Analysis of Markovian Arrivals and Service with Applications to Intermittent Overload

Isaac Grosof[*], Yige Hong[†], Mor Harchol-Balter[‡]

October 1, 2024

Abstract

In many important real-world queueing settings, arrival and service rates fluctuate over time. We consider the MAMS system, where the arrival and service rates each vary according to an arbitrary finite-state Markov chain, allowing intermittent overload to be modeled. This model has been extensively studied, and we derive results matching those found in the literature via a somewhat novel framework.

We derive a characterization of mean queue length in the MAMS system, with explicit bounds for all arrival and service chains at all loads, using our new framework. Our bounds are tight in heavy traffic. We prove even stronger bounds for the important special case of two-level arrivals with intermittent overload.

Our framework is based around the concepts of relative arrivals and relative completions, which have previously been used in studying the MAMS system, under different names. These quantities allow us to tractably capture the transient correlational effect of the arrival and service processes on the mean queue length.

1 Introduction

Basic queueing theory assumes *independent and identically-distributed (i.i.d.)* interarrival times and service durations. In reality, the arrival rate might fluctuate between periods of high arrival rate and lower arrival rate. The same is true for the service rate. Because of these fluctuations, using a formula which is designed for a fixed arrival rate and service rate can lead to highly inaccurate results.

There has been extensive work on characterizing queue length in systems where arrival rate and/or service rate fluctuate with time. This technical report presents a somewhat new framework for analyzing such systems, which we use to derive closed-form expressions for mean queue length in systems where the arrival rate and/or service rate fluctuates with time. Specifically, we allow the arrival rate and service rate to each vary according to arbitrary finite-state Markov chains, and derive closed-form expressions for this setting.

Results equivalent to ours were first derived by Yechiali and Naor [32] in the two-level arrival setting, derived by Falin and Falin [8] in the Markov-modulated arrivals setting, and the results of Falin and Falin were extended to our Markov-modulated arrivals and service setting by Dimitrov [6].

The aforementioned results are stronger than ours, as they also characterize the entire moment generating function (MGF) of the queue length in heavy traffic. We discuss these results further in Section 2.

Our framework is presented here in the hopes that it may be useful as a starting point to build future generalizations to more complex settings.

[*]Electrical and Computer Engineering Department, University of Illinois, Urbana-Champaign & Department of Industrial Engineering and Management Science, Northwestern University, izzy.grosof@northwestern.edu

[†]Computer Science Department, Carnegie Mellon University, harchol@cs.cmu.edu

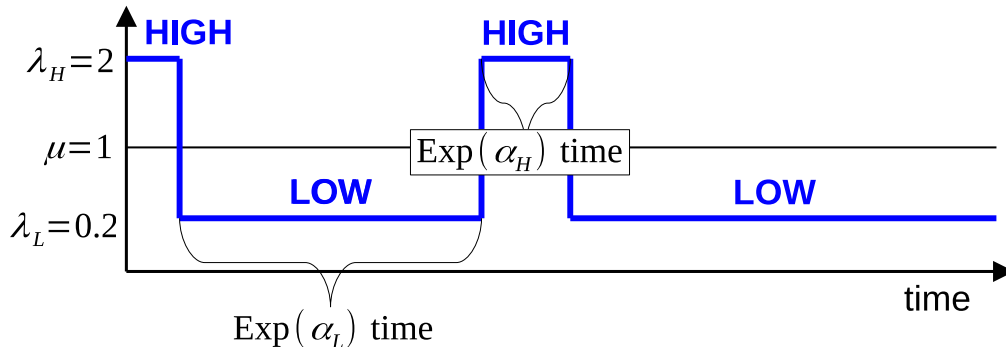[‡]Computer Science Department, Carnegie Mellon University, yigeh@andrew.cmu.edu

Figure 1: A time-varying arrival process with intermittent overload. The system spends 5/6 of its time on average with a low arrival rate of $\lambda_L = 0.2$, and 1/6 of its time with a high arrival rate of $\lambda_H = 2$. The average arrival rate is $\lambda = 0.5$, which is below the service rate of $\mu = 1$.

A specific motivating example: Intermittent overload

As a motivating example, we study the problem of intermittent overload under a two-level arrival process.

Systems are often built so that the average load, ρ, is not too high. This stems from the general understanding that mean queue length increases as a function related to $1/(1-\rho)$. Standard operating principles for computer systems practitioners recommend keeping the *average* load under 0.5 [19].

However, a system with an average load of 0.5 might not ever have an instantaneous load of 0.5, but rather it could be running at very low load most of the time, with brief periods of overload. The example shown in Fig. 1 is of a system which spends $\frac{5}{6}$ of the time at load 0.2 and $\frac{1}{6}$ of the time at load 2.0, resulting in an average load of 0.5. Because this system experiences intermittent overload, its mean queue length is *much larger* than that of a M/M/1 with load 0.5, especially when high-load and low-load periods are long.

Intermittent overload, or fluctuating load more generally, is common in compute settings [1, 4, 9, 16, 21, 29] and also in medical setting settings [3, 14]. It is the reason that capacity provisioning is difficult, and it is the motivation behind the many papers written on dynamic capacity provisioning [10, 18, 26].

The simple two-level arrivals system has been extensively studied by Yechiali and Naor [32], who derived a closed-form formula for mean queue length, as well as a tight result in heavy traffic. We reanalyze this system with our framework, proving comparable results.

Further work on the two-level system also exists, which we discuss in Section 2 [13, 30].

Our results and approach

Summary of results: This technical report proves closed-form bounds for mean queue length in the MAMS model, using a somewhat new framework, matching existing bounds in the literature. Our results are tight in heavy traffic and hold at all loads. When we specialize to the two-level system, we prove tight closed-form bounds in two additional settings: in the limit as the system switches quickly between arrival rates ($\alpha_H, \alpha_L \to \infty$), and in a setting with intermittent overload setting and slow switching between arrival rates ($\lambda_H > \mu; \alpha_H, \alpha_L \to 0$). Our main MAMS theorem is Theorem 5.1, which we use to give closed-form bounds in Corollary 5.1, and our main two-level-arrival result is Theorem 3.1, which we use to give additional closed-form bounds in Corollary 7.1.

Approach: Our framework is a somewhat novel extension of the drift method introduced by Eryilmaz and Srikant [7]. In a nutshell, the drift method involves exploiting the fact that the stationary drift of any given random variable related to system performance is zero. The difficulty in using the drift method is choosing an appropriate test function which defines the random variable. In the past, many applications of the drift method have involved i.i.d. arrivals and service. One exception is Mou and Maguluri [22], which required a complex extension to the drift method called the m-step drift method.

Because the m-step drift method is much more complicated, the convergence rate in heavy traffic proven via this method is weaker than the existing result of Falin and Falin [8]. Our new framework is finding a novel test function which allows for greater analytical tractability, matching the existing literature.

Our new test function introduces a way of thinking about Markovian systems, which we call *relative arrivals* and *relative completions*, which is an independent rediscovery of the technique introduced by Falin and Falin [8]. The basic idea is as follows: In an i.i.d. system, it is easy to quantify the expected number of arrivals by time t; call this $\mathbb{E}[A(t)]$. For a system with Markovian arrivals, such as the two-level-arrival system, the expected number of arrivals by time t depends on which arrival state the system is in at time 0: If the system is in the high load state at time 0, it will see more arrivals by time t than if it is initially in the low load state. If we let $A_H(t)$ (respectively, $A_L(t)$) denote the number of arrivals by time t for a two-level system initialized in state H (respectively, L), then we expect $\mathbb{E}[A_H(t)] > \mathbb{E}[A_L(t)]$. We devise our novel test function by explicitly quantifying this difference. Specifically, we define the relative arrival quantities $\Delta(H) := \lim_{t \to \infty} \mathbb{E}[A_H(t)] - \mathbb{E}[A(t)]$ and $\Delta(L) := \lim_{t \to \infty} \mathbb{E}[A_L(t)] - \mathbb{E}[A(t)]$ (see Definition 3.1). In the general MAMS system, we also define relative completions similarly. These quantities are the heart of our novel test function.

Structure of presentation: To illustrate our method in the simplest setting possible, we start by presenting our results for the two-level arrival process, in Section 3. This allows us to use a slightly simpler test function and also greatly simplify the notation and analysis, while being an interesting result in its own right. In the next several sections (Sections 4 to 6), we then solve the general MAMS system. In Section 7, we return to the two-level system to tighten our bounds in this setting.

Outline of paper

This technical report is organized as follows:

- Section 2: We discuss prior work on the two-level arrivals system, and on queueing systems with Markovian arrivals and/or Markovian service.

- Section 3: We introduce the two-level arrivals system, the drift method, and the relative arrivals function specialized to this setting. We prove our main result, Theorem 3.1, for the two-level arrivals system.

- Section 4: We introduce the MAMS model, and generalize the drift method and the relative arrivals and relative completions function to the MAMS setting.

- Section 5: We state our main result, Theorem 5.1, for the general MAMS system.

- Section 6: We prove our main result for the general MAMS system.

- Section 7: We prove additional bounds for the two-level arrivals system.

- Section 8: We empirically demonstrate the accuracy and tightness of our bounds via simulation.

2 Prior work

We study the two-level arrival system. None of our results are novel – all have previously appeared in the literature. We give a somewhat novel framework to derive these results.

2.1 Results for two-level arrivals system

Yechiali and Naor [32] study the same two-level system as the one studied in this technical report. The paper it derives the explicit formula for the queue length and the steady-state distribution using a generating function technique with some parameters determined by roots of cubic equations. Their equation (33) exactly matches our main two-level result, Theorem 3.1. The paper also studies the fast and slow switching settings that we study in Section 7, in their Section II, cases C and D.

There has been additional study of the two-level system by Gupta et al. [13], who prove that mean queue length decreases monotonically as switching rate increases, and proves stochastic ordering results, and by Vesilo et al. [30], who prove convexity and concavity results, additional stochastic ordering, and closing certain edge cases left open by Gupta et al. [13].

2.2 Bounds at all loads via Matrix Analytic Methods

Matrix analytic methods reduce the problem of deriving the stationary distribution of an MAMS system to a series of linear algebraic calculations [20, 23, 24, 27]. Bounds can also be derived via spectral analysis of matrices related to the underlying system.

This line of work was initiated by Neuts [23], which provided matrix equations to characterize a variety of performance metrics, including busy period lengths, effective service times, virtual waiting times, time in system, and queue length. For any given parameters of the system, appropriate matrices could be constructed.

An important follow-on work by Regterschot and de Smit [28] used the Matrix Analytic approach to derive computationally tractable bounds, approximation, and analysis of these performance quantities.

From these early matrix analytic works, an entire line of papers has followed, generalizing to more complex arrival and service processes, and improving the computational complexity of the algorithms.

A noteworthy recent result by Ciucu and Poloczek [5] gives a quite computationally efficient method to derive sharp bounds on the MGF of the queue length, for the MAMS system and a variety of other related systems, via Matrix Analytic methods.

2.3 Heavy-traffic results for MAMS system

The first heavy-traffic result in the MAMS system was by Burman and Smith [2], who studied a system with Markovian arrivals and general i.i.d. service time. They proceed via a diffusion-scaling approach, and prove a nearly-rigorous characterization of mean queue length in heavy traffic, but do not prove a key interchange-of-limits lemma which would be necessary for a fully-rigorous proof. Their heavy traffic results match our Corollary 5.1.

Next, Falin and Falin [8] study the same system as [2], and rigorously characterize its queue length distribution, including characterizing the limiting MGF in heavy traffic. When specialized to the mean, their results match our main result, Theorem 5.1. They also introduce a concept equivalent to our relative arrivals quantity, which we describe in Section 3.3. Their a_m for an arrival state m is equivalent to our $\Delta_A(m)$. Their results fully subsume ours, and use a closely related approach to our approach.

Dimitrov [6] generalize the results of Falin and Falin [8] to allow Markovian service rates, in addition to Markovian arrival rates and i.i.d. job sizes. Their model differs slightly from ours, in that they use a single chain to modulate arrival and service rates, but the key results match.

Our framework is based on the drift method of Eryilmaz and Srikant [7], incorporating the relative arrivals quantity. While the relative arrivals quantity was an independent rediscovery of the approach by Falin and Falin [8], the only prior work to use the drift method to analyze the MAMS system was Mou and Maguluri [22], which analyzing a system with Markovian Arrivals and exponential service in the heavy traffic limit. While the paper proves that the scaled mean queue length $E[Q(1-\rho)]$ converges in that limit, the paper does not give a closed form expression for the limiting value, leaving it in the form of an infinite summation which can be shown to be equivalent to the results of this technical report and of the prior work in this section.

Mou and Maguluri [22]'s key analytical technique is an extension to the drift method [7] called the "m-step drift method," which is considerably more complex than the original drift method. As a result, their bounds are somewhat weaker, resulting in a slower heavy-traffic convergence rate. By contrast, we extend the drift method in a different, simpler direction with our relative arrival and relative completion functions, while managing to prove tighter and more explicit results, matching the results of Falin and Falin [8] and Dimitrov [6] when restricted to mean queue length.
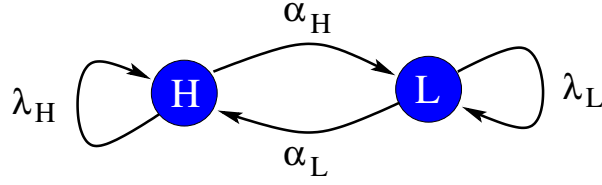
Figure 2: The arrival chain of the two-level arrival process.

