (\pi_0^{(0)} - \pi_0^{(\varepsilon)}) \frac{(\lambda^{(0)})^n}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)} \right| &= \left| -\varepsilon \frac{\mu_2}{\mu_1} \frac{(\lambda^{(0)})^n}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)} \right| \\ &= \varepsilon \frac{\mu_2}{\mu_1} \frac{(\lambda^{(0)})^n}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)} \end{aligned}$$

Let $\hat{n} = 1 + \frac{\varepsilon \lambda^{(0)}}{\theta}$. For all $n \geq \hat{n}$, $\frac{\lambda^{(0)}}{(\mu_1 + (n-1)\theta)} \leq e^{-1}$. Letting $A := \frac{\mu_2}{\mu_1} \sup_{n \leq \hat{n}} e^n \frac{(\lambda^{(0)})^n}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)}$, we have that

$$\left| (\pi_0^{(0)} - \pi_0^{(\varepsilon)}) \frac{(\lambda^{(0)})^n}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)} \right| = \varepsilon \frac{\mu_2}{\mu_1} \frac{(\lambda^{(0)})^n}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)} \leq A \varepsilon e^{-n}$$

Next we turn to Eq. (32). By factoring $(\lambda^{(0)} - \lambda^{(\varepsilon)})$ from $((\lambda^{(0)})^n - (\lambda^{(\varepsilon)})^n)$ we can perform a similar analysis,

$$\begin{aligned} \left| \pi_0^{(\varepsilon)} \frac{(\lambda^{(\varepsilon)})^n - (\lambda^{(0)})^n}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)} \right| &= \pi_0^{(\varepsilon)} \frac{(\lambda^{(0)})^n - (\lambda^{(\varepsilon)})^n}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)} \\ &= \pi_0^{(\varepsilon)} \frac{(\lambda^{(0)} - \lambda^{(\varepsilon)}) \sum_{k=0}^{n-1} (\lambda^{(0)})^k (\lambda^{(\varepsilon)})^{n-1-k}}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)} \\ &\leq \frac{(\lambda^{(0)} - \lambda^{(\varepsilon)}) n (\lambda^{(0)})^{n-1}}{\prod_{i=1}^n (\mu_1 + (i-1)\theta)} \end{aligned}$$

Because $\frac{(n+1)(\lambda^{(0)})^n}{\prod_{i=1}^{n+1}(\mu_1+(i-1)\theta)} / \frac{n(\lambda^{(0)})^{n-1}}{\prod_{i=1}^n(\mu_1+(i-1)\theta)} = \binom{n+1}{n} \frac{\lambda^{(0)}}{\mu_1+n\theta} \rightarrow 0$, there exists a $\hat{m}$ such that for all $n \geq \hat{m}$,

$$\frac{(n+1)(\lambda^{(0)})^n}{\prod_{i=1}^{n+1}(\mu_1+(i-1)\theta)} / \frac{n(\lambda^{(0)})^{n-1}}{\prod_{i=1}^n(\mu_1+(i-1)\theta)} \leq e^{-1}.$$

Letting $B := \frac{\mu_2}{\mu_1} \sup_{n \leq \hat{m}} e^n \frac{n(\lambda^{(0)})^{n-1}}{\prod_{i=1}^n(\mu_1+(i-1)\theta)}$ we have

$$\left| \pi_0^{(\varepsilon)} \frac{(\lambda^{(\varepsilon)})^n - (\lambda^{(0)})^n}{\prod_{i=1}^n(\mu_1+(i-1)\theta)} \right| \leq \frac{(\lambda^{(0)} - \lambda^{(\varepsilon)})n(\lambda^{(0)})^{n-1}}{\prod_{i=1}^n(\mu_1+(i-1)\theta)} \leq B\varepsilon e^{-n}$$

Using our bounds on Eq. (31) and Eq. (32), we now have our desired bound,

$$\begin{aligned} \left| \pi_n^{(\varepsilon)} - \pi_n^{(0)} \right| &\leq \left| (\pi_0^{(0)} - \pi_0^{(\varepsilon)}) \frac{(\lambda^{(0)})^n}{\prod_{i=1}^n(\mu_1+(i-1)\theta)} \right| + \left| \pi_0^{(\varepsilon)} \frac{(\lambda^{(\varepsilon)})^n - (\lambda^{(0)})^n}{\prod_{i=1}^n(\mu_1+(i-1)\theta)} \right| \\ &\leq A\varepsilon e^{-n} + B\varepsilon e^{-n} = \hat{C}_1 \varepsilon e^{-n} \end{aligned}$$

where $\hat{C}_1 := A + B$.

2. Bounds on $|r_{i,j}^{(0)} - r_{i,j}^{(\varepsilon)}|$.

With our bounds on $|\pi_n^{(\varepsilon)} - \pi_n^{(0)}|$ proven, we next show that

$$|r_{i,j}^{(0)} - r_{i,j}^{(\varepsilon)}| \leq \hat{C}_2 (\lambda^{(0)} - \lambda^{(\varepsilon)}). \quad (33)$$

Let $T_{i,j}$ be the random amount of time that the phase process spends in state j between starting in state i and reaching state 0. Note that $r_{i,j} = \mathbb{E}[T_{i,j}]$.

Consider a coupled pair of systems: one, $Y^{(0)}$, with arrival rate $\lambda^{(0)}$ and one, $Y^{(\varepsilon)}$, with rate $\lambda^{(\varepsilon)}$ for some $\varepsilon > 0$.

When $(Y^{(0)}, Y^{(\varepsilon)}) = (n, n)$, the pair jointly transition upward at rate $\lambda^{(\varepsilon)}$. The pair jointly transition downward at rate $\mu_1 + (n-1)\theta$. With rate $\lambda^{(0)} - \lambda^{(\varepsilon)}$, $Y^{(0)}$ transitions upward and $Y^{(\varepsilon)}$ doesn't transition. When $Y^{(0)} \neq Y^{(\varepsilon)}$, the two evolve independently. This coupling ensures that $Y^{(0)} \geq Y^{(\varepsilon)}$ almost surely.

Consider two possibilities: if the systems do not diverge before $Y^{(0)}$ reaches state 0 then the two have moved in lockstep the entire duration and have spent exactly the same time in state j . If the two processes do diverge before this point then the two processes may spend different times in state j , thus

$$\begin{aligned} r_{i,j}^{(0)} - r_{i,j}^{(\varepsilon)} &= \mathbb{E}[T_{i,j}^{(0)} - T_{i,j}^{(\varepsilon)}] = \mathbb{E}[T_{i,j}^{(0)} - T_{i,j}^{(\varepsilon)} \mid Y^{(0)}, Y^{(\varepsilon)} \text{ diverge before } Y^{(0)} = 0] \\ &\quad \cdot \mathbb{P}(Y^{(0)}, Y^{(\varepsilon)} \text{ diverge before } Y^{(0)} = 0) \end{aligned}$$

We bound both of these terms separately. First we consider the conditional mean. Let I be the (random) index where the first divergence occurs. At this

divergence, the two processes move from (I, I) to $(I, I + 1)$. Thus

$$\begin{aligned} |\mathbb{E}[T_{i,j}^{(0)} - T_{i,j}^{(\varepsilon)} \mid Y^{(0)}, Y^{(\varepsilon)} \text{ diverge before } Y^{(0)} = 0]| &= |\mathbb{E}[r_{I+1,j}^{(0)} - r_{I,j}^{(\varepsilon)}]| \\ &\leq |r_{j,j}^{(0)}| + |r_{j,j}^{(\varepsilon)}| \leq 2C_2, \end{aligned}$$

where C_2 is defined in Lemma 3.

Until a divergence occurs, the divergences occur at rate $\lambda^{(0)} - \lambda^{(\varepsilon)}$. The probability that a divergence occurs before time t is $\mathbb{P}(\text{Pois}((\lambda^{(0)} - \lambda^{(\varepsilon)})t) > 0) = 1 - \exp(-(\lambda^{(0)} - \lambda^{(\varepsilon)})t)$. Letting τ_i be the time when $Y^{(0)}$ first reaches 0,

$$\begin{aligned} \mathbb{P}(Y^{(0)}, Y^{(\varepsilon)} \text{ diverge before } Y^{(0)} = 0) &= \mathbb{E}[1 - \exp(-(\lambda^{(0)} - \lambda^{(\varepsilon)})\tau_i)] \\ &\leq (\lambda^{(0)} - \lambda^{(\varepsilon)})\mathbb{E}[\tau_i]. \end{aligned}$$

Noting that $\mathbb{E}[\tau_i]$ is the same as $\mathbf{R}^{(\varepsilon)}e^T$ in Lemma 3, we have that $\mathbb{E}[\tau_i] \leq C_3i$ where C_3 is defined in Lemma 3. Thus

$$\begin{aligned} \mathbb{P}(Y^{(0)}, Y^{(\varepsilon)} \text{ diverge before } Y^{(0)} = 0) &\leq C_3(\lambda^{(0)} - \lambda^{(\varepsilon)})i \\ \implies |r_{i,j}^{(0)} - r_{i,j}^{(\varepsilon)}| &\leq (2C_2) \cdot (C_3(\lambda^{(0)} - \lambda^{(\varepsilon)})i) = \hat{C}_3(\lambda^{(0)} - \lambda^{(\varepsilon)})i \end{aligned}$$

for $\hat{C}_3 = 2C_2C_3$.

With these bounds, we are now able to show that the hybrid-abandonment systems meet our heavy-traffic condition (Definition 1). We verify the conditions here. We begin by showing that each of $\boldsymbol{\pi}$, $\mathbf{R}$, $\mathbf{F}$, and μ_2 have the desired convergence properties and then show that $M(z^\varepsilon; \varepsilon) \rightarrow M^{(0)}$

1. As $\varepsilon \rightarrow 0$ or, equivalently, $\lambda^{(\varepsilon)} \rightarrow \lambda^{(0)}$,
 - (a) For any n , $|\pi_n^{(\varepsilon)} - \pi_n^{(0)}| \leq \hat{C}_1\varepsilon e^{-n}$ and thus $\pi_n^{(\varepsilon)} \rightarrow \pi_n^{(0)}$ where $\boldsymbol{\pi}^{(0)}$ is a valid distribution.
 - (b) For any i, j , $|r_{i,j}^{(0)} - r_{i,j}^{(\varepsilon)}| \leq \hat{C}_2(\lambda^{(0)} - \lambda^{(\varepsilon)})$ so $r_{i,j}^{(0)} \rightarrow r_{i,j}^{(\varepsilon)}$.
 - (c) $\mathbf{F}^{(\varepsilon)}$ is constant across all ε so trivially $\mathbf{F}^{(\varepsilon)}$ is convergent as $\varepsilon \rightarrow 0$.
 - (d) $\mu_2^{(\varepsilon)}$ is constant across all ε so trivially $\mu_2^{(\varepsilon)}$ is convergent as $\varepsilon \rightarrow 0$.
2. As $\varepsilon \rightarrow 0$,

$$\begin{aligned} 1 - (\mu_2^{(\varepsilon)})^{-1}\boldsymbol{\pi}^{(\varepsilon)}(\mu^{(\varepsilon)}\mathbf{I} - z^\varepsilon\mathbf{F}^{(\varepsilon)})\mathbf{R}^{(\varepsilon)}(\mu^{(\varepsilon)}\mathbf{I} - z^\varepsilon\mathbf{F}^{(\varepsilon)})e^T \\ \rightarrow 1 - (\mu_2^{(0)})^{-1}\boldsymbol{\pi}^{(0)}(\mu^{(0)}\mathbf{I} - \mathbf{F}^{(0)})\mathbf{R}^{(0)}(\mu^{(0)}\mathbf{I} - \mathbf{F}^{(0)})e^T \end{aligned}$$

We can decompose $\boldsymbol{\pi}^{(\varepsilon)}(\mathbf{I} - z^\varepsilon\mathbf{F}^{(\varepsilon)})\mathbf{R}^{(\varepsilon)}(\mathbf{I} - z^\varepsilon\mathbf{F}^{(\varepsilon)})e^T$ to isolate the effect of z^ε , $\mathbf{R}^{(\varepsilon)}$, and $\boldsymbol{\pi}^{(\varepsilon)}$. Because $\mathbf{F}^{(\varepsilon)}$ does not depend on ε in this case, we omit the dependence on ε . Similarly we replace $\mu^{(\varepsilon)}$ by μ_2 for clarity because it does not depend on ε .

$$\boldsymbol{\pi}^{(\varepsilon)}(\mu_2\mathbf{I} - z^\varepsilon\mathbf{F})\mathbf{R}^{(\varepsilon)}(\mu_2\mathbf{I} - z^\varepsilon\mathbf{F})e^T \tag{34}$$

$$\begin{aligned}
&= \boldsymbol{\pi}^{(0)}(\mu_2 \mathbf{I} - \mathbf{F}) \mathbf{R}^{(0)}(\mu_2 \mathbf{I} - \mathbf{F}) e^T \\
&+ (\boldsymbol{\pi}^{(\varepsilon)} - \boldsymbol{\pi}^{(0)})(\mu_2 \mathbf{I} - \mathbf{F}) \mathbf{R}^{(0)}(\mu_2 \mathbf{I} - \mathbf{F}) e^T
\end{aligned} \tag{35}$$

$$+ \boldsymbol{\pi}^{(\varepsilon)}(1 - z^\varepsilon) \mathbf{F} \mathbf{R}^{(0)}(\mu_2 \mathbf{I} - \mathbf{F}) e^T \tag{36}$$

$$+ \boldsymbol{\pi}^{(\varepsilon)}(\mu_2 \mathbf{I} - z^\varepsilon \mathbf{F})(\mathbf{R}^{(\varepsilon)} - \mathbf{R}^{(0)})(\mathbf{I} - \mathbf{F}) e^T \tag{37}$$

$$+ \boldsymbol{\pi}^{(\varepsilon)}(\mu_2 \mathbf{I} - z^\varepsilon \mathbf{F}) \mathbf{R}^{(\varepsilon)}(1 - z^\varepsilon) \mathbf{F} e^T \tag{38}$$

We show Eqs. (35) to (38) are each negligible in this limit. We extensively use the bounds derived in Lemma 3 and Lemma 4.

