Setup Times In Multiserver Systems

Jalani K. Williams

May 2024

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Prof. Weina Wang (CMU), Chair
Prof. Mor Harchol-Balter (CMU)
Prof. Alan Scheller-Wolf, (CMU)
Prof. Jamol Pender, (Cornell)
Prof. Bill Massey, (Princeton)

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.*

Keywords: queueing, multiserver systems, setup times, exceptional first service

To my friends and family.

Abstract

In many systems, servers do not turn on instantly; instead, a *setup time* must pass before a server can begin work. These "setup times" can wreak havoc on a system's queueing; this is especially true in modern systems, where servers are regularly turned on and off as a way to reduce operating costs (energy, labor, CO_2, etc.). To design modern systems which are both efficient *and* performant, we need to understand how setup times affect queues.

Unfortunately, despite successes in understanding setup in the single server setting, setup in the multiserver setting remains poorly understood. To circumvent the main difficulty in analyzing multiserver setup, all existing results assume that setup times are memoryless, i.e. distributed Exponentially. However, in most practical settings, setup times are close to Deterministic, and the widely used Exponential-setup assumption leads to unrealistic model behavior and a dramatic underestimation of the true harm caused by setup times.

This thesis represents a comprehensive characterization of the average waiting time in a multiserver system with *Deterministic* setup times, the M/M/k/Setup-Deterministic. In particular, we derive multiplicatively-tight lower and upper bounds on the average waiting time, demonstrating that **setup times, along with their distributions, can not be ignored; setup times can cause profound increases in waiting time, *especially* when the distribution of setup time has low variability.** Our bounds are the first closed-form bounds on waiting time in *any* many-server system with setup times, including the extensively-studied Exponential setup system. Furthermore, we use our bounds to derive a highly-accurate approximation, which we evaluate in a variety of settings. These results are made possible via our new method for bounding the expectation of a random time integral, called the Method of Intervening Stopping Times or MIST.

Acknowledgments

Special thanks to my friends and family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Setting

1.1.1 What are setup times?

In many systems, servers do not turn on instantly; instead, a *setup time* must pass before a server can begin work. For example, for applications hosted in the cloud, application replicas must take time to boot up before they can begin fulfilling requests; for overwhelmed hospitals, traveling nurses must wait to have their credentials confirmed before they can begin helping patients; for many businesses, workers must go through a long process of recruitment and onboarding before they can begin serving customers. By thinking about this "initial delay before service" as an abstract *setup time*, we can learn how setup time affects all of these systems simultaneously.

1.1.2 Why do setup times matter?

Setup times can have a significant impact on a system's queueing behavior, especially in modern systems. For systems which keep their servers *on* all the time, clearly setup times do not affect their performance. However, in many modern systems, servers are regularly turned on and off. Because servers don't turn on instantly, jobs in a system with setup times end up delayed compared to their no-setup counterparts. If setup times are long enough, this additional delay can be significant.

Nevertheless, many systems still regularly turn their servers off and on. Why? Because by doing so, one can save a considerable amount on operating costs, e.g. energy, money, CO_2, etc. That said, this cost-saving measure is only a viable option if the additional delay caused by setup times is not too large. Therefore, if we want to design systems which are simultaneously efficient *and* performant, we need a good understanding of how setup times affect queueing performance.

1.1.3 What makes the setup effect difficult to understand?

Setup times in the M/M/1. Note, though, that the additional delay caused by setup does not always manifest in an obvious way. For example, consider a simple single server queue with setup times, the M/M/1/Setup. The job which is the first to arrive to an empty system triggers the *off* server's setup, and must wait a full setup time before it is served. Likewise, every job that arrives afterward must wait in line behind the setup-triggering job, and so also partially observes the server setting up. However, even after it turns on, the fact that the server was off for a while has a lasting impression on the length of the queue; the queue is longer than it would otherwise be. Thus, setup can even affect the delay of jobs that never actually observe a server in setup.

The effect of setup on queueing behavior is made even more complex when the setup time is allowed to be a random variable, sampled independently every time setup is initiated, and when the length of each job follows an arbitrary distribution. This more complex model is called the M/G/1/Setup. Further, despite its apparent complexity, the full waiting time distribution of the M/G/1/Setup was completely characterized in 1964 by [30].

1.1.4 Why setup times are *even harder* in the multiserver setting?

Unfortunately, the effect of setup on delay is even harder to understand when multiple servers can set up at the same time. Recall that, in the single server setting, the server's setup process always completes once initiated, and there is always at least one job to work on once the server turns on. This implies that, although setup has a complex effect on job delay, the server's behavior itself is quite simple: it first initiates setup; then, once setup completes, the server begins working; then, after finishing all the work in the system, the server turns off. On the other hand, when multiple servers can *simultaneously* be *in setup*, their server states begin to interact.

In particular, via the speed of their processing, the *busy* servers indirectly control the setup behavior of the *not-busy* servers. For example, if server A is *on* while server B is setting up, then server A might finish all the work in the queue before server B even has a chance to turn *on*. As such, in the multiserver setting, it can sometimes make sense to *cancel* a server's setup process; a situation which would *never* occur in the single server setting. Of course, the opposite can also happen: if the busy servers are working much more slowly than expected, then the queue might grow large enough that we begin setting up a server that would otherwise be left *off*. This interaction between departure behavior and setup behavior is exactly what makes the setup effect so much harder to understand in the multiserver setting.

1.2 Our problem: Understanding the M/M/k/Setup

In this work, we study the effect of setup times on the average waiting time in the M/M/k/Setup, a simple variation on the classic M/M/k queue which accurately captures the complexity which arises from simultaneous setup.

1.2.1 Brief Model Description

Job Dynamics. Outside of its setup dynamics, the M/M/k/Setup behaves essentially identically to the usual M/M/k. Jobs arrive in a Poisson process to a central queue, where they wait in First-Come-First-Served order until they are served by one of k servers. The job at the head of the queue enters service whenever either 1) a server finishes setting up and turns on or 2) a server

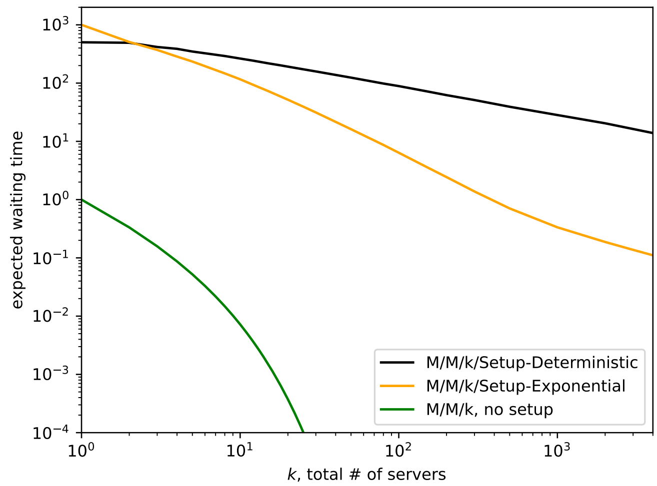

Figure 1.1: Simulation results for the M/M/k/Setup-Deterministic, M/M/k/Setup-Exponential, M/M/k (no setup), varying the number of servers k and keeping fixed the service rate $\mu = 1$, the setup time $\beta = 1000$, and the load $\rho = 0.5$. Note the high separation between all three models.

finishes its current job. Once in service, the job stays in service for an i.i.d. Exponential amount of time, after which the job departs.