3   Two-level arrivals

In this section, we will introduce and analyze the two-level arrival system, using a drift analysis based on our novel relative arrivals function. Our main result in this section in Theorem 3.1, which characterizes the mean queue length in the two-level arrival system. The techniques introduced in this section are then generalized to handle the entire MAMS model in Section 5.

We define the two-level arrivals model in Section 3.1. We introduce the drift method in Section 3.2. We introduce our novel relative arrivals function in Section 3.3. We sketch our proof in Section 3.4, introducing the function whose drift we will analyze. We characterize its drift in Section 3.5, and use it to prove our main result Section 3.6.

3.1   Two-level model

In the two-level arrivals model, jobs arrive to a single-server queue. Jobs wait in the queue in FCFS order. When a job reaches the head of the queue, it enters service. Job service times are exponentially distributed with rate μ. All jobs are indistinguishable.

Jobs arrive according to a two-level process. The two-level arrival process is depicted in Fig. 1. The arrival process consists of two arrival rates: A high rate λ_H and a low rate λ_L. The system switches between the two rates according to the switching rates α_H, the rate of switching from H to L, and α_L, the rate of switching from L to H. A Markovian view on this arrival process is shown in Fig. 2.

Let λ denote the long-term average arrival rate:

$$\lambda = \frac{\lambda_H/\alpha_H + \lambda_L/\alpha_L}{1/\alpha_H + 1/\alpha_L} = \frac{\lambda_H\alpha_L + \lambda_L\alpha_H}{\alpha_L + \alpha_H}. \tag{1}$$

Let $\rho := \lambda/\mu$ be the system load.

We will write y to denote a generic arrival state, $y \in \{H, L\}$. We will write (q, y) to denote a generic system state, where q is the queue length and y is the arrival state. Let $\bar{y}$ denote the opposite arrival state from y: $\bar{H} = L$ and $\bar{L} = H$. Let $Y(t)$ denote the arrival state at time t, and let Y be the corresponding stationary random variable. We will also consider $Y^{arrival}$, the arrival-rate-weighted arrival state:

$$P(Y^{arrival} = y) := P(Y = y)\frac{\lambda_y}{\lambda}. \tag{2}$$

It is straightforward to compute the underlying distribution for Y. We then get $Y^{arrival}$ from (2):

$$Y = \begin{cases} H & \text{w.p. } \frac{1/\alpha_H}{1/\alpha_H+1/\alpha_L} = \frac{\alpha_L}{\alpha_H+\alpha_L} \\ L & \text{w.p. } \frac{1/\alpha_L}{1/\alpha_H+1/\alpha_L} = \frac{\alpha_H}{\alpha_H+\alpha_L} \end{cases}, \qquad Y^{arrival} = \begin{cases} H & \text{w.p. } \frac{\alpha_L\lambda_H}{\alpha_H\lambda_L+\alpha_L\lambda_H} \\ L & \text{w.p. } \frac{\alpha_H\lambda_L}{\alpha_H\lambda_L+\alpha_L\lambda_H} \end{cases}. \tag{3}$$

3.2   Drift method

A key idea for analyzing queueing systems is that for any stationary random variable, its rate of increase must equal its rate of decrease. For instance, consider the work W in an M/G/1. Using the fact that

5

the rate of increase due to arriving jobs must equal the rate of decrease due to service, one can prove that ρ, the fraction of time the server is occupied, must equal $\lambda \mathbb{E}[S]$, the product of the arrival rate and the mean job size. This balance holds for any stationary random variable. In the M/G/1, by analyzing the rate of increase and decrease of W^2, one can characterize the mean work in the system.

To formalize this concept and apply it to our system, we make use of the *drift* of a random variable. In a given state, the random variable's drift is its instantaneous rate of change, in expectation over the randomness of the system. The key fact about drift is that in stationarity, the expected drift of any random variable is 0, where the expectation is taken both over the state and the future randomness of the system.

Formally, this is captured by the *instantaneous generator* G, which is the stochastic equivalent of the derivative operator. Rather than operating on a random variable directly, it is simpler to think of G as operating on a *test function*, a function f that maps system states (q, y) to real values. The instantaneous generator takes f and outputs $G \circ f$, the drift of f.

Let G specifically denote the generator operator for the two-level arrival system, which is defined as follows: For any test function $f(q, y)$,

$$G \circ f(q, y) := \lim_{t \to 0} \frac{1}{t} \mathbb{E}[f(Q(t), Y(t)) - f(q, y) \mid Q(0) = q, Y(0) = y].$$

Now, we can state the key fact about the drift:

Lemma 3.1. *Let f be a real-valued function of the MAMS system state where $\mathbb{E}[f(Q, Y_A, Y_C)] < \infty$. Then*

$$\mathbb{E}[G \circ f(Q, Y)] = 0, \tag{4}$$

where the expectation is taken over the stationary random variables Q, Y.

This result is a special case of the more general result for the MAMS system, Lemma 4.1, which we prove in Appendix C. The *drift method* relies on finding a useful test function f, such that the fact that the expected drift is zero allows us to analyze the performance of the system.

3.3 Relative arrivals

Next, we introduce a concept named the *relative arrival* function, $\Delta(i)$, which will allow us to define our needed test function for the drift method. This function was used in [8] to analyze the M/G/1 with Markov-modulated arrivals, where it was referred to as a_i.

Let $A_y(t)$ denote the number of arrivals by time t given that the arrival process starts in state $y \in \{H, L\}$.

Definition 3.1. The relative arrivals function $\Delta(y)$ is defined as follows:

$$\Delta(H) := \lim_{t \to \infty} \mathbb{E}[A_H(t)] - \lambda t, \qquad \Delta(L) := \lim_{t \to \infty} \mathbb{E}[A_L(t)] - \lambda t.$$

As we illustrate in Fig. 3, the relative arrivals function, $\Delta(y)$, represents the difference between the expected number of arrivals given that we're starting in state y and the expected total number of arrivals. As time goes to infinity, this difference converges to a finite value.

$\Delta(H)$ and $\Delta(L)$ are related as follows:

$$\Delta(H) = \frac{\lambda_H - \lambda}{\alpha_H} + \Delta(L), \qquad \Delta(L) = \frac{\lambda_L - \lambda}{\alpha_L} + \Delta(H). \tag{5}$$

To see why (5) holds, note that $\lambda_H - \lambda$ is the difference in arrival rates between H and the long-term rate, and $\frac{1}{\alpha_H}$ is the time for which that difference accrues. After this time, the system transitions to state L.
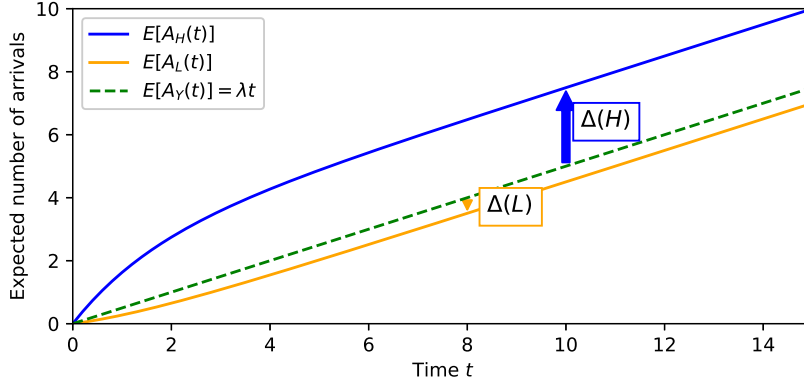
Figure 3: Expected arrivals by time t, starting in arrival states H and L, and in the steady-state Y. Relative arrivals $\Delta(H)$ and $\Delta(L)$ are the limiting differences between these functions. Setting: $\lambda_H = 2, \lambda_L = 0.2, \alpha_H = 0.5, \alpha_L = 0.1, \mu = 1, \lambda = 0.5$.

To characterize the exact values of $\Delta(H)$ and $\Delta(L)$, we can use the fact that the time-average value of Δ is zero: $\mathbb{E}[\Delta(Y)] = 0$, which we prove for the more general MAMS system in Lemma 4.3. As a result:

$$\Delta(H) = \alpha_H \frac{\lambda_H - \lambda_L}{(\alpha_H + \alpha_L)^2}, \qquad \Delta(L) = \alpha_L \frac{\lambda_L - \lambda_H}{(\alpha_H + \alpha_L)^2}. \tag{6}$$

The drift of the relative arrivals function is nicely behaved: $G \circ \Delta(y) = \lambda - \lambda_y$. This follows from Definition 3.1: The $\mathbb{E}[A_y(t)]$ term contributes λ_y to the drift, while the λt term contributes λ. This property is generalized to the MAMS system and proven in Corollary B.1.

3.4 Developing our test function

We come up with a test function $f_\Delta(q, y)$ for the drift method that combines both the queue length and the relative arrivals. Although the relative arrival function is not new, our test function is novel.

Definition 3.2. $f_\Delta(q, y) := (q + \Delta(y))^2 - \Delta(y)^2 = q(q + 2\Delta(y))$.

The intuition for the use of the test function $f_\Delta(q, y)$ is that on its own, the queue length q changes at a state-dependent rate, which is hard to analyze. Specifically, jobs arrive at the rate λ_y and complete at the rate μ, so the queue length q has drift $\lambda_y - \mu$, assuming $q > 0$.

We want to "smooth out" the instantaneous rate of change of q. Specifically, we want a function which:

- has a constant *drift*, in all system states where $q > 0$, and

- differs from q by a bounded amount.

Recall from Section 3.3 that the drift of $\Delta(y)$ is $\lambda - \lambda_y$. Thus, the function $q + \Delta(y)$ has a constant drift of $\lambda - \mu$ in all system states where $q > 0$.

Thus, $q + \Delta(y)$ is a smoothed-out proxy for the queue length q with constant drift, making it tractable for analysis. We set $f_\Delta(q, y)$ to be a smoothed-out proxy for the *square* of the queue length, as applying the generator G gives information about the expected derivative of the test function, and hence information about $\mathbb{E}[Q]$. Specifically, as we show in Lemma 3.3, the drift of $f_\Delta(q, y)$ separates cleanly: It has a $(\lambda - \mu)q$ term, and terms that do not depend on q. Then, in Theorem 3.1, we apply the fact that the drift is zero in stationarity (Lemma 3.1), allowing us to characterize $\mathbb{E}[Q]$.

7

3.5 Lemmas for the two-level system

First, let's calculate the drift of a generic test function in the two-level system.

Lemma 3.2. *For any test function f in the two-level system, its instantaneous drift in a particular system state (q, y) can be expressed as*

$$G \circ f(q, y) = \lambda_y(f(q+1, y) - f(q, y)) + \alpha_y(f(q, \bar{y}) - f(q, y)) + \mu(f(q-1+u, y) - f(q, y)),$$

where $u := (1-q)^+ = \mathbb{1}\{q = 0\}$ denotes unused service, and $\bar{y}$ is the opposite arrival state from y.

Proof. In a given state of the two-level system (q, y), there are three events that can occur: arrival, change of the arrival state y, and completion, if $q > 0$. These three events each contribute a corresponding term to the drift of the generic test function f. □

Now, we can calculate the drift for the special test function $f_\Delta(q, y) = q(q + 2\Delta(y))$.

Lemma 3.3. *For any system state (q, y) of the two-level system, the instantaneous drift is*

$$G \circ f_\Delta(q, y) = 2(\lambda - \mu)q + 2(\lambda_y - \mu(1-u))\Delta(y) + \lambda_y + \mu(1-u). \tag{7}$$

Proof. By Lemma 3.2 and the definition of $f_\Delta(q, y)$,

$$G \circ f_\Delta(q, y) = \lambda_y\big((q+1)(q+1+2\Delta(y)) - q(q+2\Delta(y))\big) \tag{8}$$
$$+ \alpha_y\big(q(q+2\Delta(\bar{y})) - q(q+2\Delta(y))\big) \tag{9}$$
$$+ \mu\big((q-1+u)(q-1+u+2\Delta(y)) - q(q+2\Delta(y))\big). \tag{10}$$

Let us simplify each term in the expression. Each term will simplify to an affine function of q: A linear term in q, plus an additive offset. The slope of some of these affine functions will depend on y, but we will show that for the overall drift, the slope does not depend on y, due to our choice of test function.

For the arrivals term, (8),

$$\lambda_y\big((q+1)(q+1+2\Delta(y)) - q(q+2\Delta(y))\big) = 2\lambda_y q + \lambda_y\big(2\Delta(y) + 1\big).$$

For the switching term, (9),

$$\alpha_y\big(q(q+2\Delta(\bar{y})) - q(q+2\Delta(y))\big) = 2\alpha_y q(\Delta_A(\bar{y}) - \Delta(y)).$$

For the completions term, (10),

$$\mu\big((q-1+u)(q-1+u+2\Delta(y)) - q(q+2\Delta(y))\big) = -2\mu q + \mu(1-u)\big(-2\Delta(y) + 1\big),$$

where we make use of the facts that $qu = 0$ and that $(1-u)^2 = (1-u)$, noting that $(1-u) \in \{0, 1\}$.