(a) Eq. (35) is asymptotically negligible.

First let us bound $(\boldsymbol{\pi}^{(\varepsilon)} - \boldsymbol{\pi}^{(0)})(\mu_2 \mathbf{I} - \mathbf{F})$. In particular, this has an analogous bound to its underlying quantities, $(\boldsymbol{\pi}^{(\varepsilon)} - \boldsymbol{\pi}^{(0)})$ (Eq. (30)).

$$\begin{aligned}
\left| [(\boldsymbol{\pi}^{(\varepsilon)} - \boldsymbol{\pi}^{(0)})(\mu^{(0)} \mathbf{I} - \mathbf{F})]_i \right| &= \left| \sum_k ((\pi_i^{(\varepsilon)} - \pi_i^{(0)})[\mu_2 \mathbf{I} - \mathbf{F}]_{i,k}) \right| \\
&= \left| (\pi_i^{(\varepsilon)} - \pi_i^{(0)})\mu_2 - (\pi_{i-1}^{(\varepsilon)} - \pi_{i-1}^{(0)})\beta_{i,i-1}\mathbb{I}\{i > 0\} \right| \\
&\leq \mu_2 \hat{C}_1 \varepsilon e^{-i} + \mu_1 \hat{C}_1 \varepsilon e^{-(i-1)} \\
&= \hat{C}_1(\mu_2 + e\mu_1)\varepsilon e^{-i}
\end{aligned}$$

Thus using the bound we just derived and the bound on $\mathbf{R}(\mu^{(\varepsilon)} \mathbf{I} - \mathbf{F}) e^T$ introduced in Section 5.5,

$$\begin{aligned}
&\left| (\boldsymbol{\pi}^{(\varepsilon)} - \boldsymbol{\pi}^{(0)})(\mu^{(0)} \mathbf{I} - \mathbf{F}) \mathbf{R}^{(0)}(\mu_2 \mathbf{I} - \mathbf{F}) e^T \right| \\
&= \left| \sum_i [(\boldsymbol{\pi}^{(\varepsilon)} - \boldsymbol{\pi}^{(0)})(\mu^{(0)} \mathbf{I} - \mathbf{F})]_i [\mathbf{R}(\mu_2 \mathbf{I} - \mathbf{F}) e^T]_i \right| \\
&\leq \sum_i ((\hat{C}_1(\mu_2 + e\mu_1)\varepsilon e^{-i}) (\tilde{C}_3 i)) = \varepsilon \hat{C}_1 \tilde{C}_3 (\mu_2 + e\mu_1) \sum_i i e^{-i}
\end{aligned}$$

Thus $(\boldsymbol{\pi}^{(\varepsilon)} - \boldsymbol{\pi}^{(0)})(\mu^{(0)} \mathbf{I} - \mathbf{F}) \mathbf{R}^{(0)}(\mu_2 \mathbf{I} - \mathbf{F}) e^T \rightarrow \varepsilon$ as $\varepsilon \rightarrow 0$.

(b) Eq. (36) is asymptotically negligible.

First, we note that

$$[\boldsymbol{\pi}^{(\varepsilon)} \mathbf{F}]_j = \sum_i \pi_i^{(\varepsilon)} \mathbf{F}_{i,j} = \pi_{i+1}^{(\varepsilon)} \beta_{i+1,i} \leq C_1 \beta_{i+1,i} e^{-(i+1)} = C_1 e^{-1} \beta_{i+1,i} e^{-i}.$$

Using our newly derived bound we can see that

$$|\boldsymbol{\pi}^{(\varepsilon)}(1 - z^\varepsilon) \mathbf{F} \mathbf{R}^{(0)}(\mathbf{I} - \mathbf{F}) e^T| = (1 - z^\varepsilon) \left| \sum_i \left([\boldsymbol{\pi}^{(\varepsilon)} \mathbf{F}]_i [\mathbf{R}^{(0)}(\mu_2 \mathbf{I} - \mathbf{F}) e^T]_i \right) \right|$$

$$\leq (1 - z^\varepsilon) \left| \sum_i \left((C_1 e^{-1} \beta_{i+1,i} e^{-i}) (\tilde{C}_3 i) \right) \right| = (1 - z^\varepsilon) \tilde{C}_3 C_1 e^{-1} \beta_{i+1,i} \sum_i e^{-i} i$$

Thus as $\varepsilon \rightarrow 0$, $(1 - z^\varepsilon) \rightarrow 0$ and $\boldsymbol{\pi}^{(\varepsilon)}(1 - z^\varepsilon) \mathbf{F} \mathbf{R}^{(0)}(\mu_2 \mathbf{I} - \mathbf{F}) e^T \rightarrow 0$
(c) Eq. (37) is asymptotically negligible.

$$\begin{aligned} & |\boldsymbol{\pi}^{(\varepsilon)}(\mu_2 \mathbf{I} - z^\varepsilon \mathbf{F})(\mathbf{R}^{(\varepsilon)} - \mathbf{R}^{(0)})(\mu_2 \mathbf{I} - \mathbf{F}) e^T| \\ &= |[\boldsymbol{\pi}^{(\varepsilon)}(\mu_2 \mathbf{I} - z^\varepsilon \mathbf{F})](\mathbf{R}^{(\varepsilon)} - \mathbf{R}^{(0)})[(\mu_2 \mathbf{I} - \mathbf{F}) e^T]| \\ &\leq |[\boldsymbol{\pi}^{(\varepsilon)}(\mu_2 \mathbf{I} - z^\varepsilon \mathbf{F})](\mathbf{R}^{(\varepsilon)} - \mathbf{R}^{(0)})[(\mu_2 + \mu_1) e^T]| \\ &\leq (\mu_1 + \mu_2) |[\boldsymbol{\pi}^{(\varepsilon)}(\mu_2 \mathbf{I} - z^\varepsilon \mathbf{F})](\mathbf{R}^{(\varepsilon)} - \mathbf{R}^{(0)}) e^T| \\ &= (\mu_1 + \mu_2) \sum_{i,j} ([\boldsymbol{\pi}^{(\varepsilon)}(\mu_2 \mathbf{I} - z^\varepsilon \mathbf{F})]_i (r_{i,j}^{(\varepsilon)} - r_{i,j}^{(0)})) \\ &\leq (\mu_1 + \mu_2) \sum_i (\tilde{C}_1 e^{-i} (\sum_j (r_{i,j}^{(\varepsilon)} - r_{i,j}^{(0)}))) \end{aligned}$$

We use two different bounds for $r_{i,j}^{(\varepsilon)} - r_{i,j}^{(0)}$. In particular we use our coupling bound, $|r_{i,j}^{(\varepsilon)} - r_{i,j}^{(0)}| \leq \hat{C}_2(\lambda^{(0)} - \lambda^{(\varepsilon)})i$, (33), for “small” j and our bounds on $|r_{i,j}^{(\varepsilon)}|, |r_{i,j}^{(0)}| \leq C_2 e^{-(j-i)^+}$ (property 2), for “large” j . We change our bound at $j = i + \ln((\lambda^{(0)} - \lambda^{(\varepsilon)})^{-1})$. Thus,

$$\begin{aligned} \left| \sum_j (r_{i,j}^{(\varepsilon)} - r_{i,j}^{(0)}) \right| &= \left| \sum_{j=0}^{\lfloor i - \ln(\lambda^{(0)} - \lambda^{(\varepsilon)}) \rfloor} (r_{i,j}^{(\varepsilon)} - r_{i,j}^{(0)}) + \sum_{j=1+\lfloor i - \ln(\lambda^{(0)} - \lambda^{(\varepsilon)}) \rfloor}^{\infty} (r_{i,j}^{(\varepsilon)} - r_{i,j}^{(0)}) \right| \\ &\leq \left| \sum_{j=0}^{\lfloor i - \ln(\lambda^{(0)} - \lambda^{(\varepsilon)}) \rfloor} (\hat{C}_2(\lambda^{(0)} - \lambda^{(\varepsilon)})i) \right| \\ &\quad + \left| \sum_{j=1+\lfloor i - \ln(\lambda^{(0)} - \lambda^{(\varepsilon)}) \rfloor}^{\infty} (C_2 e^{-(j-i)^+}) \right| \\ &\leq (i - \ln(\lambda^{(0)} - \lambda^{(\varepsilon)}))(\hat{C}_2(\lambda^{(0)} - \lambda^{(\varepsilon)})i) \\ &\quad + C_2(\lambda^{(0)} - \lambda^{(\varepsilon)}) \sum_{n=0}^{\infty} e^{-n} \\ &= \hat{C}_2(\lambda^{(0)} - \lambda^{(\varepsilon)})i^2 - \hat{C}_2 \ln(\lambda^{(0)} - \lambda^{(\varepsilon)})(\lambda^{(0)} - \lambda^{(\varepsilon)})i \\ &\quad + \frac{C_2}{1 - e^{-1}}(\lambda^{(0)} - \lambda^{(\varepsilon)}) \end{aligned}$$

With this bound on $|\sum_j (r_{i,j}^{(\varepsilon)} - r_{i,j}^{(0)})|$ we can analyze Eq. (37) directly.

$$\begin{aligned}
& \left| \boldsymbol{\pi}^{(\varepsilon)} (\mu_2 \mathbf{I} - z^\varepsilon \mathbf{F}) (\mathbf{R}^{(\varepsilon)} - \mathbf{R}^{(0)}) (\mu_2 \mathbf{I} - \mathbf{F}) e^T \right| \\
& \leq (\mu_1 + \mu_2) \sum_i (\tilde{C}_1 e^{-i} \sum_j (r_{i,j}^{(0)} - r_{i,j}^{(\varepsilon)})) \\
& \leq (\mu_1 + \mu_2) \sum_i (\tilde{C}_1 e^{-i} (\hat{C}_2 (\lambda^{(0)} - \lambda^{(\varepsilon)}) i^2 - \hat{C}_2 \ln(\lambda^{(0)} - \lambda^{(\varepsilon)}) (\lambda^{(0)} - \lambda^{(\varepsilon)}) i)) \\
& + (\mu_1 + \mu_2) (\lambda^{(0)} - \lambda^{(\varepsilon)}) \sum_i \left(\frac{C_2}{1 - e^{-1}} e^{-i} \right) \\
& = (\mu_1 + \mu_2) (\lambda^{(0)} - \lambda^{(\varepsilon)}) \sum_i (C_1 \hat{C}_2 i^2 e^{-i}) \\
& - (\mu_1 + \mu_2) (\lambda^{(0)} - \lambda^{(\varepsilon)}) \ln(\lambda^{(0)} - \lambda^{(\varepsilon)}) \sum_i (C_1 \hat{C}_2 i e^{-i}) \\
& + (\mu_1 + \mu_2) (\lambda^{(0)} - \lambda^{(\varepsilon)}) \sum_i \left(\frac{C_2}{1 - e^{-1}} e^{-i} \right)
\end{aligned}$$

Thus because $(\lambda^{(0)} - \lambda^{(\varepsilon)}) \rightarrow 0$ and $-(\lambda^{(0)} - \lambda^{(\varepsilon)}) \ln((\lambda^{(0)} - \lambda^{(\varepsilon)})) \rightarrow 0$ as $\varepsilon \rightarrow 0$, $\boldsymbol{\pi}^{(\varepsilon)} (\mu_2 \mathbf{I} - z^\varepsilon \mathbf{F}) (\mathbf{R}^{(\varepsilon)} - \mathbf{R}^{(0)}) (\mu_2 \mathbf{I} - \mathbf{F}) e^T \rightarrow 0$.

(d) Eq. (38) is asymptotically negligible.

$$\begin{aligned}
& \left| \boldsymbol{\pi}^{(\varepsilon)} (\mathbf{I} - z^\varepsilon \mathbf{F}) \mathbf{R}^{(\varepsilon)} (1 - z^\varepsilon) \mathbf{F} e^T \right| = (1 - z^\varepsilon) \left| \boldsymbol{\pi}^{(\varepsilon)} (\mathbf{I} - z^\varepsilon \mathbf{F}) \mathbf{R}^{(\varepsilon)} \mathbf{F} e^T \right| \\
& = (1 - z^\varepsilon) \left| [\boldsymbol{\pi}^{(\varepsilon)} (\mathbf{I} - z^\varepsilon \mathbf{F})] \mathbf{R}^{(\varepsilon)} [\mathbf{F} e^T] \right| \leq (1 - z^\varepsilon) \left| [\boldsymbol{\pi}^{(\varepsilon)} (\mathbf{I} - z^\varepsilon \mathbf{F})] \mathbf{R}^{(\varepsilon)} [\mu_1 e^T] \right| \\
& = \mu_1 (1 - z^\varepsilon) \left| [\boldsymbol{\pi}^{(\varepsilon)} (\mathbf{I} - z^\varepsilon \mathbf{F})] [\mathbf{R}^{(\varepsilon)} e^T] \right| \leq \mu_1 (1 - z^\varepsilon) \left| \sum_i ((\tilde{C}_1 e^{-i}) (C_3 i)) \right| \\
& = (1 - z^\varepsilon) \mu_1 \tilde{C}_1 C_3 \sum_i i e^{-i}
\end{aligned}$$

Thus as $\varepsilon \rightarrow 0$, $(1 - z^\varepsilon) \rightarrow 0$ and $\boldsymbol{\pi}^{(\varepsilon)} (\mathbf{I} - z^\varepsilon \mathbf{F}) \mathbf{R}^{(\varepsilon)} (1 - z^\varepsilon) \mathbf{F} e^T \rightarrow 0$.