Setup Dynamics. To complete our description of the M/M/k/Setup, it suffices to describe how the system controls the setup process. Servers can be in one of three states: *off*, *on*, or *in setup*. Servers turn *off* whenever they finish their current job and there are no jobs waiting in the queue. Servers turn *on* when they have remained *in setup* for a full *setup time*; in general, this setup time is some i.i.d. *random variable* which is sampled at the moment that setup is first initiated. Servers initiate and cancel *setup* based on job arrivals and departures, respectively. In particular, if a job arrives to the system and sees *off* servers, it initiates setup at one of these *off* servers (we assume every server is identical). Likewise, if the number of *in setup* servers ever exceeds the number of jobs waiting in the queue, then the system turns off the servers which have been *in setup* for the least amount of time. Using the M/M/k/Setup, we can now study the complex interactions which arise from simultaneous setup.

1.2.2 Understanding the M/M/k/Setup: State of the Art

Everyone uses the Exponential model. Despite continued academic interest, our understanding of the M/M/k/Setup is still extremely limited. Perhaps the most significant limitation is that all state-of-the-art research [15, 27, 28] assumes that setup times are Exponentially distributed. This limitation has major consequences for the utility of their work.

The Exponential model is unrealistic. We give two reasons why this limitation is so significant. First, the "Exponential setup" assumption leads to extremely unrealistic behavior in some situations. To illustrate where the breakdown in realism happens, consider a scenario where only a single server is setting up and compare it to a scenario where 100 servers begin setup at the same time. In the Exponential setup model, the 100-server system receives its first server on average 100x faster than the single-server system receives its first (and only) server. This is not a quirk of our specific example: in the Exponential model, the *longer* the system's queue is, the *more rapidly* the system's servers turn on to help *drain* that queue. In a sense, the Exponential system can rapidly "react" to increases in queue length.

The Exponential model underestimates waiting. This unrealistic "reactivity" phenomenon causes a further, more concerning, problem. We observe in Figure 1.1 that, in systems where setup times are actually closer to *Deterministic*, modeling setup times as Exponential can lead to a dramatic under-estimation of the harm caused by setup times, sometimes by many orders of magnitude. This dramatic under-estimation is the reason why, in many practical studies of the setup effect [20, 22], setup times are assumed to be Deterministic, e.g. servers take a fixed time of 2 minutes to set up.

Challenges of the Deterministic model. However, although modeling setup times as Deterministic might be more realistic, it also comes with a set of unique theoretical challenges. In the Deterministic case, one must, even in simulation, track the individual remaining setup time of *every* server that is currently setting up. In contrast, because the Exponential distribution is *memoryless*, in the Exponential case it suffices to track only the *total number* of servers setting up instead, greatly simplifying the system state. Moreover, the Exponential setup model's simplified state forms a Continuous-Time Markov Chain, a well-studied class of stochastic processes for which a number of techniques have been developed. For the Deterministic setup model, no such techniques exist.

1.3 Contributions

In this thesis, I develop the first results on the average waiting time in the M/M/k/Setup-Deterministic, demonstrating that **setup times, along with their distributions, can not be ignored; setup times can cause profound increases in waiting time, *especially* when the distribution of setup time has low variability.**

The contributions of this thesis are both theoretical and practical. On the theoretical side, in Chapters 5and 6, respectively, I derive the first lower and upper bounds on the average waiting time in the M/M/k/Setup-Deterministic. Notably, these results are the first closed-form bounds on the average waiting time in *any* M/M/k/Setup system, including the extensively-studied Exponential setup system. We obtain these bounds via a new technique for bounding random time integrals called MIST, described in Chapter 4. On the practical side, in Chapter 7, I then show how to take the components of our upper and lower bounds, and combine them to make a highly accurate approximation; an example of the approximation alongside the bounds is shown in Figure 8.3. Finally, in Chapter 8, I explore the practical implications of our findings on provisioning for modern "dynamically-scaled" multiserver systems.

1.4 Outline

Chapter 2: Prior Work. In Chapter 2, we begin our study of setup by reviewing some related work. We start by discussing the single server setting, then we move through the history of the study of setup times up to the state of the art. For each result we review, we compare and contrast their work with the main results developed in this thesis.

Chapter 3: Model. In Chapter 3, we give a more detailed description of our model. Besides reviewing the brief description we gave in this chapter, we also describe our notation and give a construction of our processes of interest using Poisson processes.

Chapter 4 : Key Ideas and Techniques. Next, in Chapter 4, we describe the key ideas and techniques of this thesis. In particular, we introduce the Method of Intervening Stopping Times; the *MIST* method. We describe the MIST method by first describing its general function, then stating its associated formal definition, then proving a key lemma which allows it to be generally applied.

Chapter 5: The Lower Bounds. In Chapter 5, we describe our first two main results (Theorems 5.1 and 5.2), both lower bounds on the average queue length in the M/M/k/Setup-Deterministic. We begin by describing in greater detail why a lower bound is needed, then proceed by stating both bounds and proving the stronger one.

Chapter 6: The Upper Bound. In Chapter 6, we describe our final main result (Theorem 6.1), an upper bound on the average queue length in the M/M/k/Setup-Deterministic. As we did in Chapter 5 with the lower bounds, we first describe why we need this upper bound. Afterwards, we give its proof.

Chapter 7: The Approximation. After proving these results, in Chapter 7, we develop an approximation to the average waiting time in the M/M/k/Setup-Deterministic. As before, we first describe why such an approximation is needed. Afterwards, we explicitly state the approximation formula and describe how to derive the approximation from the bounds in Chapters 5 and 6.

Chapter 8: Evaluation. In Chapter 8, we evaluate the direct practical implications of our work. In particular, we ask and answer three questions concerning the M/M/k/Setup:

1. How much does setup distribution matter?
2. How does our approximation's accuracy change as we vary the system parameters?
3. What impact do our results have on the practice of provisioning?