Combining the above calculations, we find that

$$G \circ f_\Delta(q, y) = 2(\lambda_y + \alpha_y(\Delta(\bar{y}) - \Delta(y)) - \mu)q \tag{11}$$
$$+ 2(\lambda_y - \mu(1-u))\Delta(y) + \lambda_y + \mu(1-u).$$

Now, we want to simplify the $\lambda_y + \alpha_y(\Delta(\bar{y}) - \Delta(y))$ coefficient in (11). To do so, let us use a relationship demonstrated in Section 3.3, namely (5):

$$\Delta(y) = \frac{\lambda_y - \lambda}{\alpha_y} + \Delta(\bar{y}) \implies \lambda = \lambda_y + \alpha_y(\Delta(\bar{y}) - \Delta(y)).$$

Thus, the linear term in (11) simplifies to a term which does not depend on y, as desired:

$$G \circ f_\Delta(q, y) = 2(\lambda - \mu)q + 2(\lambda_y - \mu(1-u))\Delta(y) + \lambda_y + \mu(1-u). \qquad □$$

3.6 Main result for two-level system

Theorem 3.1. *In the two-level arrivals system with exponential service rate, the mean queue length is:*

$$\mathbb{E}[Q] = \frac{\rho(\mathbb{E}[\Delta(Y^{arrival})] + 1)}{1 - \rho} + \mathbb{E}[\Delta(Y) \mid Q = 0], \text{ where} \tag{12}$$

$$\mathbb{E}[\Delta(Y^{arrival})] = \frac{(\lambda_H - \lambda_L)^2 \alpha_L \alpha_H}{\lambda(\alpha_H + \alpha_L)^3},$$

$$\mathbb{E}[\Delta(Y) \mid Q = 0] = \frac{\lambda_H - \lambda_L}{\alpha_H + \alpha_L} \left(P(Y = H \mid Q = 0) - \frac{\alpha_L}{\alpha_H + \alpha_L} \right).$$

Proof. We start with the result of Lemma 3.3:

$$G \circ f_\Delta(q, y) = 2(\lambda - \mu)q + 2(\lambda_y - \mu(1 - u))\Delta(y) + \lambda_y + \mu(1 - u). \tag{13}$$

Recall Lemma 3.1, the key drift-method lemma, which states that for (essentially) any test function, $f(q, y)$, we have $\mathbb{E}[G \circ f(Q, Y)] = 0$, where (Q, Y) is the stationary random variable for the system state. In particular, $\mathbb{E}[G \circ f_\Delta(Q, Y)] = 0$. We will use this fact to characterize $\mathbb{E}[Q]$. More specifically, because $f_\Delta(q, y)$ grows polynomially in q, we show it satisfies Lemma 3.1 in Lemma C.2.

Taking the expectation of (13), we find that

$$0 = \mathbb{E}[G \circ f_\Delta(Q, Y)]$$
$$= 2(\lambda - \mu)\mathbb{E}[Q] + 2\mathbb{E}_{y \sim Y}[\lambda_y \Delta(y)] - 2\mu\mathbb{E}[\Delta(y)] + 2\mu\mathbb{E}[\Delta(Y)\mathbb{1}\{Q = 0\}] + \lambda + \mu P(Q > 0)$$
$$= 2(\lambda - \mu)\mathbb{E}[Q] + 2\mathbb{E}_{y \sim Y}[\lambda_y \Delta(y)] + 2\mu\mathbb{E}[\Delta(Y)\mathbb{1}\{Q = 0\}] + \lambda + \mu P(Q > 0). \tag{14}$$

In taking this expectation, we use the facts that $\lambda := \mathbb{E}[\lambda_Y]$, that $\mathbb{E}[\Delta(Y)] = 0$, and that $u := \mathbb{1}\{q = 0\}$. Note that the $(\lambda - \mu)q$ term in (13) allows us to get a $\mathbb{E}[Q]$ term in (14).

To simplify (14), note that because the system has an exponential service rate, $P(Q > 0) = \rho = \frac{\lambda}{\mu}$, by Lemma A.1. Thus:

$$0 = (\lambda - \mu)\mathbb{E}[Q] + \mathbb{E}_{y \sim Y}[\lambda_y \Delta(y)] + (\mu - \lambda)\mathbb{E}[\Delta(Y) \mid Q = 0] + \lambda. \tag{15}$$

Recall that we defined $Y^{arrival}$ to be the arrival-weighted state of the arrival process. In particular, from (2), $P(Y^{arrival} = y) := P(Y = y)\frac{\lambda_y}{\lambda}$. As a result, $\mathbb{E}_{y \sim Y}[\lambda_y \Delta(y)] = \lambda\mathbb{E}[\Delta(Y^{arrival})]$, and (15) becomes:

$$0 = (\lambda - \mu)\mathbb{E}[Q] + \lambda\mathbb{E}[\Delta(Y^{arrival})] + (\mu - \lambda)\mathbb{E}[\Delta(Y) \mid Q = 0] + \lambda$$
$$\mathbb{E}[Q] = \frac{\rho(\mathbb{E}[\Delta(Y^{arrival})] + 1)}{1 - \rho} + \mathbb{E}[\Delta(Y) \mid Q = 0].$$

Using the values of Δ from (6) and the distribution for $Y^{arrival}$ from (3), we can compute $\mathbb{E}[\Delta(Y^{arrival})]$ and compute $\mathbb{E}[\Delta(Y) \mid Q = 0]$ up to the value of $P(Y = H \mid Q = 0)$, which we bound in Section 7. □

Note that plugging in the bound $P(Y = H \mid Q = 0) > 0$ immediately gives a strong lower bound on $\mathbb{E}[Q]$. In Section 7, we prove tight upper bounds on $P(Y = H \mid Q = 0)$ in both the quick switching limit $(\alpha_H, \alpha_L \to \infty)$ and in the slow switching limit $(\alpha_H, \alpha_L \to 0)$, giving tight results in both regimes.

Even without any bounds on $P(Y = H \mid Q = 0)$, Theorem 3.1 already tightly characterizes mean response time in the $\rho \to 1$ limit (the heavy traffic limit), holding α_H and α_L constant:

Corollary 3.1. *In the two-level arrivals system, the scaled mean queue length in the heavy traffic limit converges to* $\lim_{\rho \to 1}(1 - \rho)\mathbb{E}[Q] = \mathbb{E}[\Delta(Y^{arrival})] + 1$.

This holds because $\mathbb{E}[\Delta(Y) \mid Q = 0]$ is bounded by $\Delta(L)$ and $\Delta(H)$, which are bounded in the $\rho \to 1$ limit.

In the more general setting of Markovian arrivals and Markovian service (MAMS), we prove Theorem 5.1, a generalization of our two-level result Theorem 3.1, which can handle any number of arrival rates, as well as Markov modulated service rates. This results in a similar heavy traffic result, Corollary 5.2.

4 MAMS model and definitions

We now turn to the general setting of Markovian arrivals and Markovian service (MAMS), where the arrival process and the completions process are both Markov chains.

Like the two-level arrival system introduced in Section 3.1, the MAMS system consists of a single-server queue. Jobs wait in the queue in FCFS order. When a job reaches the head of the queue, it enters service. After some time in service, the job completes. All jobs are indistinguishable. Jobs arrive according to an arrival process which we define in Section 4.1, and complete according to a completion process which we define in Section 4.2. After introducing the MAMS model, Sections 4.3 to 4.5 generalize our discussion of the drift process, the notion of relative arrivals (and completions), and the test function from what we saw in Section 3. The main results for the MAMS model are stated in Section 5.

Our goal is to bound the mean queue length in the MAMS system, with bounds that are tight in heavy traffic. We seek to generalize Theorem 3.1, our main result for the two-level arrivals system.

4.1 Arrivals

We start by generalizing the arrival process to be a general finite-state continuous-time Markov chain. In Section 3.1, the arrival process had only two states, high arrival intensity and low arrival intensity. There were two types of updates in that arrival process: Jobs arriving, which happened with rate λ_H and λ_L, and state changes, which happened with rate α_H and α_L.

In the more general MAMS setting, we must track both the initial and final state of a state change, and we also allow for a job to arrive simultaneously with the state change.

We thus use the following more general notation: Let the state space of the arrival chain be some finite set $\mathbb{A}$. For a given pair of states $i \neq j \in \mathbb{A}$, the system may transition from i to j with an arrival, or from i to j without an arrival. We will denote the rate of the former transition as $r_{i,j,1}$, and of the later transition as $r_{i,j,0}$. In general, we will write the transition rate as $r_{i,j,a}$, where the a parameter indicates whether an arrival occurs during a transition. As a special case, self-transitions may only occur with an arrival: $r_{i,i,1}$ may be nonzero, $r_{i,i,0}$ is always zero.

For example, in the two-level system of Section 3, $\mathbb{A} = \{H, L\}$, and the nonzero transition rates are:

$$r_{H,H,1} = \lambda_H, \quad r_{L,L,1} = \lambda_L, \quad r_{H,L,0} = \alpha_H, \quad r_{L,H,0} = \alpha_L.$$

We assume that the arrival chain is irreducible. Let λ be the long-term arrival rate, as in Section 3.1. Let $Y_A(t)$ be a random variable denoting the state of the arrival chain at time t. Let Y_A be the corresponding stationary random variable. Let $Y_A^{arrival}$ be the arrival-rate-weighted arrival state, defined as follows:

$$P(Y_A^{arrival} = j) := \frac{\sum_{i \in \mathbb{A}} P(Y_A = i) r_{i,j,1}}{\lambda}.$$

Note that $Y_A(t), Y_A$, and $Y_A^{arrival}$ correspond to $Y(t), Y$, and $Y^{arrival}$ from Section 3.1. The subscript $_A$ is added to differentiate the arrival chain from the completion chain, which will use the subscript $_C$.

We will use $\cdot$ in the rate subscripts when aggregating multiple related rates. For instance, $r_{i,\cdot,\cdot} := \sum_{j,a} r_{i,j,a}$. As a special case, we will define $\lambda_i = r_{i,\cdot,1}$ to be the total instantaneous arrival rate in state i.

4.2 Completions

In the two-level arrivals model in Section 3, completions always occurred at rate $Exp(\mu)$. In the general MAMS model, we allow the completion chain to be a general finite-state continuous-time Markov chain.

Paralleling the arrival chain defined in Section 4.1, we let the state space of the completion chain be some finite set $\mathbb{C}$. For each $i, j \in \mathbb{C}$, the system may transition from i to j with a completion or without a completion, at rates $s_{i,j,0}$ and $s_{i,j,1}$ respectively. As before, $s_{i,i,1}$ may be nonzero, but $s_{i,i,0} = 0$.

We write these rates as $s_{i,j,c}$, where the c parameter denotes whether a completion occurs during a transition. Again, we assume the completion chain is irreducible.

One important note: A completion transition may occur when there are no jobs present in the system. If this happens, the completion process's state change still occurs, and the queue remains empty. We will refer to such a completion as an "unused completion."

Let μ denote the long-term rate of completions, whether used or unused.

As an example, in the two-level system there was only a single completion state, which we may arbitrarily call X. Then $\mathbb{C} = \{X\}$. The sole completion transition rate was $s_{X,X,1} = \mu$.

We define $Y_C(t), Y_C$, and Y_C^{comp} to correspond to $Y_A(t), Y_A$, and $Y_A^{arrival}$ from Section 4.1. In particular,

$$P(Y_C^{comp} = j) := \frac{\sum_{i \in \mathbb{C}} P(Y_C = i) s_{i,j,1}}{\mu}.$$

We define the load ρ to be the ratio λ/μ. Our bounds hold for all loads ρ, and are tight in the heavy-traffic limit, as $\rho \to 1$.

In our main result for the two-level arrival system, Theorem 3.1, we make use of an expectation of the form $\mathbb{E}[\cdot \mid Q = 0]$, conditioning on the system state when the queue is empty. In the MAMS setting, we condition the system state when when unused completions occur. We will write this expectation as $\mathbb{E}_U[\cdot]$. Note that in the two-level system, because of the exponential service process and the PASTA property [31], these two expectations coincide: $\mathbb{E}_U[\cdot] = \mathbb{E}[\cdot \mid Q = 0]$. We define the expectation $\mathbb{E}_U[\cdot]$ in more detail in Appendix A.

As with the arrival process, we use $\cdot$ to aggregate related rates, and define $\mu_i = s_{i,\cdot,1}$ to correspond to λ_i.

4.3   Drift method

In the general MAMS model, we continue using the drift method which we discussed in Section 3.2, using the fact that the rate of change of any random variable is zero in steady state. Formally, the rate of change of a random variable is captured by the drift of a test function. Here a test function f is any real-valued function of the state of the MAMS system, (q, i_A, i_C), where q denotes the total number of jobs, and i_A and i_C denote the states of the arrival and service processes, respectively. The drift of the test function $G \circ f(q, i_A, i_C)$ is defined as

$$G \circ f(q, i_A, i_C) := \lim_{t \to 0} \frac{1}{t} \mathbb{E}[f(Q(t), Y_A(t), Y_C(t)) - f(q, i_A, i_C) \mid Q(0) = q, Y_A(0) = i_A, Y_C(0) = i_C].$$

The operator G is called the *instantaneous generator*, which takes a test function f as input and gives the drift of the test function $G \circ f$ as output. The instantaneous generator can be seen as the stochastic equivalent of the derivative operator.

The lemma below shows that the expected value of the $G \circ f$ in steady state is zero, generalizing Lemma 3.1 to the MAMS system. We prove Lemma 4.1 in Appendix C.

Lemma 4.1. *Let f be a test function for which $\mathbb{E}[f(Q, Y_A, Y_C)] < \infty$. Then*

$$\mathbb{E}[G \circ f(Q, Y_A, Y_C)] = 0. \tag{16}$$

All test functions f that we will use with (16) will grow at a polynomial rate in q, which we show satisfy Lemma 4.1 in Lemma C.2.

Similar to Lemma 3.2, Lemma 4.2 below provides the expression for the drift of a generic function in the MAMS system. Lemma 4.2 is standard in the drift method literature, and we have specialized it to this system. We prove Lemma 4.2 in Appendix C.