3. For all ε , the phase process is an $M/M/1 + M$ queue with $\lambda > 0$ and $\theta > 0$. Thus the phase process is ergodic for all ε .

The hybrid-abandonment system therefore meets our heavy traffic conditions (Definition 1). Having show this we need only show that the system also meets the birth-death, rate bounded, and uniformly stable conditions.

5.7.2 Part 2: Birth-Death, Rate Bounded, and Uniformly Stable

The phase process, the first queue, is a birth-death process for all ε , so it immediately satisfies our birth-death condition (Definition 3). Thus, we need only show that the family is uniformly stable (Definition 4) and rate-bounded (Definition 2). With $a = b = e^{-1}$ and $N = \lceil \frac{\theta e}{\lambda_{\max}} \rceil + 1$,

1. $\sup_{n \geq N, \varepsilon} \frac{\alpha_{n, n+1}}{\alpha_{n, n-1}} = \sup_{n \geq N, \varepsilon} \frac{\lambda}{\mu_1 1\{n > 0\} + (n-1)\theta 1\{n > 0\}} \leq e^{-1}.$
2. $\sup_{n \geq N, \varepsilon} \frac{\alpha_{n, n+1}}{\alpha_{n+1, n}} = \sup_{n \geq N, \varepsilon} \frac{\lambda^{(\varepsilon)}}{\mu_1 + n\theta} \leq e^{-1}.$

Thus, the family satisfies all the conditions of Definition 4 and is uniformly stable.

Finally, let us show that the family is rate-bounded, with $T_{\max} = (\lambda_{\min})^{-1}$ and $\hat{\lambda} = \mu_1$.

1. $\sup_{n, \varepsilon} \left\{ \left(\sum_{m \neq n} \alpha_{n, m} \right)^{-1} \right\} = \sup_n \left\{ \left(\lambda^{(\varepsilon)} + \mu_1 1\{n > 0\} + (n-1)\theta 1\{n > 0\} \right)^{-1} \right\} = (\lambda_{\min})^{-1} < \infty.$
2. $\sup_{n, \varepsilon} \left(\lambda_n^{(\varepsilon)} + \beta_{n, n+1}^{(\varepsilon)} + \beta_{n, n-1}^{(\varepsilon)} \right) \leq \mu_1.$
3. $\mu^{(\varepsilon)} = \mu_2$ for all ε .

The family therefore meets all the requirements of Definition 2 and is rate-bounded.

Thus, the sequence of hybrid-abandonment queues is a birth-death family of IMAP queue that meets our heavy traffic conditions and is uniformly stable and rate bounded. $\square$

5.8 Asymptotic Independence and Exponentiality for Hybrid Abandonment Systems

Finally, we prove that the same asymptotic results hold for hybrid-abandonment systems as for the more general heavy-traffic families. This further narrows the variety of systems we consider but allows us to provide the most straightforward guarantees.

Theorem 3 *In a tandem queue with hybrid abandonment, as the arrival rate rises to a critical threshold and the second queue's idleness approaches 0, $\varepsilon Q^{(\varepsilon)}$ is asymptotically exponential with mean $M^{(0)}$ (where $M^{(0)}$ is defined in Definition 1) and asymptotically independent from $Y^{(\varepsilon)}$.*

Equivalently,

$$\lim_{\varepsilon \rightarrow 0} \bar{q}(z; \varepsilon) = \frac{1}{1 + M^{(0)} \ln z} \quad (8)$$

$$\lim_{\varepsilon \rightarrow 0} \hat{q}(z; i, \varepsilon) = \frac{\pi_i^{(0)}}{1 + M^{(0)} \ln z} \quad (9)$$

Proof Lemma 5 shows that the hybrid-abandonment systems are birth-death convergent heavy traffic families that are uniformly stable and rate bounded. Thus, by Theorem 2, the hybrid-abandonment systems have asymptotically exponentially scaled second queue lengths and asymptotically independent first and second queue lengths. $\square$

6 Numerical Study

Theorem 3 proves that in the load $\rho \rightarrow 1$ limit, the scaled second queue length $\varepsilon \mathbb{E}[Q_2]$ has mean $M^{(0)}$, is exponentially distributed, and is independent of the first queue length. We now use simulation to show that our results are approximately true even outside of the asymptotic regime and are a useful tool in managerial decision-making.

We investigate four questions in this section. We first investigate the accuracy of the approximation $\varepsilon\mathbb{E}[Q_2] \approx M^{(0)}$ or, equivalently, that $\mathbb{E}[Q_2] \approx \frac{M^{(0)}}{1-\rho}$ when ρ is large. Next, we investigate whether Q_1 and Q_2 are meaningfully correlated in this high ρ environment. We then investigate whether εQ_2 is approximately distributed as an exponential random variable with mean $M^{(0)}$. Finally, we investigate the usefulness of our heavy-traffic approximation in a common operations management task: selecting the optimal capacity in a queueing system.

To answer each question, we directly simulate the hybrid-abandonment tandem system and observe its behavior. In answering the first three of our four questions, we consider $\mu_1 \in \{1.1, 1.2, 1.5, 2, 4\}$, $\rho \in \{.8, .9, .95, .99\}$, and $\theta \in \{.1, .3, .5\}$. We consider each possible combination of μ_1 , θ_1 , and ρ and fix $\mu_2 = 1$ resulting in 60 distinct parameter sets.

Finally, when studying the approximation's use in optimal capacity selection, we additionally consider a holding cost parameter $c \in \{.5, .75, 1, 1.5, 2\}$ and let μ_2 be the decision variable in a simple optimization model.

Full details on the simulation and a link to the data are available in Appendix C.

6.1 Question 1: Accuracy of Heavy-Traffic Approximation

Theorem 3 states that as $\rho \rightarrow 1$, the scaled second queue length $(1 - \rho)Q_2$ converges to $M^{(0)}$. This gives rise to our heavy traffic approximation for $\mathbb{E}[Q_2]$, $\mathbb{E}[Q_2] \approx \frac{M^{(0)}}{1-\rho}$. In each of the 60 parameter sets described above, we simulate our hybrid-abandonment tandem system to find and report an estimate of $\mathbb{E}[Q_2]$ and the standard deviation of our estimator. We then compare our simulation estimate of $\mathbb{E}[Q_2]$ to our heavy-traffic estimate, $\frac{M^{(0)}}{1-\rho}$ and report the difference, $E[Q_2] - \frac{M^{(0)}}{1-\rho}$, and relative difference, $(\mathbb{E}[Q_2] - \frac{M^{(0)}}{1-\rho})/\mathbb{E}[Q_2]$.

To maintain readability, in Table 1 we aggregate our results based on load ρ . Specifically, rather than reporting the results when $(\mu_1, \rho, \theta) = (1.1, .8, .1)$, $(\mu_1, \rho, \theta) = (1.2, .8, .3)$, $(\mu_1, \rho, \theta) = (1.5, .8, .5)$, etc. separately, we average our four reported statistics over all of these parameter sets and report this final average. We choose to collapse along ρ values as our results are far more dependent on load ρ than on μ_1 or θ .

Table 1. Difference between empirical mean $\mathbb{E}[Q_2]$ and approximate mean $\frac{M^{(0)}}{1-\rho}$.

ρ	Average $\mathbb{E}[Q_2]$	Average SD of $\mathbb{E}[Q_2]$	Average Gap	Average Rel Gap
0.80	3.66	0.001	-0.758	-17.0%
0.90	8.08	0.003	-0.754	-8.5%
0.95	16.9	0.014	-0.765	-4.3%
0.99	87.6	0.330	-0.745	-0.9%

As $\rho \rightarrow 1$, we can notice two related trends. First, the average gap between the empirical mean and the approximate mean is roughly constant across all values of ρ at about -0.75 jobs. Thus, as the empirical mean rises with ρ , the relative gap shrinks from -17% to under -1%.

Based on the results presented in Table 1, we can conclude that when $\rho \geq 0.8$ our heavy-traffic approximation is consistently within 1 customer of the true mean queue length value. Even though the absolute difference between the approximation and the true mean queue length does not vanish as ρ increases from 0.8 toward 1, the relative difference between our approximation and the true mean queue length shrinks from around 17% toward 0% as ρ increases toward 1.

Observation 1 *The heavy-traffic approximation is very accurate for loads above 0.8. The absolute difference between the approximation and the true value is roughly constant across ρ values while the relative difference between the two shrinks toward 0 as ρ increases toward 1.*

6.2 Question 2: Validation of Non-correlation

Theorem 3 states that as $\rho \rightarrow 1$, the first and second queue lengths are asymptotically independent. To show that this is approximately true even outside of our asymptotic regime, we simulate the hybrid-abandonment tandem system to empirically estimate the correlation between Q_1 and Q_2 . In Table 2, we present the empirical Pearson correlation coefficient between Q_1 and Q_2 across various parameter sets. As in Section 6.1, we again aggregate our results along ρ values to maintain readability. We report the average correlation coefficient among the 15 parameter sets at each ρ value as well as the maximum and minimum correlation coefficient at each ρ value.

Table 2. Correlation between Q_1 and Q_2 .

ρ	Average Correlation	Max Correlation	Min Correlation
0.8	-0.042	-0.001	-0.088
0.9	-0.032	-0.001	-0.078
0.95	-0.020	-0.001	-0.057
0.99	-0.005	0.000	-0.017

Two trends in the data are immediately noticeable: first, across all our parameter sets, the correlation is extremely weak. Although there is no universally agreed benchmark for a “weak” correlation, a common standard is that of Cohen [29] which defines any correlation coefficient with magnitude less than 0.1 as very weak or negligible. The highest magnitude of a correlation coefficient we observe in any parameter set is 0.088 as seen in Table 2, thus the correlation is negligible in each of the 60 parameter sets studied.

Second, the magnitude of the correlation decreases as ρ increases toward 1. Both the maximum magnitude of the correlation and the average magnitude of the correlation decrease as ρ increases. Therefore, we can treat the two queue lengths as uncorrelated even outside of our asymptotic regime.

Observation 2 *At load above 0.8, the two queue lengths are not meaningfully correlated. The small amount of correlation that does exist between the two decreases in magnitude as the load increases.*

6.3 Question 3: Validation of Exponential Distribution

Theorem 3 states that as $\rho \rightarrow 1$, the scaled queue length $(1 - \rho)Q_2$ is asymptotically distributed as an exponential with mean $M^{(0)}$. To compare the empirical distribution of $(1 - \rho)Q_2$ outside of this limit with an exponential with mean $M^{(0)}$, we compute the Kolmogorov-Smirnov statistic or KS statistic, a form of “distance” between two distributions. The KS statistic between two random variables, X and Y is given by

$$KS(X, Y) = \sup_t |F_X(t) - F_Y(t)|.$$

Given Theorem 3, we would expect that as we consider ρ closer and closer to 1 we should see the KS gap between $(1 - \rho)Q_2$ and the exponential with mean $M^{(0)}$ should decrease toward 0.

In Table 3 we present the average and maximum KS statistics between the scaled empirical queue length and the exponential distribution with mean $M^{(0)}$ and the average standard deviation of the KS statistic of our different parameter sets aggregated by their ρ value.

Table 3. KS gap between $(1 - \rho)Q_2$ and Exponential with mean $M^{(0)}$.

ρ	Average KS Stat	Max KS Stat	Avg KS Stat Std
0.8	0.2000	0.2001	0.0003
0.9	0.1000	0.1003	0.0006
0.95	0.0500	0.0503	0.0008
0.99	0.0129	0.0222	0.0033

We can observe from Table 3 that the average and max KS statistics are approximately equal to $1 - \rho$. A gap of $1 - \rho$ is the minimum gap that we can achieve simply due to the mismatch between the discrete empirical distribution and the continuous exponential distribution it is converging to. In any stable system with load ρ , the system is idle $(1 - \rho)$ of the time, so $F_{(1-\rho)Q_2}(0) = F_{Q_2}(0) = 1 - \rho$. In contrast, the exponential distribution, a positive, continuous variable, has no mass at or below 0, so $F(0) = 0$. Thus, the KS gap between $(1 - \rho)Q_2$ and the exponential with mean $M^{(0)}$ is bounded below by $1 - \rho$.

To circumvent this floor on the KS gap we also compare our empirical distribution to a discrete random variable. A natural candidate is a geometric random variable with our desired mean. This would coincide with distribution of the queue length of a simple $M/M/1$ that has the same asymptotic behavior as our Q_2 .

In Table 4 we present the average and maximum KS statistics between the empirical queue length and the geometric distribution with mean $\frac{M^{(0)}}{1-\rho}$ and the average standard deviation of the KS statistic of our different parameter sets aggregated by their ρ value.

Table 4. KS gap between Q_2 and Geometric with mean $\frac{M^{(0)}}{1-\rho}$.