Chapter 9: Conclusion. Finally, in Chapter 9, we summarize the main results of this thesis, discuss some possible applications, and describe a few open problems.

Chapter 2

Prior Work

In Chapter 2, we discuss the body of literature which analyzes systems with setup times.

2.1 Systems without Simultaneous Setup

2.1.1 The M/G/1/Setup

The best-understood case is the single-server case. The foremost result on this model is the result of [30]; the author considers a generalization of the M/G/1 queue where, if a customer arrives while the server is idle, then they have a different service distribution than if they arrive while the server is busy. By observing that the system state at customer departure times forms a discrete Markov chain, then analyzing that embedded chain, Welch characterizes the steady-state and transient distributions of the queue length; via distributional Little's Law, this gives the same result for delay and response time. This important result has been extended in a variety of different directions, by adjusting the service discipline or arrival process[2, 3, 18]..

2.1.2 M/M/k and M/G/k with staggered setup

The easiest case of multiserver systems with setup times involves the *staggered setup* model, where at most one server can be in setup at a time, greatly simplifying the analysis. In [1], the authors obtain an expression for the steady-state distribution of queue length for the system when setup times are Exponential, using the method of difference equations. In [11] the authors simplify the solution of the M/M/k with exponential setup times considerably, an prove a decomposition result for mean delay. In [9], the decomposition result is generalized to a hyperexponential job size distribution, and shown to hold approximately for a general job size distribution.

2.2 The M/M/k/Setup-Exponential

2.2.1 M/M/k/Setup-Exponential, Approximations

All previous theoretical results that investigate an M/M/k/Setup system assume Exponential setup times. We first highlight the state-of-the-art papers concerning approximating the M/M/k/Setup-Exponential. In particular, we highlight the work in [27] and [11]. Gandhi et al. [11] seek useful intuitive approximations to the M/M/k/Setup-Exponential system. Their approximations stem from an exact analysis of the M/M/∞/Setup-Exponential system, which they then modify in various ways to capture the finite server case. The approximations in [11] work well, except when both load and setup times are moderately high ($\rho > 0.5$ and $\frac{\mu}{\alpha} > 10$).

Pender and Phung-Duc [27] consider a generalization of the M/M/k/Setup-Exponential model which includes non-stationary arrival rate and customer abandonment. Within this model, they derive a mean field approximation for the system dynamics, which they prove converges as the number of servers, k, approaches infinity.

Unlike our work, neither Pender and Phung-Duc [27] nor Gandhi et al. [11] provide explicit bounds on the delay. The approximations themselves are also not stated as an explicit function of the system parameters. Finally, neither considers Deterministic setup times.

2.2.2 M/M/k/Setup-Exponential, Exact Analysis

There are only a few results that deal with the exact analysis of the M/M/k with Exponential setup times. The most well-known are [15] and [28].

In a followup to the approximation work done in [11], in [15] the authors give the first exact analysis of the M/M/k/Setup-Exponential system. To do this, they develop the *Recursive Renewal Reward (RRR)* technique, which allows them to analyze 2-dimensional Markov chains of a certain structure. They apply this technique to the M/M/k/Setup-Exponential system, and thus provide a method for computing the time-average value of *any* function of the system state; applying this method to the correct function gives the mean and Laplace transform for the number of jobs in queue.

In [28], the author rederives the exact solutions for the queue length obtained in [15] using two different methods: an analysis using generating functions, and an analysis applying the matrix analytic method after casting the system as a quasi-birth-death process. Although these techniques appear different, the author highlights some core correspondences between them, and also between these methods and the RRR technique of [15].

Despite the fact that [15] and [28] represented the first breakthrough in our understanding of the setup effect in 50 years, their results are limited in two significant ways. First, instead of a closed-form formula for the average waiting time, the authors only derived an algorithm for computing the average waiting time. This algorithm is useful in the sense that it bypasses the need to simulate the system, but unfortunately fails to give intuition about wait times scale with system parameters. Moreover, like all of the works mentioned in this section, their work assumes that setup times are distributed Exponentially, which turns out to severely limit the utility of their results; see Section 8.1 for a detailed discussion.

2.2.3 Distributed Setting

There has also been some work on servers with setup times outside of the centralized queue setting. In [25], the authors consider a queueing system which functions much like the M/M/k/Setup-Exponential, except, instead of a central queue, each server has its own queue, and there is a central dispatcher which routes arriving jobs to one of these queues. In this model, they describe a token-based load balancing and scaling scheme called TABS, and prove that its performance (as $k \to \infty$) is asymptotically optimal. In particular, they show that the relative energy wastage and the mean delay both go to 0 under their scheme, by analyzing an appropriate fluid limit. In a followup paper, [24], the authors consider the performance of TABS in the infinite-buffer case. They give two results. First, they show that, somewhat counterintuitively, there exist parameter settings under which the TABS scheme is unstable. Second, they show that, in spite of this finite instability, for sufficiently large k, the system under TABS is stable. Moreover, its performance continues to be asymptotically optimal. In our opinion, it is best to think of these results as complementary to the body of work on the M/M/k/Setup, as one typically does when comparing distributed queueing work to centralized queueing work. Although all papers discussed deal with setup times in some capacity, the nature of the questions being asked and answered in [25] and [24] are very different from the central-queue-oriented work we discuss.

2.3 Scheduling with Setup

Gittins in the G/G/k/Setup. In [19], the authors consider a very general model queueing model, the G/G/k/Setup, and show that the scheduling performance of the Gittins policy is near-optimal in this setting. In particular, they explicitly bound the deviation from optimality of the average waiting time under the Gittins policy, showing that this "suboptimality loss" is uniformly bounded at all loads. They thus conclude that the Gittins policy is heavy-traffic optimal.

Their work differs from ours in two important ways. First, they investigate a model of setup where the setup process is never cancelled. While this may be accurate in certain situations, a reasonable amount of complexity in our problem stems from the fact that the setup process can be cancelled. Second, they are mainly concerned with bounding the performance of a scheduling policy as compared to the optimal scheduling policy. - By contrast, our principle results are concerned with directly characterizing the average waiting time. In that sense, [19] serves as an interesting study whose results are somewhat orthogonal to ours.

2.4 Prior Work on Deterministic Setup Times

M/G/2/Setup-Deterministic, with dispatching In the control literature, deterministic setup times have been incorporated into models in order to enhance realism. Hyytiä et al. [20] consider a dispatching version of the M/G/2/Setup-Deterministic model, and attempt to build near-optimal policies for the joint control of setup initiation and the dispatching of jobs. We hope that our analysis here could open the door to more fine-grained stochastic analysis of such control policies.