Lemma 4.2. *For any real-valued function f of the state of the MAMS system,*

$$G \circ f(q, i_A, i_C) = \sum_{j_A \in \mathbb{A}, a \in \{0,1\}} r_{i_A, j_A, a} \left(f(q + a, j_A, i_C) - f(q, i_A, i_C) \right)$$
$$+ \sum_{j_C \in \mathbb{C}, c \in \{0,1\}} s_{i_C, j_C, c} \left(f(q - c + u, i_A, j_C) - f(q, i_A, i_C) \right),$$

where $u = (c - q)^+$ denotes the unused service.

4.4 Relative arrivals and completions

Next, we introduce a concept named the *relative arrival* function, $\Delta(i)$, which will allow us to define our needed test function for the drift method.

In Section 3.3, we introduced the concept of the relative arrivals function, which was used in [8] to analyze the M/G/1 with Markov-modulated arrivals, where it was referred to as a_i. In the MAMS setting, we denote this relative arrivals function as $\Delta_A(i)$, and we also introduce the relative completions function $\Delta_C(i)$, defined equivalently.

Let $A_i(t)$ denote the number of arrivals by time t given that the arrival chain starts in state $i \in \mathbb{A}$ at time $t = 0$. Let $C_i(t)$ denote the number of completions by time t given that the completion chain starts in state $i \in \mathbb{C}$ at time $t = 0$.

Definition 4.1. Define the relative arrivals and relative completions functions $\Delta_A(i), \Delta_C(i)$ as follows:

$$\Delta_A(i) := \lim_{t \to \infty} \mathbb{E}[A_i(t)] - \lambda t, \qquad \Delta_C(i) := \lim_{t \to \infty} \mathbb{E}[C_i(t)] - \mu t.$$

We verify that these limits always converge to a finite value in Appendix B. These functions can be seen as the relative value function of the Markov reward process where a reward of 1 is received whenever an arrival or completion occurs, respectively. We can therefore calculate their values explicitly by solving the corresponding Poisson equation, given in Lemma B.1, along with Lemma 4.3, which states that these quantities have zero expected value in stationarity:

Lemma 4.3. $\mathbb{E}[\Delta_A(Y_A)] = 0$, $\mathbb{E}[\Delta_C(Y_C)] = 0$.

Proof. Note that $[Y_A(t) \mid Y_A(0) = Y_A] \sim Y_A$. As a result, $\mathbb{E}[A_{Y_A}(t)] = \lambda t$ for any t. Then by definition, $\mathbb{E}[\Delta_A(Y_A)] = \lim_{t \to \infty} (\mathbb{E}[A_{Y_A}(t)] - \lambda t) = 0$. Similarly, $\mathbb{E}[\Delta_C(Y_C)] = 0$. $\qquad \square$

Another important property of $\Delta_A(i)$ and $\Delta_C(i)$ are their drifts:

Corollary B.1. $G \circ \Delta_A(i_A) = \lambda - \lambda_{i_A}, G \circ \Delta_C(i_C) = \mu - \mu_{i_C}$.

Corollary B.1 follows from the Poisson equation, Lemma B.1, and it is proven in Appendix B.

4.5 Test function

In order to characterize $\mathbb{E}[Q]$, we introduce the test function $f_{\Delta_A, \Delta_C}(q, i_A, i_C)$, which generalizes the two-level arrival test function defined in Definition 3.2 to the MAMS system.

Definition 4.2. $f_{\Delta_A, \Delta_C}(q, i_A, i_C) := q(q + 2\Delta_A(i_A) - 2\Delta_C(i_C))$.

The intuition for this test function again comes from finding a smooth proxy for the queue length q. In the two-level system, we used $q + \Delta(y)$ as a proxy for q, to capture the influence of the state-dependent arrival rate in the two-level system, In the MAMS system, both the arrival rates and completion rates are state-dependent, so we use $q + \Delta_A(i_A) - \Delta_C(i_C)$ as a smooth proxy for q.

To characterize mean queue length, we need a proxy for the square of the queue length. In the two-level system, we used $q(q + 2\Delta(y))$ as our proxy for the square of queue length. In the MAMS system, $f_{\Delta_A, \Delta_C}(q, i_A, i_C)$ fulfils the same role.

12

5 Results

We now present the main of the paper: a closed-form formula for the mean queue length in the MAMS system. This will be proven in Section 6.

Theorem 5.1 (Mean queue length). *In the MAMS system, the mean queue length is given by:*

$$\mathbb{E}[Q] = \frac{\rho\mathbb{E}[\Delta_A(Y_A^{arrival})] + \mathbb{E}[\Delta_C(Y_C^{comp})] + \rho}{1 - \rho} + \mathbb{E}_U[\Delta_A(Y_A) - \Delta_C(Y_C^{comp})],$$

where $\mathbb{E}_U[\cdot]$ was defined in Section 4.2:

$$\mathbb{E}_U[\Delta_A(Y_A) - \Delta_C(Y_C^{comp}))] = \frac{\mathbb{E}[(\mu_{Y_C}\Delta_A(Y_A) - \sum_{j_C \in \mathbb{C}} s_{Y_C, j_C, 1}\Delta_C(j_C)))\mathbb{1}\{Q = 0\}]}{\mu - \lambda}.$$

Our formula consists of a fully-explicit primary term and a secondary term (the $\mathbb{E}_U[\cdot]$ term) which depends on the behavior of the MAMS system when the queue is empty, captured by the expectation over the unused service $\mathbb{E}_U[\cdot]$. Theorem 5.1 generalizes our two-level result, Theorem 3.1. The $\mathbb{E}[\Delta(Y) \mid Q = 0]$ term in Theorem 3.1 becomes the $\mathbb{E}_U[\cdot]$ term in Theorem 5.1, and terms are introduced for the Markovian service process. One can exactly compute the $\Delta_A(\cdot)$ and $\Delta_C(\cdot)$ functions by solving the corresponding Poisson equations, as discussed in Section 4.4.

As one consequence of Theorem 5.1, note that $\mathbb{E}_U[\Delta_A(Y_A) - \Delta_C(Y_C^{comp})]$ is bounded by the maximum and minimum possible values of Δ_A and Δ_C, yielding the following simple bounds:

Corollary 5.1. *Let $\Delta_A^{\max}$ and $\Delta_A^{\min}$ be the maximum and minimum values of $\Delta_A(i_A)$ over all arrival states i_A, and define $\Delta_C^{\max}$ and $\Delta_C^{\min}$ similarly. In the MAMS system, the mean queue length is bounded by*

$$\Delta_A^{\min} - \Delta_C^{\max} \le \mathbb{E}[Q] - \frac{\rho\Delta_A(Y_A^{arrival}) + \Delta_C(Y_C^{comp}) + \rho}{1 - \rho} \le \Delta_A^{\max} - \Delta_C^{\min}.$$

Proof. This corollary follows immediately from Theorem 5.1, once we recall that the arrival and completion state spaces $\mathbb{A}$ and $\mathbb{C}$ are finite, and that the relative arrivals and completions functions always converge, as discussed in Section 4.4. □

Note that these bounds converge in the $\rho \to 1$ limit, as $\Delta_A^{\max}, \Delta_A^{\min}, \Delta_C^{\max}$, and $\Delta_C^{\min}$ remain bounded in that limit, allowing us to exactly characterize the heavy traffic behavior of the system:

Corollary 5.2. *In the MAMS system, the scaled mean queue length in the heavy traffic limit converges:*

$$\lim_{\rho \to 1}(1 - \rho)\mathbb{E}[Q] = \mathbb{E}[\Delta_A(Y_A^{arrival})] + \mathbb{E}[\Delta_C(Y_C^{comp})] + 1.$$

6 Proofs for MAMS system

We will follow a similar structure to that used in Section 3:

- Lemma 6.1 sets up the formula for the drift of our chosen test function f_{Δ_A, Δ_C}, which we introduced in Section 4.5, splitting the drift into an arrivals term and a completions term.

- Lemmas 6.2 and 6.4 simplify the arrivals and completions drift terms.

- Lemmas 6.3 and 6.5 compute the steady-state expectations of the arrivals and completions drift terms.

- In Section 6.2, we prove Theorem 5.1 using our expected drift results and the fact that expected drift in steady state is zero to characterize mean queue length $\mathbb{E}[Q]$.

Lemmas 6.2 and 6.4 collectively generalize Lemma 3.3 from the two-level system to the MAMS system, while Lemmas 6.3 and 6.5 and Theorem 5.1 collectively generalize Theorem 3.1.

6.1 Key lemmas

We start by deriving an initial formula for the drift of our test function f_{Δ_A, Δ_C}, which we introduced in Section 4.5. We separate this formula into two terms: An arrivals term g_A and a completions term g_C.

Lemma 6.1 (Formula for drift of f_{Δ_A, Δ_C}). *For any MAMS system and any system state (q, i_A, i_C),*

$$G \circ f_{\Delta_A, \Delta_C}(q, i_A, i_C) = g_A(q, i_A, i_C) + g_C(q, i_A, i_C),$$

where $g_A(q, i_A, i_C) :=$ \hfill (17)

$$\sum_{j_A \in \mathbb{A}, a} r_{i_A, j_A, a} \left((q+a)(q + a + 2\Delta_A(j_A) - 2\Delta_C(i_C)) - q(q + 2\Delta_A(i_A) - 2\Delta_C(i_C)) \right),$$

and where $g_C(q, i_A, i_C) :=$ \hfill (18)

$$\sum_{j_C \in \mathbb{C}, c} s_{i_C, j_C, c} \left((q - c + u)(q - c + u + 2\Delta_A(i_A) - 2\Delta_C(j_C)) - q(q + 2\Delta_A(i_A) - 2\Delta_C(i_C)) \right).$$

Proof. Follows immediately from Lemma 4.2. $\qquad\square$

Now, we simplify the g_A term.

Lemma 6.2. *For any MAMS system and any system state (q, i_A, i_C),*

$$g_A(q, i_A, i_C) = 2\lambda q + (1 - 2\Delta_C(i_C))\lambda_{i_A} + 2\sum_{j_A \in \mathbb{A}} r_{i_A, j_A, 1}\Delta_A(j_A). \tag{19}$$

Proof. We start by simplifying $g_A(q, i_A, i_C)$. Recall the definition of g_A, (17):

$$g_A(q, i_A, i_C) = \sum_{j_A, a} r_{i_A, j_A, a} \left((q+a)(q + a + 2\Delta_A(j_A) - 2\Delta_C(i_C)) - q(q + 2\Delta_A(i_A) - 2\Delta_C(i_C)) \right). \tag{20}$$

Let us simplify the bracketed term inside the summation:

$$(q+a)(q + a + 2\Delta_A(j_A) - 2\Delta_C(i_C)) - q(q + 2\Delta_A(i_A) - 2\Delta_C(i_C))$$
$$= 2q(a + \Delta_A(j_A) - \Delta_A(i_A)) + a(a - 2\Delta_C(i_C) + 2\Delta_A(j_A)). \tag{21}$$

Let us substitute (21) back into (20):

$$g_A(q, i_A, i_C) = 2q \sum_{j_A, a} r_{i_A, j_A, a}(a + \Delta_A(j_A) - \Delta_A(i_A)) \tag{22}$$

$$+ \sum_{j_A, a} r_{i_A, j_A, a} a(a - 2\Delta_C(i_C) + 2\Delta_A(j_A)). \tag{23}$$

We will separately simplify (22) and (23).

For (22), we prove in Corollary B.1 that for any initial state i_A,

$$\sum_{j_A, a} r_{i_A, j_A, a}(a + \Delta_A(j_A) - \Delta_A(i_A)) = \lambda.$$

This follows from the definition of the relative arrivals function Δ_A.

For (23), note that $a \in \{0, 1\}$:

$$\sum_{j_A, a} r_{i_A, j_A, a} a(a - 2\Delta_C(i_C) + 2\Delta_A(j_A)) = (1 - 2\Delta_C(i_C))\lambda_{i_A} + 2\sum_{j_A} r_{i_A, j_A, 1}\Delta_A(j_A).$$

Combining our simplifications, we find that

$$g_A(q, i_A, i_C) = 2\lambda q + (1 - 2\Delta_C(i_C))\lambda_{i_A} + 2\sum_{j_A} r_{i_A, j_A, 1}\Delta_A(j_A). \qquad\square$$

Now, we characterize the expectation of g_A over the steady state of the MAMS system.