ρ	Average KS Stat	Max KS Stat	Avg KS Stat Std
0.8	0.0608	0.0810	0.0003
0.9	0.0314	0.0473	0.0006
0.95	0.0163	0.0277	0.0008
0.99	0.0097	0.0200	0.0033

Comparing the KS distance from $(1 - \rho)Q_2$ to the exponential with mean $M^{(0)}$ and the geometric with mean $\frac{M^{(0)}}{1-\rho}$, we see that the geometric with mean $\frac{M^{(0)}}{1-\rho}$ is a much better approximation. At all loads studied, both the average and maximum KS statistic observed are lower in the geometric case than the exponential case. This improvement is far larger at smaller values of ρ but even at the maximum value considered ($\rho = .99$), the geometric variable is a small but non-negligible improvement over the exponential in fitting the observed data.

Based on the data in Table 3 and Table 4, we can see that the exponential with mean $M^{(0)}$ is a good fit for the scaled queue length $(1 - \rho)Q_2$ only when ρ is fairly close to 1. The geometric with mean $\frac{M^{(0)}}{1-\rho}$ is a much better fit for Q_2 at all loads studied.

Observation 3 *The scaled distribution εQ_2 is well approximated by an exponential, and an even better fit is obtained from approximating Q_2 by a geometric distribution. The accuracy of both approximations increases significantly as the load of the system rises.*

6.4 Question 4: Application to Capacity Management

Capacity planning is a classic problem in operations management (see Kleinrock [30], Holt et al. [31]). A firm must decide how much capacity (such as machines in a factory, servers in a data center, tables in a restaurant, etc.) to buy before it observes its demand. Capacity is costly, so firms have an incentive to buy as little as necessary. However, buying too little means either not satisfying all demand or otherwise delivering significantly degraded service. Firms therefore have to carefully consider how much capacity they purchase. Consider a simple static capacity planning problem: A hybrid-abandonment tandem system serves customers and the manager must pick the optimal capacity μ_2 for the second stage server. The operator has to consider both the cost of the capacity and the cost of waiting customers' goodwill lost during the wait. We assume that capacity costs 1 per unit of capacity per unit of time and that goodwill cost accrued due to waiting at a rate of c per customer in the second queue per unit of time. We thus have an optimization problem of the form

$$\min_{\mu_2} C(\mu_2) = \min_{\mu_2} (c\mathbb{E}[Q_2] + \mu_2).$$

We define μ_2^* to be the optimizer of this problem.

Because we do not have an expression for a tractable exact $\mathbb{E}[Q_2]$ in terms of system primitives, any attempt to minimize the cost from first principles of optimization would require repeated simulation to estimate $E[Q_2]$ under different parameter sets. We can, however, easily compute our heavy traffic approximation for $\mathbb{E}[Q_2]$, $\frac{M^{(0)}}{1-\rho}$. This creates

an “approximation problem”,

$$\min_{\mu_2} \hat{C}(\mu_2) = \min_{\mu_2} (c \frac{M^{(0)}}{1 - \rho} + \mu_2) = \min_{\mu_2} (c \frac{M^{(0)}}{1 - \bar{\lambda}/\mu_2} + \mu_2),$$

where $\bar{\lambda} = (1 - \pi_0)\mu_1$ is the average arrival rate to the second queue. This approximation problem does have a simple analytical solution: $\hat{\mu}_2^* = \bar{\lambda} + \sqrt{\frac{\bar{\lambda}M^{(0)}}{c}}$.

We define the “excess cost” $\Delta = C(\hat{\mu}_2^*) - C(\mu_2^*)$ as the additional cost of running the system at $\mu_2 = \hat{\mu}_2^*$ rather than $\mu_2 = \mu_2^*$. Similarly, we define the “relative excess cost” as the relative increase in cost under the same scenario, $\Delta/C(\mu_2^*)$.

Table 5. Optimal and Approximate μ_2 Values and Costs

c	Avg μ_2^*	Avg ρ^*	Avg $\hat{\mu}_2^*$	Avg $C(\mu_2^*)$	Avg $\hat{C}(\hat{\mu}_2^*)$	Max Δ	Max $\Delta/C(\mu_2^*)$
0.5	0.530	0.403	0.550	0.8417	1.369	0.0067	0.55%
0.75	0.598	0.359	0.623	0.9820	1.762	0.0076	0.58%
1	0.658	0.324	0.684	1.101	2.132	0.0090	0.60%
1.5	0.756	0.283	0.787	1.302	2.832	0.0080	0.60%
2	0.838	0.254	0.874	1.472	3.501	0.0110	0.60%

Table 5 demonstrates two somewhat at-odds observations. Observing the Avg $C(\mu_2^*)$ and Avg $\hat{C}(\hat{\mu}_2^*)$ columns, we can see that the optimal cost in the approximation problem is consistently much larger than the optimal cost in the true problem. At different c values, the optimal approximation cost avg $\hat{C}(\hat{\mu}_2^*)$ is anywhere from 1.5 to 2 times larger than the optimal true cost avg $C(\mu_2^*)$. However, even though the optimal values in these two problems diverge starkly, the absolute excess cost Max Δ remains extremely small, below 0.0110. Across all of the parameter sets we consider, the relative excess cost Max $\Delta/C(\mu_2^*)$ is below 0.6% as well. This fits well with the results seen in Table 1 which showed that the difference between the empirical mean queue length and the approximate mean queue length is roughly constant for different values of ρ . As a result, the approximation problem seems very similar to a shifted version of the original problem.

Observation 4 *The heavy-traffic approximation for the mean queue length is a useful tool for capacity management even when optimal load ρ is comparatively small. Using the heavy-traffic approximation rather than the true mean queue length results in a similar optimal capacity level and extremely low excess cost, both in absolute and in relative terms.*

The approximation does seem to consistently overestimate the cost of running the system, however, meaning that it may not be broadly applicable for forecasting either average queue length or long run average running cost.

7 Conclusion

We present the first heavy-traffic analysis of the hybrid-abandonment tandem queue. We characterize the scaled second queue length by finding its asymptotic mean and showing it is asymptotically exponentially distributed and asymptotically independent of the first queue length. We do this by analyzing countably infinite Markovian arrival process (IMAP) queues, a class that includes our original hybrid-abandonment system. We provide sufficient conditions for the scaled queue length of an IMAP queue to be asymptotically exponential and asymptotically independent of its phase process and verify these conditions hold in the case of a hybrid-abandonment queue. Further, we show empirically that our heavy-traffic results are useful approximations beyond heavy traffic.

A promising area of future research is similar analysis of queues with infinite Markovian service processes (IMSPs), either in lieu of or in addition to IMAP arrivals. We believe IMSPs will be slightly harder to analyze due to unused service occurring when the queue is empty. Another promising extension is to provide an analogous bound on the visitation matrix $\mathbf{R}$ and the steady-state distribution of the phase process $\boldsymbol{\pi}$ for phase processes beyond birth-death processes. Of particular interest would be bounds for queueing networks to extend the heavy traffic analysis from two-stage systems to more complex queueing networks.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Number CMMI-2514091. We are thankful for the monetary support provided by the NSF and to the staff who made it possible.

Disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Code and Data Disclosure

The code and generated data used in the numerical experiments Section 6 are available at the author's GitHub repository [here](#).

Appendix A Notation

In Table [A1](#) we present a full list of notation used in this paper.

Table A1 Notation Used

Symbol	Meaning	Formula
Q Y	Queue Length Distribution Phase Process Distribution	
$\alpha_{i,j}$ $\mathbf{A}$ λ_i $p_{i,j}$ $\beta_{i,j}$ $\bar{\lambda}$ μ	Phase process transition rate from state i to state j Generator Matrix of Y In-state arrival Rate in Modulation State i Probability of a between-state arrival due to phase process transition from state i to state j Rate of queue arrivals caused by i to j transitions in phase process Long run average arrival rate Queue Service Rate	$[\alpha_{i,j}]_{i,j \in \mathbb{N}_0}$ $\alpha_{i,j} p_{i,j}$ $\mathbb{E}[\lambda_Y + \sum_{j \neq Y} \beta_{Y,j}]$
$\pi_{i,j}$ π_i $\boldsymbol{\pi}$ $\boldsymbol{\pi}_{\text{idle}}$ $\boldsymbol{\pi}_0$ ε	Steady State probability of (i, j) Steady State Probability of i Steady State Distribution vector of Y Idleness in modulation state i Steady state distribution $\boldsymbol{\pi}$ with the 0 element removed Idle Capacity	$\mathbb{P}(Y = i, Q = j)$ $\sum_{j \in \mathbb{N}_0} \pi_{i,j}$ $\{\pi_i\}_{i \in \mathbb{N}_0}$ $P(Y = i, Q = 0)$ $\{\pi_i\}_{i \in \mathbb{N}_0 / 0}$ $\mathbb{P}(Q = 0)$
$\mathbf{R}$ $\mathbf{e}^T$ $\mathbf{F}$ $\mathbf{I}$	Near-inverse matrix Ones vector Arrival Rates Matrix Identity Matrix	See Section 5.1 $(1, 1, \dots)^T$ $[F]_{i,j} = \begin{cases} \lambda_i, & i = j \\ \beta_{i,j}, & i \neq j \end{cases}$ $\text{diag}(1, 1, \dots)$
$\hat{q}(z; i)$ $\hat{\mathbf{q}}(z)$ $\bar{q}(z)$	Partial z -transform with $Y = i$ z -transform vector z -transform of Q	$\mathbb{E}[z^Q Y = i]$ $\{\hat{q}(z; i)\}_{i \in S}$ $\mathbb{E}[z^Q]$
$\cdot(\varepsilon)$ $f(z; \varepsilon)$	Quantity in the system with idleness ε in a parameterized family Function f evaluated at z where underlying terms (such as $\mathbf{F}$ or $\boldsymbol{\pi}$) are from the ε system	
$\mathbb{P}(a_1 \rightarrow a_2 \rightarrow a_3 / b_1, b_2)$	Probability the phase process, starting in a_1 , passes through a_2 then a_3 (possibly with repetition) without entering b_1 or b_2	

Appendix B Proofs

B.1 Lemma 1 Proof

Proof The element $[\mathbf{R}]_{i,j}$, which we call $r_{i,j}$ for notational convenience, is the expected time the process spends in state j between starting in state i and reaching the 0 state for the first time. We can represent this as

$$[\mathbf{R}]_{i,j} = r_{i,j} = \mathbb{E}\left[\sum_{k=1}^{N_{i,j}} T_k\right],$$

where $N_{i,j}$ is the number of visits the process makes to state j between starting in state i and reaching state 0 and T_k is the length of time spent in state j during the k^{th} visit to the state.

Because $T_1, T_2, \dots, T_k, \dots$ is a sequence of i.i.d. exponential random variables with mean $\frac{1}{-\alpha_{j,j}} < \infty$, if $\mathbb{E}[N_{i,j}] < \infty$ then $r_{i,j} = \mathbb{E}[N_{i,j}]\mathbb{E}[T_0] < \infty$ by Wald's lemma. Thus, if we can show that $\mathbb{E}[N_{i,j}] < \infty$ for all i and j , then $\mathbf{R}$ is well-posed.

First, let us show that $\mathbb{E}[N_{i,i}] < \infty$ for all $i \in \mathbb{N}_0$ and then show the same for when $i \neq j$. Because the chain is positive recurrent, with probability 1 the chain to return to state i infinitely many times after starting there. Because the chain is irreducible, there is some positive probability p_i that chain reaches state 0 before returning to state i . Thus, $N_{i,i} \sim \text{Geom}(p_i)$ and $\mathbb{E}[N_{i,i}] = \frac{1-p_i}{p_i} < \infty$.

Next consider when $i \neq j$. Consider two possibilities: starting in state i the chain either reaches 0 before j or vice versa. If the chain reaches 0 before j , then, by definition, $N_{i,j} = 0$. Conversely if the chain reaches j before 0 then the expected number of returns to j is exactly the same as if the process simply began in state j . Therefore

$$\mathbb{E}[N_{i,j}] = P(\text{Reach } j \text{ before } 0) \mathbb{E}[N_{j,j}] \leq \mathbb{E}[N_{j,j}] < \infty.$$

Thus, $\mathbb{E}[N_{i,j}] < \infty$ for all i, j and $r_{i,j} = \frac{\mathbb{E}[N_{i,j}]}{-\alpha_{j,j}} < \infty$. The $\mathbf{R}$ matrix is well posed.

Next, let us show that our proposed $\mathbf{R}$ has the near-inverse property. Notice that the quantity $N_{i,j}$ is an element of the fundamental matrix of an absorbing DTMC, particularly the embedded DTMC of our CTMC after making 0 an absorbing state. Let $\mathbf{H}_{\bar{0},\bar{0}}$ and be the one step transition matrix of the transient states in this absorbing embedded DTMC, i.e., all state except 0.

$$\mathbf{N}\mathbf{I} + \mathbf{H}_{\bar{0},\bar{0}} + (\mathbf{H}_{\bar{0},\bar{0}})^2 + \dots = \mathbf{I} + \mathbf{N}\mathbf{H}_{\bar{0},\bar{0}} = \mathbf{I} + \mathbf{H}_{\bar{0},\bar{0}}\mathbf{N}$$

or, after rearranging,

$$\mathbf{I} = \mathbf{N}(\mathbf{I} - \mathbf{H}_{\bar{0},\bar{0}}) = (\mathbf{I} - \mathbf{H}_{\bar{0},\bar{0}})\mathbf{N}.$$

This is precisely $\mathbf{N} = (\mathbf{I} - \mathbf{H}_{\bar{0},\bar{0}})^{-1}$ and therefore

$$-(\mathbf{W}_{\bar{0},\bar{0}})^{-1}\mathbf{N} = (\mathbf{W}_{\bar{0},\bar{0}})^{-1}(\mathbf{H}_{\bar{0},\bar{0}} - \mathbf{I}) = (\mathbf{A}_{\bar{0},\bar{0}})^{-1}.$$

where $\mathbf{W}_{\bar{0},\bar{0}}$ is the holding rates matrix for all states except 0 and $\mathbf{A}_{\bar{0},\bar{0}} = [\alpha_{i,j}]_{i,j \in \mathbb{N}_0/0}$.