M/M/k/Setup-Deterministic, simulation only The only work we have found which discusses the M/M/k/Setup-Deterministic model explicitly is a simulation-based thesis by Kara [22]. They observe that the mean delay in the M/M/k/Setup-Deterministic is consistently larger than that of the M/M/k/Setup-Exponential, and, as the mean setup time $\frac{1}{\alpha}$ increases, the relative increase in mean delay between the M/M/k/Setup-Deterministic and the M/M/k/Setup-Exponential also increases. We corroborate and expand on their results in Chapter 8.

Algorithms for reducing the effect of setup times on delay and energy usage Setup times are both a problem from a delay perspective and also from an energy perspective (servers utilize peak power while in setup [14]). One can of course avoid setup times altogether by always leaving servers on, but this results in wasted energy as well, since a server which is on, but idle, utilizes 60-70% of peak energy [14]. To manage power efficiently, several algorithms have been developed to reduce the costly effects of setup times. One idea is *DelayedOff*, whereby a one waits some time before turning off a server, so as to avoid a future setup time [10, 11, 14, 27]. Another idea is routing jobs to the *Most Recently Busy server (MRB)*, so as to minimize the size of the pool of servers that are turning on and off [10]. Similar to MRB is the idea of creating a rank ordering of all servers and always sending each job to the *lowest-numbered server in the rank* [14]. The goal of all such algorithms is to minimize the Energy-Response-time-Product (ERP) [10], maximize the Normalized-Performance-Per-Watt (NPPW) [8], or minimize energy given a fixed tail cutoff for response time [14]. Other ideas for minimizing delay and energy involve utilizing sleep states in servers, which require more power than being off, but have a lower setup time [12, 13].

Chapter 3

Model

In Chapter 3, we discuss our model of interest, the M/M/k/Setup-Deterministic. We begin the chapter by going through a detailed model description, then discuss how to construct the relevant stochastic processes via Poisson processes.

3.1 Detailed Model Description

The system behavior, excluding setup. As in the typical M/M/k queue, jobs arrive in a Poisson process of rate $k\lambda$ into a FCFS queue where jobs wait to be served at one of k servers. The job at the head of queue enters service whenever a server frees up, either from a job completing service or from a server finishing set up. Once a job enters service, it remains in service for Exp(μ) time before departing. We assume all the servers have identical service and setup properties. As such, we can assign each server an index from 1 to k, and without loss of generality assume that departures always occur at the busy server with the highest index; i.e., we re-index the servers when a job departs so the server with the newly departed job has the highest index among the busy servers. From here, we define the quantity $Z(t)$ to be the number of busy servers (or jobs in service) at time t, the quantity $Q(t)$ to be the number of jobs waiting in the queue at time t, and the quantity $N(t) = Q(t) + Z(t)$ to be the total number of jobs in our system. Excluding the setup dynamics, one sees that, as promised, our model behaves identically to the M/M/k queue.

The setup dynamics. From here, it suffices to describe precisely how servers will be turned *on* and *off*. We assume that each server is always in one of three states: *on*, *off*, or *in setup*. A given server remains *on* only as long as that server remains busy. In other words, a server turns *off* when it finishes its current job and the queue is empty. On the other hand, server i begins *setup* when a job arrives to the system and there are only $i - 1$ jobs in the system. Server i remains in setup until one of two events occurs: either 1) some fixed quantity β time has passed, or 2) there are fewer than i jobs in the system; accordingly, we refer to β as the *setup time* of a server. In the first case, if β time has passed without $N(t)$ dipping below i, then server i has completed its setup and begins working on the job at the head of the queue. In the second case, if the number of jobs $N(t)$ dips below i before server i completes setup, then the setup is canceled and server i turns *off*. We use $Y_i(t)$ to denote the detailed state of server i at time t. If server i is *off*, we set

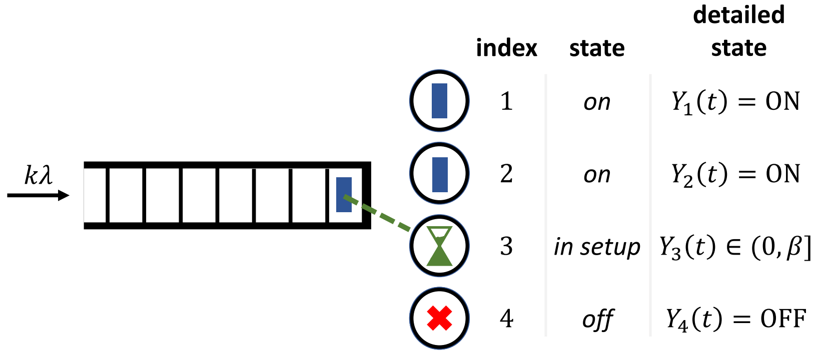

Figure 3.1: An example of M/M/k/Setup-Deterministic with $k = 4$. The state pictured has $Z(t) = 2$ busy servers, which means there are 2 jobs in service. There is $Q(t) = 1$ job in queue, and thus $N(t) = Z(t) + Q(t) = 4$ jobs in system.

$Y_i(t) = \mathrm{OFF}$; if server i is *on*, we set $Y_i(t) = \mathrm{ON}$; if server i is *in setup*, we let $Y_i(t)$ denote the remaining amount of time until server i would finish setup, if left uninterrupted. To be precise, $Y_i(t)$ is set to β when server i first initiates *setup*, and this value decreases at rate 1 until either setup completes or setup is canceled. For convenience, we assume, without loss of generality, that $\mathrm{ON} < s < \mathrm{OFF}$ for every possible remaining setup time $s \in (0, \beta]$; this ensures that the detailed state $Y_i(t)$ is non-decreasing in i. As a shorthand, we use $\boldsymbol{Y} = (Y_1(t), Y_2(t), \ldots, Y_k(t))$ to denote the vector of detailed server states.

A state descriptor. Accordingly, a Markovian state descriptor for our system at time t is $S(t) \triangleq (N(t), \boldsymbol{Y}(t))$. Note that, since one can recover the number of jobs in service $Z(t)$ from the detailed server states $\boldsymbol{Y}(t)$, one could also choose the state to be $(Q(t), \boldsymbol{Y}(t))$. Either suffices in providing a complete description of the forward dynamics of the system. Furthermore, when discussing the steady-state distribution of, say, the number of jobs $N(t)$, we use the notation $N(\infty)$.

Some important constants. We define some system parameters which are critical to system behavior. We use $\rho \triangleq \frac{\lambda}{\mu}$ to refer to the load of our system, i.e., the time-average utilization of an average server. We call the offered load $R \triangleq k\rho$; this is the time-average *number* of busy servers in our system. To enforce stability, we require that $\rho < 1$. As discussed previously, the symbol β refers to the fixed (Deterministic) setup time of a server.