Lemma 6.3 (Mean of arrivals term g_A). *For any MAMS system,*

$$\mathbb{E}[g_A(Q, Y_A, Y_C)] = \lambda(2\mathbb{E}[Q] + 2\mathbb{E}[\Delta_A(Y_A^{arrival})] + 1). \tag{24}$$

Proof. Let us begin with the simplified expression for g_A from Lemma 6.2:

$$g_A(q, i_A, i_C) = 2\lambda q + (1 - 2\Delta_C(i_C))\lambda_{i_A} + 2\sum_{j_A} r_{i_A, j_A, 1}\Delta_A(j_A).$$

$$\mathbb{E}[g_A(Q, Y_A, Y_C)] = 2\lambda\mathbb{E}[Q] + \mathbb{E}[(1 - 2\Delta_C(Y_C))\lambda_{Y_A}] \tag{25}$$

$$+ 2\sum_{i_A, j_A} P(Y_A = i_A)r_{i_A, j_A, 1}\Delta_A(j_A). \tag{26}$$

We simplify (25) and (26) separately. For (25), note that Y_A and Y_C are independent, as the arrival process and completions process update independently. Note also that $\mathbb{E}[\Delta_C(Y_C)] = 0$, as shown in Lemma 4.3, and that $\mathbb{E}[\lambda_{Y_A}] = \lambda$, which is the definition of λ. Thus, we can simplify the second term of (25):

$$\mathbb{E}[(1 - 2\Delta_C(Y_C))\lambda_{Y_A}] = \mathbb{E}[(1 - 2\Delta_C(Y_C))]\mathbb{E}[\lambda_{Y_A}] = \lambda. \tag{27}$$

As for (26), recall from Section 4.1 the definition of $Y_A^{arrival}$, the arrival-rate-weighted arrival state:

$$P(Y_A^{arrival} = j) := \frac{\sum_{i \in \mathbb{A}} P(Y_A = i)r_{i,j,1}}{\lambda}.$$

As a result, we can simplify (26):

$$\sum_{i_A, j_A} P(Y_A = i_A)r_{i_A, j_A, 1}\Delta_A(j_A) = \lambda\sum_{j_A} P(Y_A^{arrival} = j_A)\Delta_A(j_A) = \lambda\mathbb{E}[\Delta_A(Y_A^{arrival})]. \tag{28}$$

Substituting (27) and (28) into (25), we find that

$$\mathbb{E}[g_A(Q, Y_A, Y_C)] = \lambda(2\mathbb{E}[Q] + 2\mathbb{E}[\Delta_A(Y_A^{arrival})] + 1). \qquad \square$$

Now, we switch to the completions term, g_C. We start by simplifying g_C. Lemma 6.4 mirrors Lemma 6.2, but with some additional complications due to unused service when $q = 0$.

Lemma 6.4 (Simplification of the completion term g_C). *For any MAMS system and any system state* (q, i_A, i_C),

$$g_C(q, i_A, i_C) = -2\mu q + \left((1 - 2\Delta_A(i_A))\mu_{i_C} + 2\sum_{j_C \in \mathbb{C}} s_{i_C, j_C, 1}\Delta_C(j_C)\right)\mathbb{1}\{q > 0\}. \tag{29}$$

Proof. Recall the definition of $g_C(q, i_A, i_C)$ from Lemma 4.2:

$$g_C(q, i_A, i_C) := \sum_{j_C, c} s_{i_C, j_C, c}\left((q - c + u)(q - c + u + 2\Delta_A(i_A) - 2\Delta_C(j_C)) - q(q + 2\Delta_A(i_A) - 2\Delta_C(i_C))\right),$$

where $u := (c - q)^+$ is the amount of unused service caused by a specific transition. Note that $u \in \{0, 1\}$.

Let us simplify the bracketed term inside the summation:

$$(q - c + u)(q - c + u + 2\Delta_A(i_A) - 2\Delta_C(j_C)) - q(q + 2\Delta_A(i_A) - 2\Delta_C(i_C))$$

$$= 2q(-c + u - \Delta_C(j_C) + \Delta_C(i_C)) + (-c + u)(-c + u + 2\Delta_A(i_A) - 2\Delta_C(j_C))$$

$$= -2q(c + \Delta_C(j_C) - \Delta_C(i_C)) + c(c - 2\Delta_A(i_A) + 2\Delta_C(j_C)) + u(2q - 2c + u + 2\Delta_A(i_A) - 2\Delta_C(j_C)). \tag{30}$$

15

We substitute (30) back to the definition of $g_C(q, i_A, i_C)$ and get

$$g_C(q, i_A, i_C) = -2q \sum_{j_C,c} s_{i_C,j_C,c}(c + \Delta_C(j_C) - \Delta_C(i_C)) \tag{31}$$

$$+ \sum_{j_C,c} s_{i_C,j_C,c} c(c - 2\Delta_A(i_A) + 2\Delta_C(j_C)) \tag{32}$$

$$+ \sum_{j_C,c} s_{i_C,j_C,c} u(2q - 2c + u + 2\Delta_A(i_A) - 2\Delta_C(j_C)). \tag{33}$$

For the term in (31), we prove in Corollary B.1 that

$$\sum_{j_C,c} s_{i_C,j_C,c}(c + \Delta_C(j_C) - \Delta_C(i_C)) = \mu.$$

so the term in (31) can be simplified as

$$-2q \sum_{j_C,c} s_{i_C,j_C,c}(c + \Delta_C(j_C) - \Delta_C(i_C)) = -2\mu q.$$

For the term in (32), note that $c \in \{0, 1\}$, so

$$\sum_{j_C,c} s_{i_C,j_C,c} c(c - 2\Delta_A(i_A) + 2\Delta_C(j_C)) = \sum_{j_C} s_{i_C,j_C,1}(1 - 2\Delta_A(i_A) + 2\Delta_C(j_C))$$

$$= (1 - 2\Delta_A(i_A))\mu_{i_C} + 2 \sum_{j_C} s_{i_C,j_C,1}\Delta_C(j_C).$$

For the term in (33), recall that $u = (c - q)^+$, so $u \in \{0, 1\}$, and $u = 1$ if and only if $c = 1$ and $q = 0$. We thus simplify the term in (33) as

$$\sum_{j_C,c} s_{i_C,j_C,c} u(2q - 2c + u + 2\Delta_A(i_A) - 2\Delta_C(j_C))$$

$$= \sum_{j_C} s_{i_C,j_C,1}(-1 + 2\Delta_A(i_A) - 2\Delta_C(j_C))\mathbb{1}\{q = 0\}$$

$$= -\left((1 - 2\Delta_A(i_A))\mu_{i_C} + 2 \sum_{j_C} s_{i_C,j_C,1}\Delta_C(j_C)\right)\mathbb{1}\{q = 0\}.$$

Combining our simplifications, we find that

$$g_C(q, i_A, i_C) = -2\mu q + \left((1 - 2\Delta_A(i_A))\mu_{i_C} + 2 \sum_{j_C} s_{i_C,j_C,1}\Delta_C(j_C)\right)\mathbb{1}\{q > 0\}. \qquad \square$$

Lemma 6.5 (Mean of completions term g_C). *For any MAMS system,*

$$\mathbb{E}[g_C(Q, Y_A, Y_C)] = -2\mu\mathbb{E}[Q] + \lambda + 2\mu\mathbb{E}[\Delta_C(Y_C^{comp})] + 2\mathbb{E}[h_{C0}(Y_A, Y_C)\mathbb{1}\{Q = 0\}],$$

where h_{C0} is defined as follows:

$$h_{C0}(i_A, i_C) = \mu_{i_C}\Delta_A(i_A) - \sum_{j_C \in \mathbb{C}} s_{i_C,j_C,1}\Delta_C(j_C),$$

$$\mathbb{E}[h_{C0}(Y_A, Y_C)\mathbb{1}\{Q = 0\}] = (\mu - \lambda)\mathbb{E}_U[\Delta_A(Y_A) - \Delta_C(Y_C^{comp})].$$

16

Proof. By the simplified formula of $g_C(q, i_A, i_C)$ in Lemma 6.4, we have

$$\mathbb{E}[g_C(Q, Y_A, Y_C)] = -2\mu\mathbb{E}[Q] + \mathbb{E}[((1 - 2\Delta_A(Y_A))\mu_{Y_C} + 2\sum_{j_C} s_{Y_C, j_C, 1}\Delta_C(j_C))\mathbb{1}\{Q > 0\}]. \qquad (34)$$

We can decompose the second term of (34) using the fact that $\mathbb{1}\{Q > 0\} = 1 - \mathbb{1}\{Q = 0\}$. Observe that

$$\mathbb{E}[(1 - 2\Delta_A(Y_A))\mu_{Y_C} + 2\sum_{j_C} s_{Y_C, j_C, 1}\Delta_C(j_C)] = \mathbb{E}[1 - 2\Delta_A(Y_A)]\mathbb{E}[\mu_{Y_C}] + 2\mu\mathbb{E}[\Delta_C(Y_C^{comp})]$$

$$= \mu + 2\mu\mathbb{E}[\Delta_C(Y_C^{comp})],$$

where the first equality uses the independence of Y_A and Y_C, and the definition of Y_C^{comp}; and the second equality is due to $\mathbb{E}[\mu_{Y_C}] = \mu$ and $\mathbb{E}[\Delta_A(Y_A)] = 0$. Next, we deal with the term

$$\mathbb{E}[((1 - 2\Delta_A(Y_A))\mu_{Y_C} + 2\sum_{j_C} s_{Y_C, j_C, 1}\Delta_C(j_C))\mathbb{1}\{Q = 0\}]$$

$$= \mathbb{E}[\mu_{Y_C}\mathbb{1}\{Q = 0\}] + \mathbb{E}[(-2\Delta_A(Y_A)\mu_{Y_C} + 2\sum_{j_C} s_{Y_C, j_C, 1}\Delta_C(j_C))\mathbb{1}\{Q = 0\}]$$

$$= \mu - \lambda - 2\mathbb{E}[h_{C0}(Y_A, Y_C)\mathbb{1}\{Q = 0\}],$$

where the last equality is due to $\mathbb{E}[\mu_{Y_C}\mathbb{1}\{Q = 0\}] = \mu - \lambda$, which we prove in Lemma A.1. Plugging the above calculations into (34), we get

$$\mathbb{E}[g_C(Q, Y_A, Y_C)] = -2\mu\mathbb{E}[Q] + \mu + 2\mu\mathbb{E}[\Delta_C(Y_C^{comp})] - \mu + \lambda + 2\mathbb{E}[h_{C0}(Y_A, Y_C)\mathbb{1}\{Q = 0\}]$$

$$= -2\mu\mathbb{E}[Q] + \lambda + 2\mu\mathbb{E}[\Delta_C(Y_C^{comp})] + 2\mathbb{E}[h_{C0}(Y_A, Y_C)\mathbb{1}\{Q = 0\}].$$

Equivalently, we can write the last term $\mathbb{E}[h_{C0}(Y_A, Y_C)\mathbb{1}\{Q = 0\}]$ in terms of $\mathbb{E}_U[\cdot]$:

$$\mathbb{E}[h_{C0}(Y_A, Y_C)\mathbb{1}\{Q = 0\}] = \mathbb{E}[\mu_{Y_C}\Delta_A(Y_A)\mathbb{1}\{Q = 0\}] - \mathbb{E}[\sum_{j_C} s_{Y_C, j_C, 1}\Delta_C(j_C))\mathbb{1}\{Q = 0\}]$$

$$= (\mu - \lambda)\mathbb{E}_U[\Delta_A(Y_A)] - (\mu - \lambda)\mathbb{E}_U[\Delta_C(Y_C^{comp})]$$

$$= (\mu - \lambda)\mathbb{E}_U[\Delta_A(Y_A) - \Delta_C(Y_C^{comp})]. \qquad \square$$

6.2   Proof of main result: Theorem 5.1

Theorem 5.1. *In the MAMS system, the mean queue length is given by:*

$$\mathbb{E}[Q] = \frac{\rho\mathbb{E}[\Delta_A(Y_A^{arrival})] + \mathbb{E}[\Delta_C(Y_C^{comp})] + \rho}{1 - \rho} + \mathbb{E}_U[\Delta_A(Y_A) - \Delta_C(Y_C^{comp})],$$

where recall the definition of $\mathbb{E}_U[\cdot]$ in Section 4.2 that

$$\mathbb{E}_U[\Delta_A(Y_A) - \Delta_C(Y_C^{comp}))] = \frac{\mathbb{E}[(\mu_{Y_C}\Delta_A(Y_A) - \sum_{j_C \in \mathbb{C}} s_{Y_C, j_C, 1}\Delta_C(j_C)))\mathbb{1}\{Q = 0\}]}{\mu - \lambda}.$$

Proof. Recall Lemma 6.1, $G \circ f_{\Delta_A, \Delta_C}(q, i_A, i_C) = g_A(q, i_A, i_C) + g_C(q, i_A, i_C)$. Because $\mathbb{E}[G \circ f_{\Delta_A, \Delta_C}(Q, Y_A, Y_C)] = 0$ by Lemma 4.1, we have

$$\mathbb{E}[g_A(Q, Y_A Y_C)] + \mathbb{E}[g_C(Q, Y_A, Y_C)] = 0.$$

Using Lemmas 6.3 and 6.5, we get

$$\lambda(2\mathbb{E}[Q] + 2\mathbb{E}[\Delta_A(Y_A^{arrival})] + 1)$$

$$- 2\mu\mathbb{E}[Q] + \lambda + 2\mu\mathbb{E}[\Delta_C(Y_C^{comp})] + 2(\mu - \lambda)\mathbb{E}_U[\Delta_A(Y_A) - \Delta_C(Y_C^{comp})] = 0.$$

17

Rearranging the terms,

$$2(\mu - \lambda)\mathbb{E}[Q] = 2\lambda\mathbb{E}[\Delta_A(Y_A^{arrival})] + 2\mu\mathbb{E}[\Delta_C(Y_C^{comp})] + 2\lambda + 2(\mu - \lambda)\mathbb{E}_U[\Delta_A(Y_A) - \Delta_C(Y_C^{comp})]$$

$$\mathbb{E}[Q] = \frac{\lambda\mathbb{E}[\Delta_A(Y_A^{arrival})] + \mu\mathbb{E}[\Delta_C(Y_C^{comp})] + \lambda}{\mu - \lambda} + \mathbb{E}_U[\Delta_A(Y_A) - \Delta_C(Y_C^{comp})]$$

$$= \frac{\rho\Delta_A(Y_A^{arrival}) + \Delta_C(Y_C^{comp}) + \rho}{1 - \rho} + \mathbb{E}_U[\Delta_A(Y_A) - \Delta_C(Y_C^{comp})]. \qquad \square$$

7 Bounds on two-level system when $Q = 0$

We proved Theorem 5.1, a general result characterizing mean queue length in the MAMS system. As a corollary, in Corollary 5.1, we proved explicit bounds on mean queue length in the MAMS system. In the two-level arrival systems, stronger results can be obtained because we know more about the dynamics of the system. Specifically, recall that we proved Theorem 3.1, which characterized mean queue length $\mathbb{E}[Q]$, with the only remaining uncertainty being the probability $P(Y = H \mid Q = 0)$ that the system is in the H arrival state while the queue is empty.