Thus,

$$r_{i,j} = \begin{cases} [(\mathbf{A}_{\bar{0},\bar{0}})^{-1}]_{i,j}, & i, j \in \mathbb{N}_0/0 \\ 0, & \text{else} \end{cases}$$

The steady-state balance equation provides that

$$\pi_0(\alpha_{0,1}, \alpha_{0,2}, \dots) + (\pi_1, \pi_2, \dots)\mathbf{A}_{\bar{0},\bar{0}} = 0$$

or, equivalently,

$$(\alpha_{0,1}, \alpha_{0,2}, \dots)(\mathbf{A}_{\bar{0},\bar{0}})^{-1} = -\frac{1}{\pi_0}(\pi_1, \pi_2, \dots).$$

Thus, we can now construct our matrix $\mathbf{A}\mathbf{R}$. For $i, j \in \mathbb{N}_0/0$, $[\mathbf{A}_{\bar{0},\bar{0}}\mathbf{R}]_{i,j} = \mathbb{I}\{i = j\}$. When $i = 0$, $[\mathbf{A}_{\bar{0},\bar{0}}\mathbf{R}]_{i,j} = 0$ because $[\mathbf{R}]_{0,j} = 0$ for all $j \in \mathbb{N}_0$. Lastly, when $i \in \mathbb{N}_0/0$, $[\mathbf{A}_{\bar{0},\bar{0}}\mathbf{R}]_{i,0} = \frac{\pi_i}{\pi_0}$. We have our desired result that

$$\mathbf{A}\mathbf{R} = \begin{bmatrix} 0 & -\frac{1}{\pi_0}\boldsymbol{\pi}_{\bar{0}} \\ 0 & \mathbf{I} \end{bmatrix}.$$

□

B.2 Lemma 2 Proof

Proof We begin with the balance equation for the IMAP queue. For any state (y, q) ,

$$\begin{aligned} \sum_{i \neq y} \alpha_{i,y} (p_{i,y} \pi_{i,q-1} \mathbb{I}\{q > 0\} + (1 - p_{i,y}) \pi_{i,q}) + \lambda_y \pi_{y,q-1} \mathbb{I}\{q > 0\} + \mu \pi_{y,q+1} \\ = \left(\sum_{i \neq y} \alpha_{y,i} + \lambda_y + \mu \mathbb{I}\{q > 0\} \right) \pi_{y,q} \end{aligned}$$

Multiplying both sides of the equation by z^q and then summing over all values of q we can convert the above balance equation into a partial z -transform balance equation.

$$\begin{aligned} \sum_{q=0}^{\infty} \left(\sum_{i \neq y} \alpha_{i,y} (p_{i,y} \pi_{i,q-1} \mathbb{I}\{q > 0\} + (1 - p_{i,y}) \pi_{i,q}) + \lambda_y \pi_{y,q-1} \mathbb{I}\{q > 0\} + \mu \pi_{y,q+1} \right) z^q = \\ \sum_{q=0}^{\infty} \left(\sum_{i \neq y} \alpha_{i,y} (p_{i,y} \pi_{i,q-1} \mathbb{I}\{q > 0\} + (1 - p_{i,y}) \pi_{i,q}) \right) z^q \\ + \sum_{q=0}^{\infty} (\lambda_y \pi_{y,q-1} \mathbb{I}\{q > 0\} z^q) + \sum_{q=0}^{\infty} (\mu \pi_{y,q+1} z^q) = \\ \sum_{n=0}^{\infty} \left(\sum_{i \neq y} \alpha_{i,y} (p_{i,y} \pi_{i,n}) z^{n+1} \right) + \sum_{q=0}^{\infty} \left(\sum_{i \neq y} \alpha_{i,y} (1 - p_{i,y}) \pi_{i,q} \right) z^q \\ + \sum_{n=0}^{\infty} (\lambda_y \pi_{y,n} z^{n+1}) + \sum_{n=0}^{\infty} (\mu \pi_{y,n}) z^{n-1} - \mu \pi_{y,0} z^{-1} = \\ z \sum_{i \neq y} \alpha_{i,y} p_{i,y} \hat{q}(z; i) + \sum_{i \neq y} \alpha_{i,y} (1 - p_{i,y}) \hat{q}(z; i) + z \hat{q}(z; y) + z^{-1} \mu (\hat{q}(z; y) - \pi_{y,0}) \\ = (\lambda_y z + \frac{\mu}{z}) \hat{q}(z; y) + \sum_{i \neq y} \alpha_{i,y} (p_{i,y} z + (1 - p_{i,y})) \hat{q}(z; i) - \frac{\mu}{z} \pi_{y,0} \end{aligned}$$

and

$$\begin{aligned} \sum_{q=0}^{\infty} \left(\left(\sum_{i \neq y} \alpha_{y,i} + \lambda_y + \mu \mathbb{I}\{q > 0\} \right) \pi_{y,q} z^q \right) &= \sum_{q=0}^{\infty} \left(\left(\sum_{i \neq y} \alpha_{y,i} + \lambda_y + \mu \right) \pi_{y,q} z^q \right) - \mu \pi_{y,0} \\ &= \left(\sum_{i \neq y} \alpha_{y,i} + \lambda_y + \mu \right) \sum_{q=0}^{\infty} (\pi_{y,q} z^q) - \mu \pi_{y,0} \\ &= \left(\sum_{i \neq y} \alpha_{y,i} + \lambda_y + \mu \right) \hat{q}(z; y) - \mu \pi_{y,0} \end{aligned}$$

Re-equating the two we have

$$\begin{aligned} \left(\lambda_y z + \frac{\mu}{z} \right) \hat{q}(z; y) + \sum_{i \neq y} \alpha_{i,y} (p_{i,y} z + (1 - p_{i,y})) \hat{q}(z; i) - \frac{\mu}{z} \pi_{y,0} \\ = \left(\sum_{i \neq y} \alpha_{y,i} + \lambda_y + \mu \right) \hat{q}(z; y) - \mu \pi_{y,0} \end{aligned}$$

We can simplify this by adding

$\alpha_{y,y} (p_{y,y} z + (1 - p_{y,y})) \hat{q}(z; y) = \alpha_{y,y} \hat{q}(z; y)$ and $\sum_i \alpha_{i,y} (p_{i,y} (1 - z)) \hat{q}(z; i)$ to both sides to find

$$\begin{aligned} & \left(\lambda_y z + \frac{\mu}{z} \right) \hat{q}(z; y) + \sum_i \alpha_{i,y} (p_{i,y} + (1 - p_{i,y})) \hat{q}(z; i) - \frac{\mu}{z} \pi_{y,0} \\ &= (\lambda_y + \mu) \hat{q}(z; y) + \sum_i \alpha_{i,y} (p_{i,y} (1 - z)) \hat{q}(z; i) - \mu \pi_{y,0} \end{aligned}$$

Rearranging terms we have

$$\begin{aligned} & \sum_i \alpha_{i,y} \hat{q}(z; i) \\ &= (\lambda_y + \mu) \hat{q}(z; y) - \left(\lambda_y z + \frac{\mu}{z} \right) \hat{q}(z; y) + \sum_i \alpha_{i,y} (p_{i,y} (1 - z)) \hat{q}(z; i) - \mu \pi_{y,0} + \frac{\mu}{z} \pi_{y,0} \\ &= \lambda_y (1 - z) \hat{q}(z; y) + \mu \left(1 - \frac{1}{z} \right) \hat{q}(z; y) + \sum_i \alpha_{i,y} (p_{i,y} (1 - z)) \hat{q}(z; i) - \mu \left(1 - \frac{1}{z} \right) \pi_{y,0} \end{aligned}$$

We can collect each of these equations into a single matrix-form balance equation

$$\hat{\mathbf{q}}(z) \mathbf{A} = \hat{\mathbf{q}}(z) \left(\mathbf{F} (1 - z) + \mu \left(1 - \frac{1}{z} \right) \mathbf{I} \right) - \mu \left(1 - \frac{1}{z} \right) \boldsymbol{\pi}_{\text{idle}}$$

Right multiplying both sides by $\mathbf{R}$ we have the following relationship

$$\hat{\mathbf{q}}(z) = \hat{\mathbf{q}}(z; 0) \frac{\boldsymbol{\pi}}{\pi_0} + \hat{\mathbf{q}}(z) \left(\mathbf{F} (1 - z) + \mu \left(1 - \frac{1}{z} \right) \mathbf{I} \right) \mathbf{R} - \mu \left(1 - \frac{1}{z} \right) \boldsymbol{\pi}_{\text{idle}} \mathbf{R} \quad (\text{B1})$$

Eq. (B1) provides us $\hat{\mathbf{q}}(z)$ in terms of $\hat{\mathbf{q}}(z; 0)$ and $\hat{\mathbf{q}}(z)$ itself. By substituting our equation for $\hat{\mathbf{q}}(z)$ into itself we can yield an equation that places more emphasis on $\hat{\mathbf{q}}(z; 0)$. This iterated equation is our near-independence equation

$$\begin{aligned} \hat{\mathbf{q}}(z) &= \hat{\mathbf{q}}(z; 0) \frac{\boldsymbol{\pi}}{\pi_0} \left(\mathbf{I} + \left(\mathbf{F} (1 - z) + \mu \left(1 - \frac{1}{z} \right) \mathbf{I} \right) \mathbf{R} \right) \\ &+ \hat{\mathbf{q}}(z) \left(\left(\mathbf{F} (1 - z) + \mu \left(1 - \frac{1}{z} \right) \mathbf{I} \right) \mathbf{R} \right)^2 - \mu \left(1 - \frac{1}{z} \right) \boldsymbol{\pi}_{\text{idle}} \mathbf{R} \left(\mathbf{F} (1 - z) + \mu \left(1 - \frac{1}{z} \right) \mathbf{I} \right) \mathbf{R} \\ &- \mu \left(1 - \frac{1}{z} \right) \boldsymbol{\pi}_{\text{idle}} \mathbf{R} \end{aligned}$$

or, after collecting terms,

$$\hat{\mathbf{q}}(z) = \hat{\mathbf{q}}(z; 0) \frac{\boldsymbol{\pi}}{\pi_0} + \hat{\mathbf{q}}(z; 0) \frac{\boldsymbol{\pi}}{\pi_0} \left(\mathbf{F} (1 - z) + \mu \left(1 - \frac{1}{z} \right) \mathbf{I} \right) \mathbf{R} + \left(1 - \frac{1}{z} \right) \bar{\mathbf{Y}}(z)$$

If we instead right multiply our matrix-form balance equation by $\mathbf{e}^T = (1, 1, \dots)^T$ we have a flow balance equation

$$0 = \hat{\mathbf{q}}(z) \left(\mathbf{F} (1 - z) + \mu \left(1 - \frac{1}{z} \right) \mathbf{I} \right) \mathbf{e}^T - \mu \left(1 - \frac{1}{z} \right) \varepsilon$$

where $\boldsymbol{\pi}_{\text{idle}} \mathbf{e}^T = \varepsilon$.

Substituting our near-independence equation for $\hat{\mathbf{q}}(z)$ into our flow balance equation we have

$$\left(\hat{\mathbf{q}}(z; 0) \frac{\boldsymbol{\pi}}{\pi_0} + \hat{\mathbf{q}}(z; 0) \frac{\boldsymbol{\pi}}{\pi_0} \left(\mathbf{F} (1 - z) + \mu \left(1 - \frac{1}{z} \right) \mathbf{I} \right) \mathbf{R} + \left(1 - \frac{1}{z} \right) \bar{\mathbf{Y}}(z) \right)$$

$$\cdot \left(\mathbf{F}(1-z) + \mu \left(1 - \frac{1}{z} \right) \mathbf{I} \right) e^T - \mu \left(1 - \frac{1}{z} \right) \varepsilon = 0$$

Isolating $\hat{q}(z; 0)$ we find

$$\begin{aligned} \hat{q}(z; 0) &= \pi_0 \frac{\mu \left(1 - \frac{1}{z} \right) \varepsilon - \left(1 - \frac{1}{z} \right)^2 \bar{\mathbf{Y}}(z) (-z\mathbf{F} + \mu\mathbf{I}) e^T}{\left(1 - \frac{1}{z} \right) (\mu\varepsilon + \bar{\lambda}(1-z)) + \left(1 - \frac{1}{z} \right)^2 \boldsymbol{\pi}(-z\mathbf{F} + \mu\mathbf{I}) \mathbf{R}(-z\mathbf{F} + \mu\mathbf{I}) e^T} \\ &= \pi_0 \frac{1 - \mu^{-1}\varepsilon^{-1} \left(1 - \frac{1}{z} \right) \bar{\mathbf{Y}}(z) (-z\mathbf{F} + \mu\mathbf{I}) e^T}{1 + \varepsilon^{-1} \frac{\bar{\lambda}}{\mu} (1-z) + \mu^{-1}\varepsilon^{-1} \left(1 - \frac{1}{z} \right) \boldsymbol{\pi}(-z\mathbf{F} + \mu\mathbf{I}) \mathbf{R}(-z\mathbf{F} + \mu\mathbf{I}) e^T} \end{aligned}$$

or, after condensing terms,

$$\hat{q}(z; 0) = \pi_0 \frac{1 + U(z)}{1 + M \ln(z) / \varepsilon + V(z)}$$

□

B.3 Lemma 4 Proof

Proof As seen in Theorem 1, $\mathbf{R}$ and $\boldsymbol{\pi}$ often appear right multiplied by the matrix $\left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right)$. We can use the bounds obtained in Lemma 3 above to derive a number of related bounds on these related quantities.