Busy period notation. Our results can be stated more concisely with two quantities related to a busy period of an M/M/1 queue. We give the notation below. We use $T^{\text{busy}}(n, j)$ to denote the expectation of the random *length* of an $M/M/1$ busy period with arrival rate $k\lambda$, service rate $k\lambda + \mu j$, and which starts with n jobs in the system. Likewise, we use $I^{\text{busy}}(n, j)$ to denote expectation of the random *time integral of the number of jobs* within the $M/M/1$ over the same period. Explicitly, we have

$$T^{\text{busy}}(n, j) = \frac{n}{\mu j} \tag{3.1}$$

and

$$I^{\text{busy}}(n, j) = \frac{n}{\mu j}\left[\frac{n+1}{2} + \frac{R}{j} + 1\right]. \tag{3.2}$$

3.2 Construction

We now discuss how we formally construct this system using Poisson processes; being explicit here will prove useful when we make coupling arguments in the future.

The arrival and departure processes. We take the number of jobs that have arrived at time t to be $\Pi_A(t)$, where Π_A is a Poisson process of rate $k\lambda$. In a slight abuse of notation, we let $\Pi_A([a, b])$ denote the number of arrivals that occur in the interval $[a, b]$; we apply the same extension to all other counting processes mentioned here. We set the potential departure process of, say, server i to be $\Pi_i(t)$, where Π_i is a Poisson process of rate μ. A potential departure from server i only "counts" if server i is busy when that potential departure occurs, i.e., if the number of busy servers $Z(t) \geq i$ at the time. Thus, the total number of departures from our system by time t is

$$D(t) \triangleq \sum_{i=1}^{k} \int_0^t \mathbf{1}\{Z(s) \geq i\}\, \mathrm{d}\Pi_i(s),$$

where these integrals are with respect to the Π_i's as counting processes.

The number of busy servers $Z(t)$. To find the number of busy servers $Z(t)$, one could count the number of setup completion events that have occurred so far and the number of server shutoffs that have occurred so far; this description is a bit difficult to work with. Alternatively, one can see from the initial description of setup dynamics that server i is *on* at time t if and only if the total number of jobs $N(s) \geq i$ for all $s \in [t - \beta, t]$, where one should recall that β is the setup time. An easier description of $Z(t)$ follows:

$$Z(t) = \min\left(k, \min_{s \in [t-\beta, t]} N(s)\right).$$

A departure operator. We can extend our departure process $D(t)$ to a departure operator $\mathcal{D}[f(s)](\mathcal{I})$ which takes a function $f(s) \in \{0, 1, \ldots, k\}$ defined on some interval $\mathcal{I}$ and computes the number of departures that would occur in that interval provided that the number of busy servers $Z(s) = f(s)$, i.e.

$$\mathcal{D}[f(s)]((a, b]) \triangleq \sum_{i=1}^{k} \int_a^b \mathbf{1}\{f(s) \geq i\}\, \mathrm{d}\Pi_i(s).$$

Note that we can write the total number of departures using our newly-defined operator as $D(t) = \mathcal{D}[Z(s)]([0, t])$.

Chapter 4

Key Ideas and Techniques

In this chapter, we describe the key techniques underlying the major results of this thesis. We introduce these techniques separately; they will then be made applied directly throughout the thesis. We describe each technique by first describing its general function, then stating its associated formal definition, then proving the lemma which makes its application general.

4.1 The Method of Intervening Stopping Times (MIST)

4.1.1 Why we need it

The basic function of this lemma is to bound the expected time integral between two random events in some Markov system, an initial event and a final event. This type of problem arises often in the study of stochastic systems. In particular, within this thesis, we are mainly concerned with bounding the average waiting time $\mathbb{E}\left[T_Q\right]$ in the M/M/k/Setup-Deterministic. By applying Little's Law [17] and the Renewal Reward theorem [17], we can translate the problem of bounding the average waiting time into the problem of bounding the time integral of the queue length $Q(t)$ between a specially-defined time 0 state and another stopping time X, i.e.

$$\mathbb{E}\left[T_Q\right] = \frac{1}{k\lambda}\mathbb{E}\left[Q(\infty)\right] = \frac{1}{k\lambda}\frac{\mathbb{E}\left[\int_0^X Q(t)\mathrm{d}t\right]}{\mathbb{E}\left[X\right]} = \frac{1}{k\lambda}\frac{\mathbb{E}\left[\int_0^X Q(t)\mathrm{d}t\right]}{\mathbb{E}\left[\int_0^X 1\mathrm{d}t\right]}. \tag{4.1}$$

From here, it suffices to bound the expectation of these time integrals; this is what the Intervening Stopping Time Lemma, Lemma 4.1, does.

4.1.2 What it does

The basic idea of this lemma is to break up our random time interval of interest into a *random number* of smaller, more manageable pieces. We do this by defining *intervening events*, moments where something special happens to the system state that gives us an opportunity to characterize the system's behavior. From there, we can define a "small piece" of time as the time in between these intervening events. For example, in this work, it can often be useful to analyze the system around time points where the number of jobs $N(t)$ gets large. Because we work in a system with setup times, if we have a lot of jobs for a long enough period of time, then, by the end of that long period of time, we can guarantee that a lot of servers are turned on.

By performing this decomposition of the integral into smaller pieces, we reduce our initial bounding problem to showing two facts:

- First, we must show that the time integral of these smaller pieces is not too big; in this thesis, we typically use martingale arguments combined with worst-case coupling arguments to prove this fact.
- Second, we must show that not too many of these these smaller pieces actually occur. For this "not too many" condition, it's particularly helpful if we can show that, if the i-th intervening event has occurred, then the $(i+1)$-th event has at most a constant probability of occurring.

By formalizing our notion of events using stopping times and applying some ideas from Wald's equation, we obtain the Intervening Stopping Time Lemma, Lemma 4.1, which we now state and prove.

4.1.3 IST Lemma: Statement and Proof

Lemma 4.1 (Intervening Stopping Time Lemma). *Given a starting stopping time T_0, an ending stopping time P, and a collection of intervening stopping times $(T_i : i \in \mathbb{Z}^+)$, define the random variable F to be such that $T_F \leq P < T_{F+1}$. Now, given some time-varying random variable $Y_t \geq 0$ which is a function of the underlying Markov state of the system $\mathcal{S}(t)$, suppose that:*

1. $\mathbb{E}\left[\int_{T_0}^{\min(T_1,P)} Y_t dt \middle| \mathcal{F}_{T_0}\right] \leq G_0(\mathcal{S}(T_0))$,
2. $\mathbb{E}\left[\int_{T_i}^{\min(T_{i+1},P)} Y_t dt \middle| \mathcal{F}_{T_i}, F \geq i\right] \leq G_i + B \cdot \mathbb{E}\left[\min(T_{i+1}, P) - T_i | \mathcal{F}_{T_i}, F \geq i\right]$,
3. *and* $\Pr\left(F \geq i | \mathcal{F}_{T_i}, F \geq i-1\right) \leq 1 - p_i$,

where G_0 is also some function of the system state, and the G_i's, the p_i's, and B are all constants (possibly depending on system parameters).