In [32], it is shown that the exact value of this probability can be obtained by solving a cubic equation, which leads to the exact value of the mean queue length. In this section, we focus on proving explicit upper bounds on $P(Y = H \mid Q = 0)$, as stated in Theorem 7.1. These bounds complement the straightforward lower bound $P(Y = H \mid Q = 0) \geq 0$.

Theorem 7.1. *In the two-level arrival system, $P(Y = H \mid Q = 0)$ is bounded as follows:*

$$P(Y = H \mid Q = 0) \leq P(Y = H) = \frac{\alpha_L}{\alpha_L + \alpha_H}. \tag{35}$$

Additionally, if the system is intermittently overloaded ($\lambda_H > \mu$), the following bound holds:

$$P(Y = H \mid Q = 0) \leq \frac{1}{1 - \rho} \frac{1}{\lambda_H - \mu} \frac{\alpha_H \alpha_L}{\alpha_H + \alpha_L}. \tag{36}$$

We will prove Theorem 7.1 in Section 7.1.

These bounds are tight in different asymptotic regimes: (35) is tight in the $\alpha_H, \alpha_L \to \infty$ limit, where the arrival state switches rapidly, while (36) is tight in the $\alpha_H, \alpha_L \to 0$ limit, as the arrival state switches slowly. These bounds complement our heavy-traffic result for the two-level system, Corollary 3.1.

Using these bounds alongside Theorem 3.1, we can derive tight bounds on mean queue length in the two-level arrival system:

Corollary 7.1. *In the two-level arrivals system with exponential service rate, the mean queue length is bounded as follows:*

$$\mathbb{E}[Q] \geq \frac{\rho}{1 - \rho}\left(\frac{(\lambda_H - \lambda_L)^2 \alpha_L \alpha_H}{\lambda(\alpha_H + \alpha_L)^3} + 1\right) - \frac{\lambda_H - \lambda_L}{\alpha_H + \alpha_L}\frac{\alpha_L}{\alpha_H + \alpha_L}, \tag{37}$$

$$\mathbb{E}[Q] \leq \frac{\rho}{1 - \rho}\left(\frac{(\lambda_H - \lambda_L)^2 \alpha_L \alpha_H}{\lambda(\alpha_H + \alpha_L)^3} + 1\right). \tag{38}$$

If the system is intermittently overloaded ($\lambda_H > \mu$), then additionally

$$\mathbb{E}[Q] \leq \frac{\rho}{1 - \rho}\left(\frac{(\lambda_H - \lambda_L)^2 \alpha_L \alpha_H}{\lambda(\alpha_H + \alpha_L)^3} + 1\right) - \frac{\lambda_H - \lambda_L}{\alpha_H + \alpha_L}\frac{\alpha_L}{\alpha_H + \alpha_L}\left(1 - \frac{\alpha_H}{(1 - \rho)(\lambda_H - \mu)}\right). \tag{39}$$

Proof. These bounds follow from combining Theorem 3.1 with bounds on $P(Y = H \mid Q = 0)$. (37) follows from the fact that $P(Y = H \mid Q = 0) \geq 0$. (38) follows from (35). (39) follows from (36). $\square$

In the non-intermittently-overloaded case ($\lambda_H < \mu$), prior work has proven basic bounds on $\mathbb{E}[Q]$. Gupta et al. [13] proved that $\mathbb{E}[Q]$ is bounded below by the mean queue length of a M/M/1 with arrival rate λ and completion rate μ, and bounded above by a weighted mixture of two M/M/1s: one with arrival rate λ_H and one with arrival rate λ_L. While these bounds are not tight, it is challenging to use our techniques to prove tighter results in this setting, which we leave to future work.

7.1 Proof of Theorem 7.1

Proof. We start by considering the general two-level arrival system, with no restriction on λ_H and μ. We will prove (35):

$$P(Y = H \mid Q = 0) \le P(Y = H) = \frac{\alpha_L}{\alpha_L + \alpha_H}.$$

To prove this result, we make use of prior work by Gupta et al. [13] to bound the probability $P(Q = 0 \mid Y = L)$, the probability that the queue is empty given that the arrival rate is low.

Let π_0^L be the probability that the queue is empty at the moment when the system switches from L to H. By [13, Theorem 3], $P(Q = 0 \mid Y = L) = \pi_0^L$. During the proof of [13, Theorems 6 and 7], it is proven that there exists some $\chi \in (0, 1)$ such that $\pi_0^L / \pi_0^H = (\mu - \chi\lambda_L)/(\mu - \chi\lambda_H)$.

Because $\lambda_L \le \lambda_H$, we know that $\pi_0^L \ge \pi_0^H$, and hence that $P(Q = 0 \mid Y = L) \ge P(Q = 0 \mid Y = H)$.

To quantify $P(Q = 0)$, note that in all systems, the fraction of potential service events which go unused is $1 - \rho$. In the two-level arrival system, because the completion process is exponential, it is therefore the case that $P(Q = 0) = 1 - \rho$.

Thus, we know that $P(Q = 0 \mid Y = L) \ge 1 - \rho$, and that $P(Q = 0 \mid Y = H) \le 1 - \rho$. We can therefore bound $P(Y = H \mid Q = 0)$:

$$P(Y = L \mid Q = 0) \ge P(Y = L) = \frac{\alpha_H}{\alpha_H + \alpha_L}, \quad P(Y = H \mid Q = 0) \le P(Y = H) = \frac{\alpha_L}{\alpha_H + \alpha_L}.$$

Now, consider the intermittently overloaded two-level arrival system, where $\lambda_H > \mu$. We will prove (36):

$$P(Y = H \mid Q = 0) \le \frac{1}{1 - \rho} \frac{1}{\lambda_H - \mu} \frac{\alpha_H \alpha_L}{\alpha_H + \alpha_L}.$$

The intuition behind this bound is that if $\lambda_H > \mu$, the queue length grows throughout each $Y = H$ interval, and hence the queue is rarely empty (Q is rarely 0) during such an interval.

To make this idea rigorous, consider an overloaded M/M/1 with arrival rate λ_H greater than the service rate μ, in which the queue is empty at time 0 ($Q(0) = 0$). The expected total amount of time for which $Q(t) = 0$ in the overloaded M/M/1 system is finite: by a standard calculation, one can show that this expected duration is $1/(\lambda_H - \mu)$. That amount of time is an upper bound on the expected amount of time for which $Q = 0$ during each $Y = H$ interval in the two-level system.

Thus, the expected time that two-level system spends in the state $y = H, q = 0$ during each $y = H$ interval is at most $1/(\lambda_H - \mu)$, regardless of the initial queue length at the beginning of the interval. A larger initial queue length and a finite interval length can only decrease the expected time for which the system is empty.

To bound $P(Y = H \wedge Q = 0)$, we use a renewal-reward argument, with two kinds of cycles: A *switching cycle*, consisting of a $y = H$ interval followed by a $y = L$ interval, and a *zero-switching cycle*, which is a renewal cycle with a renewal point whenever the system changes from $y = L$ to $y = H$ at a moment when $Q = 0$. Let Z be a random variable denoting the number of switching cycles within each zero-switching cycle. Note that Z has a finite mean because the two-level system is positive recurrent with $\lambda < \mu$.

We now apply the renewal-reward theorem [17], with the renewal period being the zero-switching cycles, and the reward being the time spent in the state $Y = H, Q = 0$:

$$P(Y = H \wedge Q = 0) = \frac{\mathbb{E}[Y = H \wedge Q = 0 \text{ duration during a zero-switching cycle}]}{\mathbb{E}[\text{Length of a zero-switching cycle}]}.$$
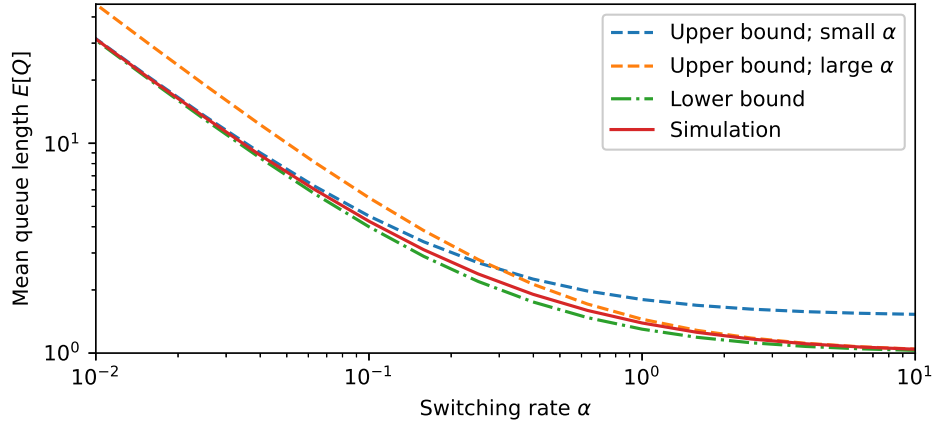
19

Figure 4: Simulation and bounds for mean queue length $\mathbb{E}[Q]$ in two-level system with intermittent overload, under varying switching rates α_H, α_L. Setting: $\lambda_H = 2, \lambda_L = 0.2, \alpha_H = 5\alpha_L, \mu = 1, \rho = 0.5$. Overall switching rate $\alpha := (\alpha_H + \alpha_L)/2$. Bounds given in Corollary 7.1: The upper bound which is tight for small α is (39), for large α is (38), and the lower bound is (37). Simulated 10^8 arrivals.

The lengths of the switching cycles within the zero-switching cycles are i.i.d. with expected length $\frac{1}{\alpha_H} + \frac{1}{\alpha_L}$, and the length of zero-switching cycle is a stopping time with mean $E[Z](\frac{1}{\alpha_H} + \frac{1}{\alpha_L})$. The amount of time for which $Y = H \wedge Q = 0$ during the switching cycles are correlated, but only via the length of the queue at the start of the cycle. Regardless of that queue length, the expected time for which $Y = H \wedge Q = 0$ per switching cycle is at most $1/(\lambda_H - \mu)$. As a result,

$$P(Y = H \wedge Q = 0) \leq \frac{\mathbb{E}[Z]/(\lambda_H - \mu)}{\mathbb{E}[Z](1/\alpha_H + 1/\alpha_L)} = \frac{1/(\lambda_H - \mu)}{1/\alpha_H + 1/\alpha_L}.$$

Recall that $P(Q = 0) = 1 - \rho$, because the system has exponential service. As a result,

$$P(Y = H \mid Q = 0) \leq \frac{1}{1 - \rho} \frac{1/(\lambda_H - \mu)}{1/\alpha_H + 1/\alpha_L} = \frac{1}{1 - \rho} \frac{1}{\lambda_H - \mu} \frac{\alpha_H \alpha_L}{\alpha_H + \alpha_L}. \square$$

Note that (36) scales linearly with α_H, α_L, so it converges to 0 in the $\alpha_H, \alpha_L \to 0$ limit.

8   Simulation

We have characterized the mean queue length for the MAMS system, with bounds given in Corollary 5.1. We have proven even tighter bounds for the two-level arrival system in Corollary 7.1.

In this section, we simulate mean queue length for the two-level arrival system and a more general MAMS system, and compare our simulations to our bounds, to validate our bounds and demonstrate their tightness.

In Fig. 4, we simulate an intermittently overloaded two-level arrivals system under a variety of switching rates α_H, α_L. In Fig. 4, we have moderate overall load, because $\rho = 0.5$, but significant periods of intermittent overload, because $\lambda_H = 2 > \mu = 1$.