1. $\left| [\pi^{(\varepsilon)} \left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right)]_j \right| \leq \tilde{C}_1 a^j$ for some $\tilde{C}_1 > 0$.

$$\begin{aligned} \left| [\pi^{(\varepsilon)} \left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right)]_j \right| &= \left| \sum_{m=0}^{\infty} \pi_m [\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)}]_{m,j} \right| \\ &= \left| \pi_{j-1} [\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)}]_{j-1,j} + \pi_j [\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)}]_{j,j} + \pi_{j+1} [\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)}]_{j+1,j} \right| \\ &= \left| \pi_{j-1} [\alpha_{j-1,j} p_{j-1,j}] + \pi_j [\mu^{(\varepsilon)} - z^\varepsilon \lambda_j] + \pi_{j+1} [\alpha_{j+1,j} p_{j+1,j}] \right| \\ &\leq \pi_{j-1} |\alpha_{j-1,j} p_{j-1,j}| + \pi_j \left| \mu^{(\varepsilon)} - z^\varepsilon \lambda_j \right| + \pi_{j+1} |\alpha_{j+1,j} p_{j+1,j}| \\ &\leq \pi_{j-1} \hat{\lambda} + \pi_j \left(\hat{\mu} + \hat{\lambda} \right) + \pi_{j+1} \hat{\lambda} \\ &\leq C_1 a^{j-1} \hat{\lambda} + C_1 a^j \left(\hat{\mu} + \hat{\lambda} \right) + C_1 a^{j+1} \hat{\lambda} \\ &\leq \left(C_1 a^{-1} \hat{\lambda} + C_1 \left(\hat{\mu} + \hat{\lambda} \right) + C_1 a \hat{\lambda} \right) a^j \\ &= \tilde{C}_1 a^j \end{aligned}$$

for $\tilde{C}_1 = C_1 a^{-1} \hat{\lambda} + C_1 \left(\hat{\mu} + \hat{\lambda} \right) + C_1 a \hat{\lambda}$

2. $\left| [\mathbf{R}^{(\varepsilon)} \left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right)]_{i,j} \right| \leq \tilde{C}_2 \tilde{b}^{(j-i)^+}$ for some $\tilde{C}_2 > 0$.

$$\left| [\mathbf{R}^{(\varepsilon)} \left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right)]_{i,j} \right| = \left| \sum_{m=0}^{\infty} r_{i,m}^{(\varepsilon)} [\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)}]_{m,j} \right|$$

$$\begin{aligned}
&= \left| r_{i,j-1}^{(\varepsilon)} (-z^\varepsilon \alpha_{j-1,j} p_{j-1,j}) + r_{i,j}^{(\varepsilon)} (\mu^{(\varepsilon)} - z^\varepsilon \lambda_j) + r_{i,j+1}^{(\varepsilon)} (-z^\varepsilon \alpha_{j+1,j} p_{j+1,j}) \right| \\
&\leq r_{i,j-1}^{(\varepsilon)} |-z^\varepsilon \alpha_{j-1,j} p_{j-1,j}| + r_{i,j}^{(\varepsilon)} |\mu^{(\varepsilon)} - z^\varepsilon \lambda_j| + r_{i,j+1}^{(\varepsilon)} |-z^\varepsilon \alpha_{j+1,j} p_{j+1,j}| \\
&\leq C_2 \tilde{b}^{(j-1-i)^+} \hat{\lambda} + C_2 \tilde{b}^{(j-i)^+} (\hat{\lambda} + \hat{\mu}) + C_2 \tilde{b}^{(j+1-i)^+} \hat{\lambda} \leq \tilde{C}_2 \tilde{b}^{(j-i)^+}
\end{aligned}$$

for $\tilde{C}_2 = C_2 \tilde{b}^{-1} \hat{\lambda} + C_2 (\hat{\lambda} + \hat{\mu}) + C_2 \tilde{b}^{-1} \hat{\lambda}$.

$$3. \left| [\mathbf{R}^{(\varepsilon)} (\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)}) e^T]_i \right| \leq \tilde{C}_3 i \text{ for some } \tilde{C}_3 > 0.$$

$$\begin{aligned}
&\left| [\mathbf{R}^{(\varepsilon)} (\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)}) e^T]_i \right| = \left| \sum_{j=0}^{\infty} [\mathbf{R}^{(\varepsilon)} (\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)})]_{i,j} \right| \\
&\leq \left| \sum_{j=0}^{i-1} [\mathbf{R}^{(\varepsilon)} (\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)})]_{i,j} \right| + \left| \sum_{j=i}^{\infty} [\mathbf{R}^{(\varepsilon)} (\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)})]_{i,j} \right| \\
&\leq \tilde{C}_2 i + \tilde{C}_2 \sum_{j=i}^{\infty} \tilde{b}^{j-i} \leq \tilde{C}_2 i + \tilde{C}_2 \frac{1}{1-\tilde{b}} \leq \tilde{C}_3 i,
\end{aligned}$$

for $\tilde{C}_2 = 1 + \frac{1}{1-\tilde{b}}$.

With these extended bounds in place, we're finally ready to show that the birth-death uniform stability conditions imply our original uniform stability conditions and, in turn, imply our desired heavy-traffic results.

1. For all $z \in (0, 1)$, as $\varepsilon \rightarrow 0$:

$$(a) (1 - z^{-\varepsilon}) \hat{\mathbf{q}}(z^\varepsilon; \varepsilon) \left((-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I}) \mathbf{R}^{(\varepsilon)} \right)^2 (-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu \mathbf{I}) e^T \rightarrow 0.$$

Because $(1 - z^{-\varepsilon}) \rightarrow 0$ as $\varepsilon \rightarrow 0$, if

$$\left| \hat{\mathbf{q}}(z^\varepsilon; \varepsilon) \left((-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I}) \mathbf{R}^{(\varepsilon)} \right)^2 (-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu \mathbf{I}) e^T \right|$$

is uniformly bounded for all z, ε then the convergence of $(1 - z^{-\varepsilon})$ to 0 implies the product converges to 0 as well. Utilizing the bounds we derived above we can see that

$$\begin{aligned}
&\left| \hat{\mathbf{q}}(z^\varepsilon; \varepsilon) \left((-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I}) \mathbf{R}^{(\varepsilon)} \right)^2 (-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu \mathbf{I}) e^T \right| \\
&= \left| \sum_{i,j} [\hat{\mathbf{q}}(z^\varepsilon; \varepsilon) (-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I})]_i [\mathbf{R}^{(\varepsilon)} (-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I})]_{i,j} \right. \\
&\quad \left. \cdot [\mathbf{R}^{(\varepsilon)} (-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I}) e^T]_j \right|
\end{aligned}$$

$$\begin{aligned}
&\leq \left| \sum_{i,j} \left(\tilde{C}_1 a^i \right) \left(\tilde{C}_2 \tilde{b}^{(j-i)^+} \right) \left(\tilde{C}_3 j \right) \right| = \tilde{C}_1 \tilde{C}_2 \tilde{C}_3 \left| \sum_{i=0}^{\infty} \left(a^i \left(\sum_{j=0}^{i-1} j + \sum_{j=i}^{\infty} j \tilde{b}^{j-i} \right) \right) \right| \\
&= \tilde{C}_1 \tilde{C}_2 \tilde{C}_3 \left| \sum_{i=0}^{\infty} \left(a^i \left(\sum_{j=0}^{i-1} j + \sum_{j=i}^{\infty} (j-i) \tilde{b}^{j-i} + \sum_{j=i}^{\infty} i \tilde{b}^{j-i} \right) \right) \right| \\
&= \tilde{C}_1 \tilde{C}_2 \tilde{C}_3 \left| \sum_{i=0}^{\infty} \left(a^i \left(\sum_{j=0}^{i-1} j + \sum_{k=0}^{\infty} k \tilde{b}^k + i \sum_{k=0}^{\infty} \tilde{b}^k \right) \right) \right| \\
&\leq \tilde{C}_1 \tilde{C}_2 \tilde{C}_3 \left| \sum_{i=0}^{\infty} \left(a^i \left(\frac{(i-1)(i-2)}{2} + \frac{\tilde{b}}{(1-\tilde{b})^2} + i \frac{1}{1-\tilde{b}} \right) \right) \right| \\
&= \tilde{C}_1 \tilde{C}_2 \tilde{C}_3 \left| \sum_{i=0}^{\infty} (a^i O(i^2)) \right| < \infty
\end{aligned}$$

(b) $\mu^{(\varepsilon)} (1 - z^\varepsilon) \pi_0^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu \mathbf{I} \right) e^T \rightarrow 0.$

Because $\mu^{(\varepsilon)} (1 - z^\varepsilon) \rightarrow 0$ as $\varepsilon \rightarrow 0$, if

$$\left| \pi_0^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T \right|$$

is uniformly bounded for all z, ε then the convergence of $\mu^{(\varepsilon)} (1 - z^\varepsilon)$ to 0 implies the product converges to 0 as well. Utilizing the bounds we derived above we can see that

$$\begin{aligned}
&\left| \pi_0^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T \right| \\
&= \left| \sum_{i,j} \pi_{i,0}^{(\varepsilon)} [\mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right)]_{i,j} [\mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T]_j \right| \\
&\leq \left| \sum_{i,j} (C_1 a^i) \left(\tilde{C}_2 \tilde{b}^{(j-i)^+} \right) \left(\tilde{C}_3 j \right) \right| = \tilde{C}_1 \tilde{C}_2 \tilde{C}_3 \left| \sum_{i=0}^{\infty} \left(a^i \left(\sum_{j=0}^{i-1} j + \sum_{j=i}^{\infty} j \tilde{b}^{j-i} \right) \right) \right| \\
&= C_1 \tilde{C}_2 \tilde{C}_3 \left| \sum_{i=0}^{\infty} \left(a^i \left(\sum_{j=0}^{i-1} j + \sum_{j=i}^{\infty} (j-i) \tilde{b}^{j-i} + \sum_{j=i}^{\infty} i \tilde{b}^{j-i} \right) \right) \right| \\
&= C_1 \tilde{C}_2 \tilde{C}_3 \left| \sum_{i=0}^{\infty} \left(a^i \left(\sum_{j=0}^{i-1} j + \sum_{k=0}^{\infty} k \tilde{b}^k + i \sum_{k=0}^{\infty} \tilde{b}^k \right) \right) \right| \\
&\leq \tilde{C}_1 \tilde{C}_2 \tilde{C}_3 \left| \sum_{i=0}^{\infty} \left(a^i \left(\frac{(i-1)(i-2)}{2} + \frac{\tilde{p}}{(1-\tilde{p})^2} + i \frac{1}{1-\tilde{p}} \right) \right) \right|
\end{aligned}$$

$$= C_1 \tilde{C}_2 \tilde{C}_3 \left| \sum_{i=0}^{\infty} (a^i O(i^2)) \right| < \infty$$

$$(c) \quad \pi_0^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T \rightarrow 0.$$

$$\begin{aligned} & \left| \pi_0^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T \right| = \left| \sum_{i=0}^{\infty} \pi_{i,0}^{(\varepsilon)} [\mathbf{R}^{(\varepsilon)} \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T]_i \right| \\ & \leq \left| \sum_{i=0}^{\infty} \pi_{i,0}^{(\varepsilon)} \tilde{C}_3 i \right| = \left| \sum_{i=0}^{\lfloor \log \frac{1}{\varepsilon} \rfloor} \pi_{i,0}^{(\varepsilon)} \tilde{C}_3 i + \sum_{i=\lfloor \log \frac{1}{\varepsilon} \rfloor + 1}^{\infty} \pi_{i,0}^{(\varepsilon)} \tilde{C}_3 i \right| \\ & \leq \left| \sum_{i=0}^{\lfloor \log \frac{1}{\varepsilon} \rfloor} \pi_{i,0}^{(\varepsilon)} \tilde{C}_3 \log \frac{1}{\varepsilon} + \sum_{i=\lfloor \log \frac{1}{\varepsilon} \rfloor + 1}^{\infty} \pi_{i,0}^{(\varepsilon)} \tilde{C}_3 i \right| \leq \tilde{C}_3 \left| \varepsilon \log \frac{1}{\varepsilon} \right| + \tilde{C}_3 \left| \sum_{i=\lfloor \log \frac{1}{\varepsilon} \rfloor + 1}^{\infty} \pi_{i,0}^{(\varepsilon)} i \right| \\ & \leq \tilde{C}_3 \left| \varepsilon \log \frac{1}{\varepsilon} \right| + \tilde{C}_3 \left| \sum_{i=\lfloor \log \frac{1}{\varepsilon} \rfloor + 1}^{\infty} C_1 a^i i \right| \end{aligned}$$

Because $\varepsilon \log \frac{1}{\varepsilon} \rightarrow 0$ as $\varepsilon \rightarrow 0$, we only need to establish that the second term also goes to 0. Notice that $\sum_{i=0}^{\infty} i a^i$ is a convergent sum because $0 < a < 1$. Thus as $\lfloor \log \frac{1}{\varepsilon} \rfloor + 1$ goes to ∞ as $\varepsilon \rightarrow 0$, the “tail sum” $\sum_{i=\lfloor \log \frac{1}{\varepsilon} \rfloor + 1}^{\infty} i a^i$ must also go to 0. Therefore both terms approach 0 and we have our desired result.