Then,

$$\mathbb{E}\left[\int_{T_0}^{P} Y_t dt\right] \leq \mathbb{E}\left[G_0\left(\mathcal{S}(T_0)\right)\right] + \Pr\left(F > 0\right) \sum_{j=1}^{\infty} G_j \prod_{i=2}^{j} (1 - p_i) + B \cdot \mathbb{E}\left[P - T_0\right].$$

Proof.

We begin with a manipulation of the integral, finding

$$\begin{aligned}\int_{T_0}^{P} Y_t \mathrm{d}t &= \int_{T_0}^{\min(T_1,P)} Y_t \mathrm{d}t + \sum_{i=1}^{\infty} \int_{\min(T_i,P)}^{\min(T_{i+1},P)} Y_t \mathrm{d}t \\ &= \int_{T_0}^{\min(T_1,P)} Y_t \mathrm{d}t + \sum_{i=1}^{\infty} \mathbf{1}_{T_i<P} \int_{T_i}^{\min(T_{i+1},P)} Y_t \mathrm{d}t.\end{aligned}$$

Applying linearity of expectation and the tower property, we find that

$$\begin{aligned}
&\mathbb{E}\left[\int_{T_0}^{P} Y_t \mathrm{d}t\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\int_{T_0}^{\min(T_1,P)} Y_t \mathrm{d}t \middle| \mathcal{F}_{T_0}\right]\right] + \sum_{i=1}^{\infty} \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{T_i<P}\int_{T_i}^{\min(T_{i+1},P)} Y_t \mathrm{d}t \middle| \mathcal{F}_{T_i}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\int_{T_0}^{\min(T_1,P)} Y_t \mathrm{d}t \middle| \mathcal{F}_{T_0}\right]\right] + \sum_{i=1}^{\infty} \mathbb{E}\left[\mathbf{1}_{T_i<P}\mathbb{E}\left[\int_{T_i}^{\min(T_{i+1},P)} Y_t \mathrm{d}t \middle| \mathcal{F}_{T_i}\right]\right].
\end{aligned}$$

Noting that the event $\{T_i < P\} = \{F \geq i\}$, we have

$$\begin{aligned}
&= \mathbb{E}\left[\mathbb{E}\left[\int_{T_0}^{\min(T_1,P)} Y_t \mathrm{d}t \middle| \mathcal{F}_{T_0}\right]\right] + \sum_{i=1}^{\infty} \mathbb{E}\left[\mathbf{1}_{F\geq i}\mathbb{E}\left[\int_{T_i}^{\min(T_{i+1},P)} Y_t \mathrm{d}t \middle| \mathcal{F}_{T_i}\right]\right] \\
&\leq \mathbb{E}\left[G_0\left(S(T_0)\right)\right] + \sum_{i=1}^{\infty} \mathbb{E}\left[\mathbf{1}_{F\geq i}\left(G_i + B\cdot\mathbb{E}\left[\min(T_{i+1},P) - T_i | \mathcal{S}(T_i), F\geq i\right]\right)\right] \\
&= \mathbb{E}\left[G_0\left(S(T_0)\right)\right] + B\cdot\mathbb{E}\left[P - T_0\right] + \sum_{i=1}^{\infty} G_i \Pr\left(F\geq i\right).
\end{aligned}$$

Applying our final assumption to bound $\Pr\left(F \geq i\right)$,

$$\begin{aligned}
\Pr\left(F\geq i\right) &= \Pr\left(F>0\right)\prod_{j=2}^{i}\Pr\left(F\geq j | F\geq j-1\right) \\
&= \Pr\left(F>0\right)\prod_{j=2}^{i}\mathbb{E}\left[\Pr\left(F\geq j \middle| F\geq j-1, \mathcal{F}_{T_{j-1}}\right)\right] \\
&\leq \Pr\left(F>0\right)\prod_{j=2}^{i}\mathbb{E}\left[1-p_j\right] \\
&= \Pr\left(F>0\right)\prod_{j=2}^{i}(1-p_j).
\end{aligned}$$

Applying this final result, we find

$$\mathbb{E}\left[\int_{T_0}^{P} Y_t \mathrm{d}t\right] \leq \mathbb{E}\left[G_0\left(\mathcal{S}(T_0)\right)\right] + \Pr\left(F>0\right)\sum_{j=1}^{\infty} G_j \prod_{i=2}^{j}(1-p_i) + B\cdot\mathbb{E}\left[P-T_0\right],$$

as desired. □

With Lemma 4.1 proven, we are ready to apply the MIST method to obtain our main results.

Chapter 5

The Lower Bounds

In this chapter, we discuss two of our results, our lower bounds on the average waiting time in the M/M/k/Setup-Deterministic. First, we discuss why these lower bounds are needed, then state both bounds, then prove the stronger and more recent bound.

5.1 Why we need a lower bound

From a provisioning standpoint, a lower bound tells us what system parameters **necessary** to achieve a certain average waiting time. Accordingly, we now discuss our two lower bounds. The first lower bound that we present, Theorem 5.1 (the main result of [32]), is the first-ever result bounding the average waiting time in the M/M/k/Setup-Deterministic. Notably, it is also the first closed-form result bounding the average waiting time in any M/M/k/Setup system. The second lower bound that we present, Theorem 5.2 (one of the two main results in [31]), is an improvement of Theorem 5.1. The improved lower bound now applies to systems with an arbitrarily large numbers of servers k, removes an unnecessary and restrictive condition on the system parameters, and also has a far simpler proof. We state both theorems, but only prove the improved theorem.

5.2 The First Lower Bound

We now state the first lower bound for the average queue length in the M/M/k/Setup-Deterministic, from [32].