Fig. 4 shows that our bounds in Corollary 7.1 are tight in both the fast-switching limit (large α, short periods of time in each arrival state) and the slow-switching limit (small α, long periods of time in each arrival state), with our bounds tightly and accurately constraining the mean queue length in both limits. Specifically, (37) and (38) are tight in the fast-switching limit, while (37) and (39) are tight in the slow-switching limit. Moreover, our bounds are reasonably tight at all switching rates α, even
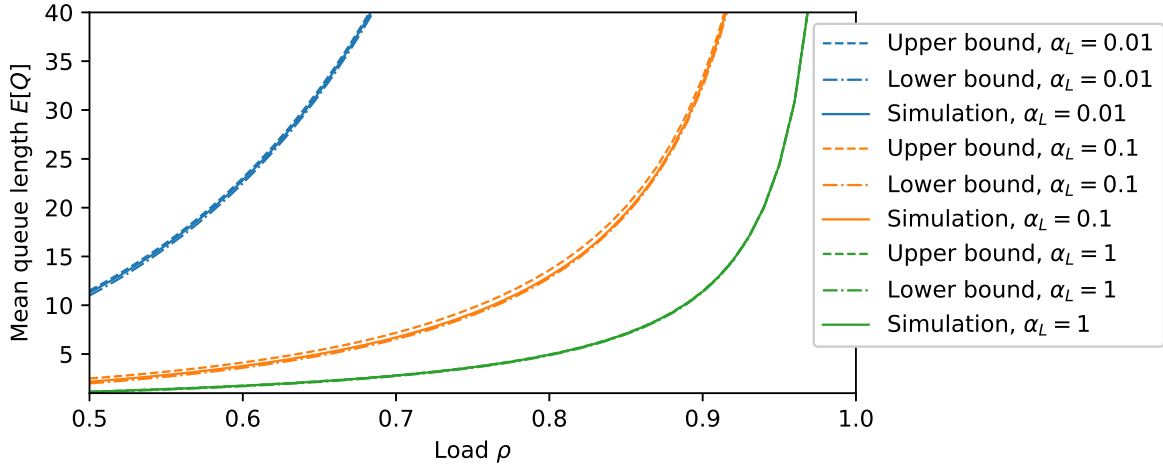
Figure 5: Simulation and bounds for mean queue length $\mathbb{E}[Q]$ in two-level system with intermittent overload, under varying load ρ, for three different values of α_L. Setting: $\lambda_H = 4\rho, \lambda_L = 0.4\rho, \alpha_H = 5\alpha_L, \mu = 1$. Bounds given in Corollary 7.1: Upper bound is minimum of (38) and (39). Simulated 10^8 arrivals.
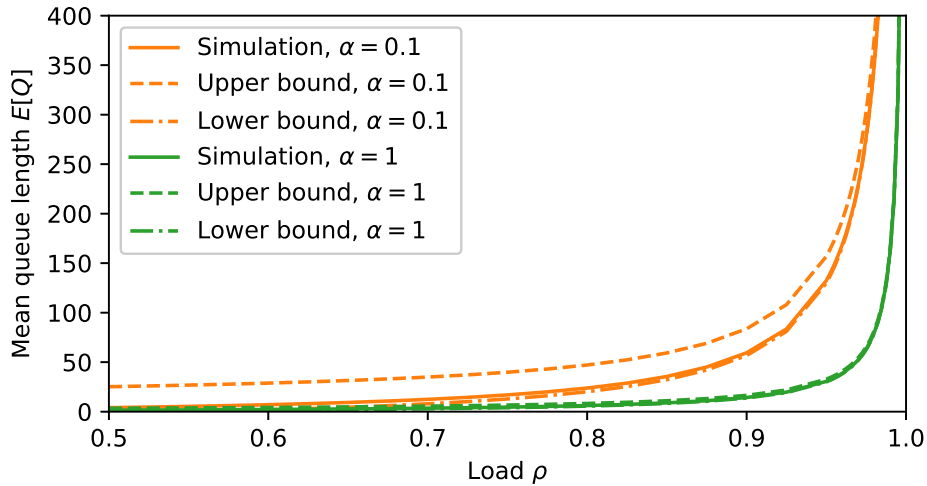

Figure 6: Setting: MAMS queue with three arrivals levels: $[0.3\rho, 2\rho, 2.2\rho]$, and three completions levels: $[0.5, 1.0, 3.0]$. The system remains in each arrival state and each service state for time $Exp(\alpha)$, then moves cyclically to the next rate in the list, wrapping around. Bounds given in Corollary 5.1. Simulated 10^9 arrivals.

outside the asymptotic regime, with a maximum absolute error of 0.5 jobs, and a maximum relative error of 23%.

In Fig. 5, we simulate a two-level system under a variety of loads ρ, and under three different pairs of switching rates α_H, α_L. We compare the simulated mean response time against our bounds from Corollary 7.1. In every configuration of load ρ and switching rates α_H, α_L simulated in Fig. 5, the absolute gap between our upper and lower bounds is at most 1 job. The relative error between our bounds drops towards zero as the system moves towards heavy traffic ($\rho \to 1$), in each setting.

In Fig. 6, we simulate a MAMS system with three levels of arrival rates and three levels of completion rates, under varying loads ρ. The system cycles between three possible arrival rates and three possible completion rates. Recall that as with all MAMS systems, the arrival chain and completions chain evolve independently. The switching rate from each arrival and completion state to the next is a fixed rate α. In Fig. 6, we show two switching rates α. We compare the simulated mean queue lengths $\mathbb{E}[Q]$ against our bounds from Corollary 5.1. Fig. 6 illustrates that our bounds are tight in heavy traffic, even for the more complex MAMS system, confirming the heavy-traffic limit result in Corollary 5.2. Unlike the two-level arrival setting illustrated in Figs. 4 and 5, our bounds are not as tight in the general MAMS setting in light traffic (i.e., for small ρ), especially for slower switching settings (low α). Proving more precise bounds on $\mathbb{E}[Q]$ for general MAMS systems in light traffic is left for future work.

9   Conclusion

We analyze the Markovian arrivals Markovian service (MAMS) system, using a somewhat novel framework based on the drift method and the concepts of relative arrivals and relative completions. While the concepts of relative arrivals and relative completions are a rediscovery of existing techniques [8], their application via the drift method [7] is novel to this work. Using this framework, we capture the impact of the correlated arrivals and completions on mean queue length. We derive the explicit bounds on mean queue length in the MAMS system, with bounds that are tight in heavy traffic, matching prior results.

Moreover, in the important special case of the two-level arrivals system under intermittent overload, we prove significantly stronger bounds which are tight both in the fast-switching and slow-switching limits, again matching prior results.

An important direction for future work is to analyze MAMS systems with infinite arrival and/or completions chains. Such systems do not appear to have been studied in prior work. Our proof of the main theorem, Theorem 5.1, holds in such systems, assuming appropriate conditions on the arrivals and completions chains, but it does not immediately imply a tight bound on mean queue length.

10   Acknowledgements

References

[1] Akramul Azim, Shreyas Sundaram, and Sebastian Fischmeister. An efficient periodic resource supply model for workloads with transient overloads. In *2013 25th Euromicro Conference on Real-Time Systems*, pages 249–258, 2013. doi: 10.1109/ECRTS.2013.34.

[2] David Y. Burman and Donald R. Smith. An asymptotic analysis of a queueing system with markov-modulated arrivals. *Operations Research*, 34(1):105–119, 1986. doi: 10.1287/opre.34.1.105.

[3] Carri W. Chan, Jing Dong, and Linda V. Green. Queues with time-varying arrivals and inspections with applications to hospital discharge policies. *Operations Research*, 65(2):469–495, 2017. doi: 10.1287/opre.2016.1536.

[4] Inho Cho. *Mitigating Compute Congestion for Low Latency Datacenter RPCs*. PhD thesis, Massachusetts Institute of Technology, 2023.

[5] Florin Ciucu and Felix Poloczek. Two extensions of kingman's gi/g/1 bound. *Proc. ACM Meas. Anal. Comput. Syst.*, 2(3), dec 2018. doi: 10.1145/3287322. URL https://doi.org/10.1145/3287322.

[6] Mitko Dimitrov. Single-server queueing system with markov-modulated arrivals and service times. *Pliska Stud. Math. Bulg*, 20:53–62, 2011.

[7] Atilla Eryilmaz and Rayadurgam Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems*, 72:311–359, 2012.

[8] Gennadi Falin and Anatoli Falin. Heavy traffic analysis of m/g/1 type queueing systems with markov-modulated arrivals. *Top*, 7(2):279–291, 1999.

[9] Paulo R. Farah and Cristina D. Murta. A transient overload generator for web servers. In *2008 IEEE International Performance, Computing and Communications Conference*, pages 119–126, 2008. doi: 10.1109/PCCC.2008.4745140.

[10] Anshul Gandhi, Mor Harchol-Balter, Ram Raghunathan, and Michael A. Kozuch. Autoscale: Dynamic, robust capacity management for multi-tier data centers. *ACM Trans. Comput. Syst.*, 30 (4), nov 2012. ISSN 0734-2071. doi: 10.1145/2382553.2382556. URL https://doi.org/10.1145/2382553.2382556.

[11] Peter W Glynn, Assaf Zeevi, et al. Bounding stationary expectations of Markov processes. *Markov processes and related topics: a Festschrift for Thomas G. Kurtz*, 4:195–214, 2008.

[12] Isaac Grosof, Yige Hong, Mor Harchol-Balter, and Alan Scheller-Wolf. The RESET and MARC techniques, with application to multiserver-job analysis. *Performance Evaluation*, 162:102378, 2023. ISSN 0166-5316. doi: https://doi.org/10.1016/j.peva.2023.102378. URL https://www.sciencedirect.com/science/article/pii/S0166531623000482.

[13] Varun Gupta, Mor Harchol-Balter, Alan Scheller Wolf, and Uri Yechiali. Fundamental characteristics of queues with fluctuating load. In *Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '06/Performance '06, page 203–215, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933190. doi: 10.1145/1140277.1140301. URL https://doi.org/10.1145/1140277.1140301.

[14] Madhav Erraguntla Guy L. Curry, Hiram Moya and Amarnath Banerjee. Transient queueing analysis for emergency hospital management. *IISE Transactions on Healthcare Systems Engineering*, 12(1):36–51, 2022. doi: 10.1080/24725579.2021.1933655.

[15] Bruce Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied Probability*, 14(3):502–525, 1982. doi: 10.2307/1426671.

[16] Jorgen Hansson and Sang H Son. Overload management in RTDBS. In *Real-Time Database Systems: Architecture and Techniques*, pages 125–139. Springer, 2001.

[17] Mor Harchol-Balter. *Performance modeling and design of computer systems: Queueing theory in action*. Cambridge University Press, 2013.

[18] Anushree Kaushik, Gulista Khan, and Priyank Singhal. Cloud energy-efficient load balancing: A green cloud survey. In *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)*, pages 581–585, 2022. doi: 10.1109/SMART55829.2022.10046686.

[19] Dave Kearns. CPU utilization: When to start getting worried. https://www.networkworld.com/article/2341220/cpu-utilization–when-to-start-getting-worried.html, 2003. Accessed: 2023-10-16.

[20] Guy Latouche and Vaidyanathan Ramaswami. *Introduction to matrix analytic methods in stochastic modeling*. Society for Industrial and Applied Mathematics, 1999.

[21] Shiyu Liu, Ahmad Ghalayini, Mohammad Alizadeh, Balaji Prabhakar, Mendel Rosenblum, and Anirudh Sivaraman. Breaking the Transience-Equilibrium nexus: A new approach to datacenter packet transport. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*, pages 47–63, 2021.

[22] Shancong Mou and Siva Theja Maguluri. Heavy traffic queue length behaviour in a switch under Markovian arrivals, 2020.

[23] Marcel F Neuts. The M/M/1 queue with randomly varying arrival and service rates. *Management Science*, 15(4):139–157, 1978.

[24] MF Neuts. Matrix-geometric solutions in stochastic models: An algorithmic approach. *Johns Hopkins University Press*, 1981.

[25] Martin L Puterman. Markov decision processes. *Wiley Series in Probability and Statistics*, 1994.

[26] Chenhao Qu, Rodrigo N. Calheiros, and Rajkumar Buyya. Auto-scaling web applications in clouds: A taxonomy and survey. 51(4), jul 2018. ISSN 0360-0300. doi: 10.1145/3148149. URL https://doi.org/10.1145/3148149.

[27] V. Ramaswami. The N/G/1 queue and its detailed analysis. *Advances in Applied Probability*, 12 (1):222–261, 1980. doi: 10.2307/1426503.

[28] G. J. K. Regterschot and J. H. A. de Smit. The queue m—g—1 with markov modulated arrivals and services. *Mathematics of Operations Research*, 11(3):465–483, 1986. doi: 10.1287/moor.11.3.465.

[29] Bianca Schroeder and Mor Harchol-Balter. Web servers under overload: How scheduling can help. *ACM Trans. Internet Technol.*, 6(1):20–52, feb 2006. ISSN 1533-5399. doi: 10.1145/1125274. 1125276. URL https://doi.org/10.1145/1125274.1125276.

[30] Rein Vesilo, Mor Harchol-Balter, and Alan Scheller-Wolf. Scaling properties of queues with time-varying load processes: Extensions and applications. *Probability in the Engineering and Informational Sciences*, 36(3):690–731, 2022. doi: 10.1017/S0269964821000048.

[31] Ronald W. Wolff. Poisson arrivals see time averages. *Operations Research*, 30(2):223–231, 1982. doi: 10.1287/opre.30.2.223.

[32] U. Yechiali and P. Naor. Queuing problems with heterogeneous arrivals and service. *Operations Research*, 19(3):722–734, 1971. doi: 10.1287/opre.19.3.722.

A Expectation over unused service

We now formally define the expectation $\mathbb{E}_U[\cdot]$, the *expectation at moments of unused service.* Unused service occurs at rate μ_{i_C} whenever $q = 0$.

For any test function $f(q, i_A, i_C)$, we define its expectation at moments of unused service $\mathbb{E}_U[\cdot]$ as

$$\mathbb{E}_U[f(Q, Y_A, Y_C)] := \frac{\mathbb{E}[\mu_{Y_C} f(Q, Y_A, Y_C) \mathbb{1}\{Q = 0\}]}{\mathbb{E}[\mu_{Y_C} \mathbb{1}\{Q = 0\}]} = \frac{\mathbb{E}[\mu_{Y_C} f(Q, Y_A, Y_C) \mathbb{1}\{Q = 0\}]}{\mu - \lambda},$$

where the second equality holds because $\mathbb{E}[\mu_{Y_C} \mathbb{1}\{Q = 0\}] = \mu - \lambda$ by Lemma A.1.