$$(d) \quad (1 - z^{-\varepsilon}) \pi^{(\varepsilon)} \left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right) \mathbf{R}^{(\varepsilon)} e^T \rightarrow 0$$

Because $(1 - z^{-\varepsilon}) \rightarrow 0$ as $\varepsilon \rightarrow 0$, if $\left| \pi^{(\varepsilon)} \left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right) \mathbf{R}^{(\varepsilon)} e^T \right|$ is uniformly bounded for all z, ε then their product also goes to 0.

$$\begin{aligned} & \left| \pi^{(\varepsilon)} \left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right) \mathbf{R}^{(\varepsilon)} e^T \right| = \left| \sum_{i=0}^{\infty} [\pi^{(\varepsilon)} \left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right)]_i [\mathbf{R}^{(\varepsilon)} e^T]_i \right| \\ & \leq \left| \sum_{i=0}^{\infty} \left(\tilde{C}_1 a^i \cdot C_3 i \right) \right| = \tilde{C}_1 C_3 \left| \frac{a}{(1-a)^2} \right| < \infty \end{aligned}$$

$$(e) \quad (1 - z^{-\varepsilon}) \hat{\mathbf{q}}(z^\varepsilon; \varepsilon) \left(\left(\mu^{(\varepsilon)} \mathbf{I} - z^\varepsilon \mathbf{F}^{(\varepsilon)} \right) \mathbf{R}^{(\varepsilon)} \right)^2 e^T \rightarrow 0.$$

Because $(1 - z^{-\varepsilon}) \rightarrow 0$ as $\varepsilon \rightarrow 0$, if

$$\left| \hat{\mathbf{q}}(z^\varepsilon; \varepsilon) \left(\left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) \mathbf{R}^{(\varepsilon)} \right)^2 \left(-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I} \right) e^T \right|$$

is uniformly bounded for all z, ε then the convergence of $(1 - z^{-\varepsilon})$ to 0 implies the product converges to 0 as well. Utilizing the bounds we derived above we

can see that

$$\begin{aligned}
& \left| \hat{\mathbf{q}}(z^\varepsilon; \varepsilon) \left((-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I}) \mathbf{R}^{(\varepsilon)} \right)^2 e^T \right| \\
&= \left| \sum_{i,j} [\hat{\mathbf{q}}(z^\varepsilon; \varepsilon) (-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I})]_i [\mathbf{R}^{(\varepsilon)} (-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I})]_{i,j} [\mathbf{R}^{(\varepsilon)} e^T]_j \right| \\
&\leq \left| \sum_{i,j} (\tilde{C}_1 a^i) (\tilde{C}_2 \tilde{b}^{(j-i)^+}) (C_3 j) \right| \\
&= \tilde{C}_1 \tilde{C}_2 C_3 \left| \sum_{i=0}^{\infty} \left(a^i \left(\sum_{j=0}^{i-1} j + \sum_{j=i}^{\infty} j \tilde{b}^{j-i} \right) \right) \right| \\
&= \tilde{C}_1 \tilde{C}_2 C_3 \left| \sum_{i=0}^{\infty} \left(a^i \left(\sum_{j=0}^{i-1} j + \sum_{j=i}^{\infty} (j-i) \tilde{b}^{j-i} + \sum_{j=i}^{\infty} i \tilde{b}^{j-i} \right) \right) \right| \\
&= \tilde{C}_1 \tilde{C}_2 C_3 \left| \sum_{i=0}^{\infty} \left(a^i \left(\sum_{j=0}^{i-1} j + \sum_{k=0}^{\infty} k \tilde{b}^k + i \sum_{k=0}^{\infty} \tilde{b}^k \right) \right) \right| \\
&\leq \tilde{C}_1 \tilde{C}_2 C_3 \left| \sum_{i=0}^{\infty} \left(a^i \left(\frac{(i-1)(i-2)}{2} + \frac{\tilde{b}}{(1-\tilde{b})^2} + i \frac{1}{1-\tilde{b}} \right) \right) \right| \\
&= \tilde{C}_1 \tilde{C}_2 C_3 \left| \sum_{i=0}^{\infty} (a^i O(i^2)) \right| < \infty
\end{aligned}$$

(f) $\mu^{(\varepsilon)} (1 - z^\varepsilon) \pi_{\text{idle}}^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} (\mu^{(\varepsilon)} I - z^\varepsilon \mathbf{F}^{(\varepsilon)}) \mathbf{R}^{(\varepsilon)} e^T \rightarrow 0.$

Because $\mu^{(\varepsilon)} (1 - z^\varepsilon) \rightarrow 0$ as $\varepsilon \rightarrow 0$, if $|\pi_{\text{idle}}^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} (\mu^{(\varepsilon)} I - z^\varepsilon \mathbf{F}^{(\varepsilon)}) \mathbf{R}^{(\varepsilon)} e^T|$ is uniformly bounded for all z, ε then the convergence of $\mu^{(\varepsilon)} (1 - z^\varepsilon)$ to 0 implies the product converges to 0 as well. Utilizing the bounds we derived above we can see that

$$\begin{aligned}
& \left| \pi_{\text{idle}}^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} (\mu^{(\varepsilon)} I - z^\varepsilon \mathbf{F}^{(\varepsilon)}) \mathbf{R}^{(\varepsilon)} e^T \right| \\
&= \left| \sum_{i,j} \pi_{i,0}^{(\varepsilon)} [\mathbf{R}^{(\varepsilon)} (-z^\varepsilon \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I})]_{i,j} [\mathbf{R}^{(\varepsilon)} e^T]_j \right| \\
&\leq \left| \sum_{i,j} (C_1 a^i) (\tilde{C}_2 p^{(j-i)^+}) (C_3 j) \right| \\
&= C_1 \tilde{C}_2 C_3 \left| \sum_{i=0}^{\infty} \left(a^i \left(\sum_{j=0}^{i-1} j + \sum_{j=i}^{\infty} j p^{j-i} \right) \right) \right|
\end{aligned}$$

$$\begin{aligned}
&= C_1 \tilde{C}_2 C_3 \left| \sum_{i=0}^{\infty} \left(a^i \left(\sum_{j=0}^{i-1} j + \sum_{j=i}^{\infty} (j-i) p^{j-i} + \sum_{j=i}^{\infty} i p^{j-i} \right) \right) \right| \\
&= C_1 \tilde{C}_2 C_3 \left| \sum_{i=0}^{\infty} \left(a^i \left(\sum_{j=0}^{i-1} j + \sum_{k=0}^{\infty} k p^k + i \sum_{k=0}^{\infty} p^k \right) \right) \right| \\
&\leq C_1 \tilde{C}_2 C_3 \left| \sum_{i=0}^{\infty} \left(a^i \left(\frac{(i-1)i}{2} + \frac{p}{(1-p)^2} + i \frac{1}{1-p} \right) \right) \right| \\
&= C_1 \tilde{C}_2 C_3 \left| \sum_{i=0}^{\infty} (a^i O(i^2)) \right| < \infty
\end{aligned}$$

(g) $\mu^{(\varepsilon)} \boldsymbol{\pi}_{\text{idle}}^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} e^T \rightarrow 0.$

$$\begin{aligned}
\left| \mu^{(\varepsilon)} \boldsymbol{\pi}_0^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} e^T \right| &= \left| \mu^{(\varepsilon)} \sum_{i=0}^{\infty} \pi_{i,0}^{(\varepsilon)} [\mathbf{R}^{(\varepsilon)} e^T]_i \right| \leq \left| \mu^{(\varepsilon)} \sum_{i=0}^{\infty} \pi_{i,0}^{(\varepsilon)} C_3 i \right| \\
&= |\mu^{(\varepsilon)}| \cdot \left| \sum_{i=0}^{\lfloor \log \frac{1}{\varepsilon} \rfloor} \pi_{i,0}^{(\varepsilon)} C_3 i + \sum_{i=\lfloor \log \frac{1}{\varepsilon} \rfloor + 1}^{\infty} \pi_{i,0}^{(\varepsilon)} C_3 i \right| \\
&\leq |\mu^{(\varepsilon)}| \cdot \left| \sum_{i=0}^{\lfloor \log \frac{1}{\varepsilon} \rfloor} \pi_{i,0}^{(\varepsilon)} C_3 \log \frac{1}{\varepsilon} + \sum_{i=\lfloor \log \frac{1}{\varepsilon} \rfloor + 1}^{\infty} \pi_{i,0}^{(\varepsilon)} C_3 i \right| \\
&\leq |\mu^{(\varepsilon)}| \cdot C_3 \left| \varepsilon \log \frac{1}{\varepsilon} \right| + |\mu^{(\varepsilon)}| \cdot C_3 \left| \sum_{i=\lfloor \log \frac{1}{\varepsilon} \rfloor + 1}^{\infty} \pi_{i,0}^{(\varepsilon)} i \right| \\
&\leq |\mu^{(\varepsilon)}| \cdot C_3 \left| \varepsilon \log \frac{1}{\varepsilon} \right| + |\mu^{(\varepsilon)}| \cdot C_3 \left| \sum_{i=\lfloor \log \frac{1}{\varepsilon} \rfloor + 1}^{\infty} C_1 a^i i \right|
\end{aligned}$$

Because $\varepsilon \log \frac{1}{\varepsilon} \rightarrow 0$ and $\mu^{(\varepsilon)} \rightarrow \mu^{(0)} \in (0, \infty)$ as $\varepsilon \rightarrow 0$, we only need to establish that the second term also goes to 0. Notice that $\sum_{i=0}^{\infty} i a^i$ is a convergent sum because $0 < a < 1$. Thus as $\lfloor \log \frac{1}{\varepsilon} \rfloor + 1$ goes to ∞ as $\varepsilon \rightarrow 0$, the “tail sum” $\sum_{i=\lfloor \log \frac{1}{\varepsilon} \rfloor + 1}^{\infty} i a^i$ must also go to 0. Therefore both terms approach 0 and we have our desired result.

2. For all $z \in (0, 1)$ and all $j \in \mathbb{N}_0$ as $\varepsilon \rightarrow 0$:

(a) $[(1 - z^{-\varepsilon}) \boldsymbol{\pi}^{(\varepsilon)} (\mu^{(\varepsilon)} I - z \mathbf{F}^{(\varepsilon)}) \mathbf{R}^{(\varepsilon)}]_j \rightarrow 0.$

$$\left| [\boldsymbol{\pi}^{(\varepsilon)} (\mu^{(\varepsilon)} I - z \mathbf{F}^{(\varepsilon)}) \mathbf{R}^{(\varepsilon)}]_j \right| = \left| \sum_{i=0}^{\infty} [\boldsymbol{\pi}^{(\varepsilon)} (\mu^{(\varepsilon)} I - z \mathbf{F}^{(\varepsilon)})]_{i r_{i,j}^{(\varepsilon)}} \right|$$

$$\begin{aligned}
&\leq \left| \sum_{i=0}^{\infty} \tilde{C}_1 a^i C_3 p^{(j-i)^+} \right| \leq \left| \sum_{i=0}^j \tilde{C}_1 a^i C_3 p^{(j-i)^+} + \sum_{i=j+1}^{\infty} \tilde{C}_1 a^i C_3 \right| \\
&\leq \left| \sum_{i=0}^j \tilde{C}_1 a^i C_3 p^{(j-i)^+} \right| + \left| \sum_{i=j+1}^{\infty} \tilde{C}_1 a^i C_3 \right| \leq \left| \sum_{i=0}^j \tilde{C}_1 a^i C_3 p^{(j-i)^+} \right| + \left| \tilde{C}_1 C_3 \frac{q^{j+1}}{1-q} \right| < \infty
\end{aligned}$$

Thus because $(1 - z^{-\varepsilon}) \rightarrow 0$ as $\varepsilon \rightarrow 0$ and $\left| [\boldsymbol{\pi}^{(\varepsilon)} \left(\mu^{(\varepsilon)} \mathbf{I} - z^{\varepsilon} \mathbf{F}^{(\varepsilon)} \right) \mathbf{R}^{(\varepsilon)}]_j \right|$ is bounded uniformly across ε, z , their product goes to 0 as $\varepsilon \rightarrow 0$.