Theorem 5.1 (First Lower Bound On Average Queue Length)**.** *In an M/M/k/Setup-Deterministic system with load ρ, setup time $\beta \geq 1000\frac{1}{\mu}$, and offered load $R \triangleq k\rho \geq 128$, if the setup time $\beta \geq \frac{1}{\mu}\log^2(k\rho)$, then the average queue length $\mathbb{E}\left[Q(\infty)\right]$ is lower bounded by*

$$\mathbb{E}\left[Q(\infty)\right] \geq \frac{\frac{1}{2}\beta^2\frac{\mu\sqrt{R}}{2} + I^{\text{busy}}\left(\left[(\mu\beta-1)\,\frac{\sqrt{R}}{2} - k(1-\rho)\right]^+, k-R\right)}{C_1^{(\text{old})}\left(3\beta+\frac{1}{\mu}\right) + \beta + C_2^{(\text{old})}\frac{\mu\beta\sqrt{R}}{\mu k(1-\rho)} + C_3^{(\text{old})}\frac{1}{\mu}\log\left(C_2^{(\text{old})}\mu\beta\sqrt{R}\right)},$$

where $C_1^{(\text{old})}$, $C_2^{(\text{old})}$, and $C_3^{(\text{old})}$ are absolute constants.

5.3 The New Lower Bound

5.3.1 The New Lower Bound: Theorem Statement

After tightening and clarifying our techniques into the MIST method of Chapter 4, we obtained the following lower bound; its proof follows.

Theorem 5.2 (Improved Lower Bound on Average Queue Length)**.** *In an M/M/k/Setup-Deterministic, let the stabilizing number $R \triangleq k\rho \triangleq\geq 100$ and $\beta \geq 100\frac{1}{\mu}$. Then,*

$$\mathbb{E}\left[Q(\infty)\right] \geq \frac{L_1\beta^2\sqrt{R} + I^{\text{busy}}\left(\left[L_1\beta\sqrt{R} - (k-R)\right]^+, k-R\right)}{2.08\beta + \frac{1}{\mu}\frac{F_1\beta\sqrt{R}}{k-R} + \frac{1}{\mu}\frac{3}{2}\ln(\beta) + \frac{1}{\mu}\ln(F_1D_1) + \frac{2}{\mu} + \left[D_2 + \frac{D_3}{\sqrt{R}}\right]\max\left(\frac{1}{D_1\sqrt{\mu\beta}}, \frac{1}{\sqrt{R}}\right)},$$

where L_1, F_1, D_1, D_2, and D_3 are absolute constants.

5.3.2 The New Lower Bound: Proof Outline.

Basic Structure. We prove Theorem 5.2 via the MIST method. As noted in Chapter 4, we begin by applying the Renewal-Reward theorem to the queue length $Q(t)$, defining our renewal points as those points in time where the $(R+1)$-th server turns off. Defining time 0 to be one of these points, and defining the cycle time $X \triangleq \min\left\{t > 0 : Z(t^-) = R+1, Z(t) = R\right\}$ as the next point, this gives

$$\mathbb{E}\left[Q(\infty)\right] = \frac{\mathbb{E}\left[\int_0^X Q(t)\mathrm{d}t\right]}{\mathbb{E}\left[X\right]}.$$

To obtain our lower bound, it suffices to lower bound the numerator and upper bound the denominator of this fraction, i.e. lower bound $\mathbb{E}\left[\int_0^X Q(t)\mathrm{d}t\right]$ and upper bound $\mathbb{E}\left[X\right]$. The time integral lower bound is handled by Lemma 5.1, which we state at the end of this section. The cycle length upper bound is split into two separate lemmas: Lemma 5.2 upper bounds the length of the cycle's "first part" and Lemma 5.3 bounds the length of its "second part".

Decomposition into phases. However, before we state or prove these lemmas, we first discuss the decomposition of the renewal cycle $[0, X)$ into two parts; one might think of this as a "miniature" application of the MIST method. We begin by noting that the end of the renewal cycle is moment when the $(R+1)$-th server turns off. Since the $(R+1)$-th server is off at the start of a renewal period, we can break the renewal cycle into two phases based on whether the $(R+1)$-th server has turned on yet. Formally, we define the accumulation time $T_A \triangleq \min\left\{t > 0 : Z(t) = R+1\right\}$ as the first moment that the $(R+1)$-th server turns on. From here, we can focus separately on the accumulation phase, from time 0 to time T_A, and the draining phase, from time T_A to time X.

With this decomposition, we can now state our main lemmas. Their proofs follow in sequence afterwards.

Lemma 5.1 (Lower bound on Cycle Integral)**.** *Define busy period integral $I^{\text{busy}}(x,z)$ as*

$$I^{\text{busy}}(x,z) \triangleq \frac{x}{\mu z}\left[\frac{x+1}{2} + \frac{1}{1-\frac{k\lambda}{k\lambda+\mu z}}\right] = \frac{x}{\mu z}\left[\frac{x+1}{2} + \frac{R}{z}\right].$$

For the time integral of the queue length $Q(t)$ over a renewal cycle, we have

$$\mathbb{E}\left[\int_0^X Q(t)dt\right] \geq \frac{1}{2}\mu\beta^2 L_1\sqrt{R} + I^{\text{busy}}\left(\left[\mu\beta L_1\sqrt{R} - (k-R)\right]^+, k-R\right).$$

Lemma 5.2 (Upper bound on Accumulation Phase Length)**.** *Recall that*

$$T_A \triangleq \min\{t>0 : Z(t) \geq R+1\}$$

is the amount of time until the $(R+1)$-th server turns on. Then we can bound the expectation $\mathbb{E}[T_A]$ by

$$\mathbb{E}[T_A] \leq e^{\frac{1}{24R\beta}}\left(\sqrt{1+\frac{1}{2R\beta}}\right)\left[1+\frac{e^{\frac{1}{12R}}}{\sqrt{2\mu\beta}}\right]\beta \leq 1.08 * \beta.$$

Lemma 5.3 (Upper bound on Draining Phase Length)**.** *Recall that the accumlation time*

$$T_A \triangleq \min\{t>0 : Z(t) \geq R+1\}$$

is the amount of time until the $(R+1)$-th server turns on and the cycle time X is the moment when it turns off. Then, one can bound $\mathbb{E}[X-T_A]$ by

$$\mathbb{E}[X-T_A] \leq \beta + \frac{1}{\mu}\frac{F_1\beta\sqrt{R}}{k-R} + \frac{1}{\mu}\frac{3}{2}\ln(\beta) + \frac{1}{\mu}\ln(F_1 D_1) + \frac{2}{\mu} + \left[D_2 + \frac{D_3}{\sqrt{R}}\right]\max\left(\frac{1}{D_1\sqrt{\mu\beta}}, \frac{1}{\sqrt{R}}\right),$$

where F_1, D_1, D_2, and D_3 are constants not depending on system parameters.