In the special case of exponential service times, $\mu_{i_C} = \mu$ for any $i_C \in \mathbb{C}$, so

$$\mathbb{E}_U[f(Q, Y_A, Y_C)] = \mathbb{E}[f(Q, Y_A, Y_C) \mid Q = 0] = \frac{\mathbb{E}[f(Q, Y_A, Y_C) \mathbb{1}\{Q = 0\}]}{1 - \rho}.$$

We will also write $\mathbb{E}_U[f(Y_C^{comp})]$, which is defined as the expectation of $f(i_C)$ with respect to the distribution of Y_C right after unused service, i.e., the completions that happen during $Q = 0$. Formally,

$$\mathbb{E}_U[f(Y_C^{comp})] := \mathbb{E}_U\left[\sum_{j_C \in \mathbb{C}} \frac{s_{Y_C, j_C, 1}}{\mu_{Y_C}} f(j_C)\right] = \frac{\mathbb{E}\left[\sum_{j_C \in \mathbb{C}} s_{Y_C, j_C, 1} f(j_C) \mathbb{1}\{Q = 0\}\right]}{\mathbb{E}\left[\mu_{Y_C} \mathbb{1}\{Q = 0\}\right]}.$$

Note the similarity with the definition of Y_C^{comp} in Section 4.2.

We now characterize the rate of *unused service* in the system, namely completion transitions that occur when the queue is empty:

Lemma A.1. *The rate of unused service is $\mu - \lambda$: $\mathbb{E}_{i \sim Y_C}[\mu_i \mathbb{1}\{Q = 0\}] = \mu - \lambda$.*

Proof. The long-run arrival rate is λ. The long-run rate of completion transitions is μ. Nonetheless, the long-run arrival and completion rate must be equal, because the system is stable. Thus the long-run completion rate is λ. The only way that a completion transition can not cause a completion is if the queue is empty, causing unused service. As a result, the rate of unused service must be $\mu - \lambda$. □

Note that by Lemma A.1, in systems with exponential service such as the two-level arrivals system, $P(Q > 0) = \lambda/\mu = \rho$.

B Relative arrivals and completions

First, we verify that Δ_A and Δ_C, the relative arrivals and relative completions functions, are well-defined and finite. Note that Δ_A is the relative value function of a average-reward Markov Reward process with Markov chain matching the arrival process and reward equal to the arrival rate λ_i. By Puterman [25, Section 8.2.1], this relative value is well-defined and finite for any finite-state Markov chain. Δ_C is equivalent.

Next, we establish two systems of equations, equivalent to the Poisson equation for MDPs, which can be used to calculate $\Delta_A(i)$ in each state $i \in \mathbb{A}$ (or $\Delta_C(i)$ in each state $i \in \mathbb{C}$), up to an additive constant. Incorporating Lemma 4.3, which states that $\mathbb{E}[\Delta_A(Y_A)] = 0$ (or $\mathbb{E}[\Delta_C(Y_A)] = 0$), uniquely characterizes $\Delta_A(i)$ (or $\Delta_C(i)$).

Lemma B.1. *For any arrivals chain, the relative arrivals function Δ_A and relative completions function Δ_C satisfy the following systems of equations:*

$$\forall i \in \mathbb{A}, \quad \Delta_A(i) = \frac{\lambda_i - \lambda}{r_{i, \cdot, \cdot}} + \sum_{j \in \mathbb{A}} \frac{r_{i, j, \cdot}}{r_{i, \cdot, \cdot}} \Delta_A(j), \quad \forall i \in \mathbb{C}, \quad \Delta_C(i) = \frac{\mu_i - \mu}{s_{i, \cdot, \cdot}} + \sum_{j \in \mathbb{C}} \frac{s_{i, j, \cdot}}{s_{i, \cdot, \cdot}} \Delta_C(j). \quad (40)$$

Proof. Let us start with Definition 4.1: $\Delta_A(i) := \lim_{t\to\infty} \mathbb{E}[A_i(t)] - \lambda t$, where $A_i(t)$ denotes the number of arrivals by time t, given that the arrivals chain starts from state i.

Let us split up $\mathbb{E}[A_i(t)] - \lambda t$, the expression inside the limit, into two time periods: Until the arrival chain next undergoes a transition, and after that time. The transition occurs after $Exp(r_{i,\cdot,\cdot})$ time, or $\frac{1}{r_{i,\cdot,\cdot}}$ expected time. Prior to the transition, $\mathbb{E}[A_i(t)] - \lambda t$ accrues at rate $\lambda_i - \lambda$. We use Y_A^{next} to denote the state after the transition. Then we have the expression:

$$\Delta_A(i) = \frac{\lambda_i - \lambda}{r_{i,\cdot,\cdot}} + \lim_{t\to\infty} \mathbb{E}[A_{Y_A^{next}}(t)] - \lambda t = \frac{\lambda_i - \lambda}{r_{i,\cdot,\cdot}} + \mathbb{E}[\Delta_A(Y_A^{next})].$$

Finally, note that $P(Y_A^{next} = j) = \frac{r_{i,j,\cdot}}{r_{i,\cdot,\cdot}}$, allowing us to derive the equations for $\Delta_A(i)$ in (40). The equations for $\Delta_C(i)$ can be proved similarly. $\qquad\square$

As a corollary of Lemma B.1, we characterize the change in the value of $\Delta_A(i)$ when a transition occurs in the arrival chain, and the change in the value of $\Delta_C(i)$ when a transition occurs in the completion chain.

Corollary B.1. *In the MAMS system, the effect of arrival (or completion) on relative arrival (or relative completion) satisfies:*

$$\forall i \in \mathbb{A}, \qquad \sum_{j \in \mathbb{A}, a \in \{0,1\}} r_{i,j,a}(\Delta_A(j) - \Delta_A(i) + a) = \lambda, \tag{41}$$

$$\forall i \in \mathbb{C}, \qquad \sum_{j \in \mathbb{C}, c \in \{0,1\}} s_{i,j,c}(\Delta_C(j) - \Delta_C(i) + c) = \mu. \tag{42}$$

Equivalently, for any $i \in \mathbb{A}$, $G \circ \Delta_A(i) = \lambda - \lambda_i$; for any $i \in \mathbb{C}$, $G \circ \Delta_C(i) = \mu - \mu_i$.

Proof. By (40) in Lemma B.1, for any $i \in \mathbb{A}$, $\Delta_A(i) = \frac{\lambda_i - \lambda}{r_{i,\cdot,\cdot}} + \sum_{j \in \mathbb{A}} \frac{r_{i,j,\cdot}}{r_{i,\cdot,\cdot}} \Delta_A(j)$. Multiplying both sides by $r_{i,\cdot,\cdot}$, rearranging the terms, and expanding $r_{i,\cdot,\cdot}, r_{i,j,\cdot}$, and λ_i, we get (41). (42) follows similarly.

It can be convenient to express (41) and (42) in terms of $G \circ \Delta_A$ and $G \circ \Delta_C$. Note that by Lemma 4.2, $G \circ \Delta_A(i) = \sum_{j \in \mathbb{A}} r_{i,j,\cdot}(\Delta_A(j) - \Delta_A(i))$. Thus, $G \circ \Delta_A(i) = \lambda - \lambda_i$ and $G \circ \Delta_C(i) = \mu - \mu_i$. $\qquad\square$

C Generator and drift method

Lemma 4.1. *Let f be a test function for which $\mathbb{E}[f(Q, Y_A, Y_C)] < \infty$. Then*

$$\mathbb{E}[G \circ f(Q, Y_A, Y_C)] = 0.$$

Proof. We apply Proposition 3 of [11]. To prove Lemma 4.1 using that proposition, we only need to verify that the total transition rate of the system is uniformly bounded.

For each state (q, i_A, i_C), the total transition rate of the MAMS system is the total transition rate of the arrival process and the completion process, i.e., $\sum_{j_A \in \mathbb{A}, a} r_{i_A, j_A, a} + \sum_{j_C \in \mathbb{C}, c} s_{i_C, j_C, c}$, which is finite for any $i_A \in \mathbb{A}$ and $i_C \in \mathbb{C}$. Because $\mathbb{A}$ and $\mathbb{C}$ are both finite sets, the total transition rates of the MAMS system are uniformly bounded. $\qquad\square$

Next, we will give sufficient condition for proving that $\mathbb{E}[f(Q, Y_A, Y_C)] < \infty$, allowing us to apply Lemma 4.1 to the test functions we use in this paper. To prove our sufficient condition, we need [15, Theorem 2.3], restated below using the state representation of the MAMS system.

Lemma C.1. *Consider a Markov chain with uniformly bounded total transition rates, and a Lyapunov function V that satisfies the conditions below for any state (q, i_A, i_C): $V(q, i_A, i_C) \geq 0$; there exists $b, \gamma > 0$ such that whenever $V(q, i_A, i_C) \geq b$,*

$$G \circ V(q, i_A, i_C) \leq -\gamma; \tag{43}$$

there exists $d > 0$ such that for any initial state, (q, i_A, i_C), and any state after one step of transition, (q', j_A, j_C),

$$V(q', j_A, j_C) - V(q, i_A, i_C) \leq d. \tag{44}$$

Then there exists $\theta > 0$ such that $\mathbb{E}[e^{\theta V(Q, Y_A, Y_C)}] < \infty$.

We also need a characterization of the drift of a generic test function in the MAMS system, Lemma 4.2.

Lemma 4.2. *For any real-valued function f of the state of the MAMS system,*

$$G \circ f(q, i_A, i_C) = \sum_{j_A \in \mathbb{A}, a \in [0,1]} r_{i_A, j_A, a} \left(f(q + a, j_A, i_C) - f(q, i_A, i_C) \right)$$

$$+ \sum_{j_C \in \mathbb{C}, c \in [0,1]} s_{i_C, j_C, c} \left(f(q - c + u, i_A, j_C) - f(q, i_A, i_C) \right),$$

where $u = (c - q)^+$ denotes the unused service.

Proof. Because the generator is the stochastic equivalent of the derivative operator, to get an expression for $G \circ f(q, i_A, i_C)$, we examine all types of transitions in the MAMS system.

For each $j_A \in \mathbb{A}$ and $a \in \{0, 1\}$, with rate $r_{i_A, j_A, a}$, the state of the arrival process changes to j_A accompanied by a arrivals. This type of transition changes the state to $(q + a, j_A, i_C)$, so it contributes $r_{i_A, j_A, a} \left(f(q + a, j_A, i_C) - f(q, i_A, i_C) \right)$ to $G \circ f(q, i_A, i_C)$.

Similarly, for each $i_C \in \mathbb{C}$ and $c \in \{0, 1\}$, with rate $s_{i_C, j_C, c}$, the state of the completion process changes to j_C accompanied by c completions. This transition changes the state to $((q - c)^+, i_A, j_C)$, so it contributes $s_{i_C, j_C, c} \left(f((q - a)^+, i_A, j_C) - f(q, i_A, i_C) \right)$ to $G \circ f(q, i_A, i_C)$. Let $u = (c - q)^+$, then $(q - a)^+ = q - a + u$.

We sum up all of the above terms to obtain the formula for $G \circ f(q, i_A, i_C)$. $\qquad\square$

We are now ready to state and prove our sufficient condition for $\mathbb{E}[f(Q, Y_A, Y_C)] < \infty$ using Lemma C.1 and Lemma 4.2. Lemma C.2 is analogous to Lemma A.2 of [12].

Lemma C.2. *Let f be a real-valued function of the state of the MAMS system, (q, i_A, i_C). Suppose $f(q, i_A, i_C)$ grows at a polynomial rate in q. Suppose that $\lambda < \mu$ in the MAMS system. Then f has finite expectation in steady state: $\mathbb{E}[f(Q, Y_A, Y_C)] < \infty$.*

Proof. We will verify conditions of Lemma C.1. Define $V(q, i_A, i_C) = (q + \Delta_A(i_A) - \Delta_C(i_C))^+$. Let $\Delta_{\max} = \max\{\max_{i_A \in \mathbb{A}} |\Delta_A(i_A)|, \max_{i_C \in \mathbb{C}} |\Delta_C(i_C)|\}$. Letting $b = 1 + 4\Delta_{\max}$ and $\gamma = \mu - \lambda$. Consider any state (q, i_A, i_C) such that $V(q, i_A, i_C) \geq b$.

- By the definition of $V(q, i_A, i_C)$, we have $q \geq 1 + 2\Delta_{\max}$, so $V(q, i_A, i_C) = q + \Delta_A(i_A) - \Delta_C(i_C)$.

- Moreover, for any state (q', j_A, j_C) reachable after one step of transition from (q, i_A, i_C), we must have $V(q', j_A, j_C) \geq q - 1 + \Delta_A(j_A) - \Delta_C(j_C) \geq 0$, which implies that $V(q', j_A, j_C) = q' + \Delta_A(j_A) - \Delta_C(j_C)$.

By Lemma 4.2,

$$G \circ V(q, i_A, i_C) = \sum_{j_A \in \mathbb{A}, a} r_{i_A, j_A, a}(a + \Delta_A(j_A) - \Delta_A(i_A)) + \sum_{j_C \in \mathbb{C}, c} s_{i_C, j_C, c}(-c + u - \Delta_C(j_C) + \Delta_C(i_C)).$$

Because $q \geq 1 + 2\Delta_{\max}$ and $c \leq 1$, we have $u = (c - q)^+ = 0$. Therefore, Corollary B.1 implies that $G \circ V(q, i_A, i_C) = \lambda - \mu < 0$. This proves the condition (43). It is not hard to see that the other condition, (44), holds with $d = 1 + 4\Delta_{\max}$. Then Theorem 2.3 of [15] implies the existence of $\theta > 0$ such that $\mathbb{E}\left[e^{\theta V(q, i_A, i_C)}\right] < \infty$. Observe that $e^{\theta V(q, i_A, i_C)}$ grows exponentially fast in q. Therefore, for any f that grows at a polynomial rate in q,

$$f(q, i_A, i_C) = O(e^{\theta V(q, i_A, i_C)}), \quad \mathbb{E}[f(Q, Y_A, Y_C)] < \infty. \qquad\square$$