$$(b) \left[\mu^{(\varepsilon)} (1 - z^{-\varepsilon}) \hat{\mathbf{q}}(z^{\varepsilon}; \varepsilon) \left((-z^{\varepsilon} \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I}) \mathbf{R} \right)^2 \right]_j \rightarrow 0$$

$$\begin{aligned}
&\left| [\hat{\mathbf{q}}(z^{\varepsilon}; \varepsilon) ((-z^{\varepsilon} \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I}) \mathbf{R})^2]_j \right| \\
&= \left| \sum_{i,k} [\hat{\mathbf{q}}(z^{\varepsilon}; \varepsilon) (-z^{\varepsilon} \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I})]_i \cdot [\mathbf{R}^{(\varepsilon)} (-z^{\varepsilon} \mathbf{F} + \mu^{(\varepsilon)} \mathbf{I})]_{i,k} \cdot r_{k,j}^{(\varepsilon)} \right| \\
&= a \left| \sum_{i,k} \tilde{C}_1 a^i \cdot \tilde{C}_3 \tilde{b}^{(k-i)^+} \cdot C_3 \tilde{b}^{(j-k)^+} \right| = a \left| \sum_{i=0}^{\infty} \sum_{k=0}^{\infty} (\tilde{C}_1 a^i \cdot \tilde{C}_3 \tilde{b}^{(k-i)^+} \cdot C_3 \tilde{b}^{(j-k)^+}) \right| \\
&= a \left| \sum_{i=0}^{\infty} \tilde{C}_1 a^i \left(\sum_{k=0}^{i-1} (\tilde{C}_3 \tilde{b}^{(k-i)^+} \cdot C_3 \tilde{b}^{(j-k)^+}) + \sum_{k=i}^{\infty} (\tilde{C}_3 \tilde{b}^{(k-i)^+} \cdot C_3 \tilde{b}^{(j-k)^+}) \right) \right| \\
&= a \tilde{C}_1 \tilde{C}_3 C_3 \left| \sum_{i=0}^{\infty} a^i \left(\sum_{k=0}^{i-1} (\tilde{b}^{(k-i)^+} \cdot \tilde{b}^{(j-k)^+}) + \sum_{k=i}^{\infty} (\tilde{b}^{(k-i)^+} \cdot \tilde{b}^{(j-k)^+}) \right) \right| \\
&= a \tilde{C}_1 \tilde{C}_3 C_3 \left| \sum_{i=0}^{\infty} a^i \left(\sum_{k=0}^{i-1} (1) + \sum_{k=i}^{\infty} (\tilde{b}^{(k-i)^+}) \right) \right| = a \tilde{C}_1 \tilde{C}_3 C_3 \left| \sum_{i=0}^{\infty} a^i \left(i + \frac{1}{1-\tilde{b}} \right) \right| \\
&= \tilde{C}_1 \tilde{C}_3 C_3 \left| \frac{a}{(1-a)^2} + \frac{1}{(1-\tilde{b})(1-a)} \right| < \infty
\end{aligned}$$

Thus because $(1 - z^{-\varepsilon}) \rightarrow 0$ as $\varepsilon \rightarrow 0$ and $[\hat{\mathbf{q}}(z^{\varepsilon}; \varepsilon) \left((-z^{\varepsilon} \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I}) \mathbf{R} \right)^2]_j$ is bounded uniformly across ε, z , their product goes to 0 as $\varepsilon \rightarrow 0$.

$$(c) \left[\mu^{(\varepsilon)} (1 - z^{-\varepsilon}) \boldsymbol{\pi}_0^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} \left((-z^{\varepsilon} \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I}) \mathbf{R} \right)^2 \right]_j \rightarrow 0.$$

$$\begin{aligned}
&\left| [\boldsymbol{\pi}_0^{(\varepsilon)} \mathbf{R}^{(\varepsilon)} (-z^{\varepsilon} \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I}) \mathbf{R}^{(\varepsilon)}]_j \right| \\
&= \left| \sum_{i,k} \pi_{i,0}^{(\varepsilon)} \cdot [\mathbf{R}^{(\varepsilon)} (-z^{\varepsilon} \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I})]_{i,k} \cdot r_{k,j}^{(\varepsilon)} \right|
\end{aligned}$$

$$\begin{aligned}
&= a \left| \sum_{i,k} \tilde{C}_1 a^i \cdot \tilde{C}_3 \tilde{b}^{(k-i)^+} \cdot C_3 \tilde{b}^{(j-k)^+} \right| \\
&= a \left| \sum_{i=0}^{\infty} \sum_{k=0}^{\infty} (\tilde{C}_1 a^i \cdot \tilde{C}_3 \tilde{b}^{(k-i)^+} \cdot C_3 \tilde{b}^{(j-k)^+}) \right| \\
&= a \left| \sum_{i=0}^{\infty} \tilde{C}_1 a^i \left(\sum_{k=0}^{i-1} (\tilde{C}_3 \tilde{b}^{(k-i)^+} \cdot C_3 \tilde{b}^{(j-k)^+}) + \sum_{k=i}^{\infty} (\tilde{C}_3 \tilde{b}^{(k-i)^+} \cdot C_3 \tilde{b}^{(j-k)^+}) \right) \right| \\
&= a \tilde{C}_1 \tilde{C}_3 C_3 \left| \sum_{i=0}^{\infty} a^i \left(\sum_{k=0}^{i-1} (\tilde{b}^{(k-i)^+} \cdot \tilde{b}^{(j-k)^+}) + \sum_{k=i}^{\infty} (\tilde{b}^{(k-i)^+} \cdot \tilde{b}^{(j-k)^+}) \right) \right| \\
&= a \tilde{C}_1 \tilde{C}_3 C_3 \left| \sum_{i=0}^{\infty} a^i \left(\sum_{k=0}^{i-1} (1) + \sum_{k=i}^{\infty} (\tilde{b}^{(k-i)^+}) \right) \right| \\
&= a \tilde{C}_1 \tilde{C}_3 C_3 \left| \sum_{i=0}^{\infty} a^i \left(i + \frac{1}{1-\tilde{b}} \right) \right| \\
&= \tilde{C}_1 \tilde{C}_3 C_3 \left| \frac{a}{(1-a)^2} + \frac{1}{(1-\tilde{b})(1-a)} \right| < \infty
\end{aligned}$$

Thus because $(1 - z^{-\varepsilon}) \rightarrow 0$ as $\varepsilon \rightarrow 0$ and $|\pi_0^{(\varepsilon)} R^{(\varepsilon)}(-z^{\varepsilon} \mathbf{F}^{(\varepsilon)} + \mu^{(\varepsilon)} \mathbf{I}) R^{(\varepsilon)}]_j|$ is uniformly bounded across all z, ε , their product also goes to 0 as $\varepsilon \rightarrow 0$.

(d) $[\mu^{(\varepsilon)} \pi_0^{(\varepsilon)} \mathbf{R}^{(\varepsilon)}]_j \rightarrow 0$.

$$\begin{aligned}
\left| [\mu^{(\varepsilon)} \pi_0^{(\varepsilon)} \mathbf{R}^{(\varepsilon)}]_j \right| &\leq \left| \mu^{(\varepsilon)} \right| \left| \sum_{i=0}^{\infty} \pi_{i,0}^{(\varepsilon)} r_{i,j}^{(\varepsilon)} \right| \leq \left| \mu^{(\varepsilon)} \right| \left| \sum_{i=0}^{\infty} \pi_{i,0}^{(\varepsilon)} r_{j,j}^{(\varepsilon)} \right| \\
&= \left| \mu^{(\varepsilon)} \right| \left| r_{j,j}^{(\varepsilon)} \right| \left| \sum_{i=0}^{\infty} \pi_{i,0}^{(\varepsilon)} \right| = \varepsilon \left| \mu^{(\varepsilon)} \right| \left| r_{j,j}^{(\varepsilon)} \right|
\end{aligned}$$

Because $|\mu^{(\varepsilon)}|$ and $|r_{j,j}^{(\varepsilon)}|$ approach constants $|\mu^{(0)}|$ and $|r_{j,j}^{(0)}|$ as $\varepsilon \rightarrow 0$, $\varepsilon |\mu^{(\varepsilon)}| |r_{j,j}^{(\varepsilon)}| \rightarrow 0$.

□

Appendix C Simulation Details

To validate the heavy-traffic approximation for the mean queue length, we simulate the hybrid-abandonment system for 10^8 jobs at each parameter set (ρ, θ, μ_1) and record the second queue length at each arrival to the first queue. By PASTA [32], these are samples of the time-average second queue length and provide an easy way to estimate the mean queue length and the standard deviation of our estimator.

To validate that the second queue length is approximately exponentially distributed and that the queues are not meaningfully correlated, we again simulated the system

for 10^8 jobs at each parameter set but instead broke the 10^8 job-long runs into 20 “batches” of $5 \cdot 10^6$ job-long runs. Each of these batches yielded one estimate of our desired quantity, either the correlation of the queue lengths or the KS statistic between the second queue length and the exponential or geometric random variable.

Finally, to test the efficacy of our approximation in capacity planning, we simulate the system for 40 batches of 10^5 jobs each at each of 400 different μ_2 values for each parameter set (λ, θ, μ_1) . Because the relative cost parameter c does not affect the queue length at any system configuration, we did not need to rerun the system at each c value. Instead, the existing queue length data could be reused for each differing c value.

Unaggregated data and full simulation code can be found at <https://github.com/CJCurtis/HybridAbandonmentSimulator>.

References

- [1] Jackson, J.R.: Networks of waiting lines. *Operations research* **5**(4), 518–521 (1957)
- [2] Reed, J., Ward, A.R.: A diffusion approximation for a generalized jackson network with reneging. In: *Proceedings of the 42nd Annual Allerton Conference on Communication, Control, and Computing* (2004)
- [3] Zayas-Caban, G., Xie, J., Green, L.V., Lewis, M.E.: Policies for physician allocation to triage and treatment in emergency departments. *IIE Transactions on Healthcare Systems Engineering* **9**(4), 342–356 (2019)
- [4] Zayas-Cabán, G., Cochran, A.L.: Scheduling servers in a two-stage queue with abandonments and costs. *Probability in the Engineering and Informational Sciences* **37**(3), 833–851 (2023)
- [5] Zayas-Cabán, G., Xie, J., Green, L.V., Lewis, M.E.: Dynamic control of a tandem system with abandonments. *Queueing Systems* **84**, 279–293 (2016)
- [6] Andradóttir, S., Ayhan, H.: Optimal server assignment in queues with flexible servers and abandonments. *Stochastic Systems* **13**(3), 360–376 (2023)
- [7] Choudhury, G.L., Whitt, W.: Heavy-traffic asymptotic expansions for the asymptotic decay rates in the BMAP/G/1 queue. *Stochastic Models* **10**(2), 453–498 (1994)
- [8] Yechiali, U., Naor, P.: Queuing problems with heterogeneous arrivals and service. *Operations Research* **19**(3), 722–734 (1971)
- [9] Burman, D.Y., Smith, D.R.: An asymptotic analysis of a queueing system with Markov-modulated arrivals. *Operations Research* **34**(1), 105–119 (1986)
- [10] Falin, G., Falin, A.: Heavy traffic analysis of M/G/1 type queueing systems with

- Markov-modulated arrivals. *Top* **7**(2), 279–291 (1999)
- [11] Dimitrov, M.: Single-server queueing system with Markov-modulated arrivals and service times. *Pliska Studia Mathematica Bulgarica* **20**(1), 53–62 (2011)
 - [12] Mitrani, I.: Approximate solutions for heavily loaded Markov-modulated queues. *Performance Evaluation* **62**(1-4), 117–131 (2005)
 - [13] Thorsdottir, H., Verloop, I.M.: Markov-modulated M/G/1-type queue in heavy traffic and its application to time-sharing disciplines. *Queueing Systems* **83**(1), 29–55 (2016)
 - [14] Mou, S., Maguluri, S.T.: Heavy-traffic queue length behavior in a switch under Markovian arrivals. *Advances in Applied Probability* **56**(3), 1106–1152 (2024)
 - [15] Grosof, I., Hong, Y., Harchol-Balter, M., Scheller-Wolf, A.: The reset and marc techniques, with application to multiserver-job analysis. *Performance Evaluation* **162**, 102378 (2023)
 - [16] Hurtado-Lange, D., Grosof, I.: Markov modulated JSQ in heavy traffic via the poisson equation (2026)
 - [17] Neuts, M.F.: The M/M/1 queue with randomly varying arrival and service rates. *Opsearch* **15**, 139–157 (1978)
 - [18] Neuts, M.F.: *Matrix-geometric Solutions in Stochastic Models: an Algorithmic Approach*. Johns Hopkins University Press, Baltimore, MD (1994)
 - [19] Latouche, G., Ramaswami, V.: *Introduction to Matrix Analytic Methods in Stochastic Modeling*. SIAM, Philadelphia, PA (1999)
 - [20] Ramaswami, V.: The N/G/1 queue and its detailed analysis. *Advances in Applied Probability* **12**(1), 222–261 (1980)
 - [21] Regterschot, G., De Smit, J.: The queue M|G|1 with Markov modulated arrivals and services. *Mathematics of operations research* **11**(3), 465–483 (1986)
 - [22] Ciucu, F., Poloczek, F.: Two extensions of kingman’s gi/g/1 bound. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* **2**(3), 1–33 (2018)
 - [23] Wang, J., Abouee-Mehrizi, H., Baron, O., Berman, O.: Tandem queues with impatient customers. *Performance Evaluation* **135**, 102011 (2019)
 - [24] Ramaswami, V., Taylor, P.G.: Some properties of the rate operators in level dependent quasi-birth-and-death processes with countable number of phases. *Stochastic Models* **12**(1), 143–164 (1996)

- [25] Latouche, G., Nguyen, G.T., Taylor, P.G.: Queues with boundary assistance: the effects of truncation. *Queueing Systems* **69**(2), 175–197 (2011)
- [26] Kroese, D., Scheinhardt, W., Taylor, P.: Spectral properties of the tandem Jackson network, seen as a quasi-birth-and-death process (2004)
- [27] Lian, Z., Zhao, N.: Departure processes and busy periods of a tandem network. *Operational Research* **11**(3), 245–257 (2011)
- [28] He, Q.-M.: *Fundamentals of Matrix-analytic Methods* vol. 365. Springer, New York, NY (2014)
- [29] Cohen, J.: *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates, Hillsdale, NJ (1988)
- [30] Kleinrock, L.: *Queueing Systems. Volume I: Theory*. Wiley, New York (1975)
- [31] Holt, C.C., Modigliani, F., Muth, J.F., Simon, H.A.: *Planning Production, Inventories, and Workforce*. Prentice-Hall, Englewood Cliffs, NJ (1960)
- [32] Wolff, R.W.: Poisson arrivals see time averages. *Operations Research* **30**(2), 223–231 (1982) <https://doi.org/10.1287/opre.30.2.223>