5.3.3 Proof of Lemma 5.1: Lower Bound on Cycle Integral.

Lemma 5.1 Proof Outline

Basic Strategy. First, we split the first phase $[0, T_A)$ into epochs, where epoch i begins when the number of busy servers $Z(t)$ first drops to $R-i$, and an epoch ends either when the next epoch starts or when the first phase ends. Our goal will be to analyze a specific "significant" epoch. In particular, we say that an epoch is long if it lasts for longer than a setup time β. Because the accumulation phase ends when the $(R+1)$-th server turns on, at least one epoch must be long. We use L to denote the index of the *first* long epoch. From here, we argue via a martingale/coupling argument that the expected time integral over the first β time in epoch L is at least $\frac{1}{2}\beta^2\mathbb{E}[L]$. To bound the integral afterwards, we couple the behavior of the total number of jobs $N(t)$ to the queue length in an M/M/1 queue with arrival rate $k\lambda$ and departure rate $k\mu$.

Formalization. Define the stopping time $\tau_i \triangleq \min\{t \geq 0 : N(t) \leq R - u\}$ as the beginning of epoch i. We say that the epoch *occurs* is $\tau_i < T_A$, and define the end of epoch i as $\gamma_i \triangleq \min(\tau_{i+1}, T_A)$ the moment when either epoch $i+1$ begins or when the first phase ends. If epoch i occurs, we say it is long if $\gamma_i - \tau_i \geq \beta$. Let $L \triangleq \min\{i \in \mathbb{N} : \gamma_i - \tau_i \geq \beta\}$ be the index of the first long epoch. It suffices to show two claims; we state and prove them in sequence.

5.3.4 Lower Bound on Integral until $\tau_L + \beta$.

We show the following claim.

Claim 5.1. *Let L be the index of the first long epoch. Then,*

$$\mathbb{E}\left[\int_0^{\tau_L+\beta} Q(t)dt\right] \geq \frac{1}{2}\mu\beta^2 L_1\sqrt{R}, \tag{5.1}$$

where L_1 is some absolute constant.

Claim 5.1 Proof Strategy. First, we show that the initial integral is bounded by

$$\mathbb{E}\left[\int_0^{\tau_L+\beta} Q(t)\mathrm{d}t\right] \geq \frac{1}{2}\mu\beta^2\mathbb{E}[L]. \tag{5.2}$$

Afterwards, we give a bound on $\mathbb{E}[L]$, showing that

$$\mathbb{E}[L] \geq L_1\sqrt{R}. \tag{5.3}$$

Proof of (5.2)**, Bound in terms of** $\mathbb{E}[L]$**.**
To show (5.2), we first condition on whether $L \geq i$, giving

$$\begin{aligned}\mathbb{E}\left[\int_0^{\tau_L+\beta} Q(t)\mathrm{d}t\right] &= \sum_{i=0}^{\infty}\mathbb{E}\left[\int_{\tau_i}^{\min(\tau_i+\beta,\tau_{i+1})} Q(t)\mathrm{d}t\mathbf{1}_{L\geq i}\right]\\ &= \sum_{i=0}^{\infty}\mathbb{E}\left[\int_{\tau_i}^{\min(\tau_i+\beta,\tau_{i+1})} Q(t)\mathrm{d}t\middle|\mathcal{F}_{\tau_i}\right]\Pr(L \geq i).\end{aligned}$$

To further develop this conditional expectation, we note that during the interval $[\tau_i, \min(\tau_i+\beta, \tau_{i+1}))$, the system must have exactly $Z(t) = R - i$ busy servers running, meaning that $Q(t) = N(t) - (R - i)$. Defining a coupled process $\tilde{Q}(t)$ as

$$\tilde{Q}(t) = A(\tau_i, t) - \mathcal{D}[R-i](\tau_i, t),$$

we see that $Q(t)$ and $\tilde{Q}(t)$ coincide during the interval in question. Moreover, one can redefine the stopping time $\gamma = \tau_{i+1}$ as $\min\left\{t > \tau_i : \tilde{Q}(t) = -1\right\}$. Noting that $Q(\min(\gamma, t)) = -1$ for any time $t > \gamma$, we find that

$$\begin{aligned}
\int_{\tau_i}^{\min(\tau_i+\beta,\tau_{i+1})} Q(t)\mathrm{d}t &= \int_{\tau_i}^{\min(\tau_i+\beta,\tau_{i+1})} \tilde{Q}(t)\mathrm{d}t \\
&= \int_{\tau_i}^{\min(\tau_i+\beta,\tau_{i+1})} \tilde{Q}\left(\min\left(t,\tau_{i+1}\right)\right)\mathrm{d}t + \int_{\min(\tau_i+\beta,\tau_{i+1})}^{\tau_i+\beta} \left(\tilde{Q}\left(\min\left(t,\tau_{i+1}\right)\right)+1\right)\mathrm{d}t \\
&= \int_{\tau_i}^{\tau_i+\beta} \tilde{Q}\left(\min\left(t,\tau_{i+1}\right)\right)\mathrm{d}t + \left[\beta - \min\left(\beta, \tau_{i+1}-\tau_i\right)\right] \\
&\geq \int_{\tau_i}^{\tau_i+\beta} \tilde{Q}\left(\min\left(t,\tau_{i+1}\right)\right)\mathrm{d}t.
\end{aligned}$$

Taking the conditional expectation at time τ_i, we find

$$\mathbb{E}\left[\int_{\tau_i}^{\tau_i+\beta} \tilde{Q}\left(\min\left(t,\tau_{i+1}\right)\right)\mathrm{d}t \middle| \mathcal{F}_{\tau_i}\right] = \int_{\tau_i}^{\tau_i+\beta} \mathbb{E}\left[\tilde{Q}\left(\min\left(t,\tau_{i+1}\right)\right)\middle|\mathcal{F}_{\tau_i}\right]\mathrm{d}t.$$

Noting that $V_L(t) = \tilde{Q}(t) - \mu i\left[t-\tau_i\right]$ is a martingale, and that $\min\left(t,\tau_{i+1}\right)$ is an almost-surely bounded stopping time, we have that

$$\begin{aligned}
\tilde{Q}(\tau_i) = V_L(\tau_i) &= 0 \\
&= \mathbb{E}\left[V_L\left(\min\left(t,\tau_{i+1}\right)\right)|\tau_i\right] \\
&= \mathbb{E}\left[\tilde{Q}\left(\min\left(t,\tau_{i+1}\right)\right)\middle|\mathcal{F}_{\tau_i}\right] - \mu i\mathbb{E}\left[\min\left(t,\tau_{i+1}\right)|\mathcal{F}_{\tau_i}\right].
\end{aligned}$$

Since

$$\mathbb{E}\left[\min\left(t,\tau_{i+1}\right)|\mathcal{F}_{\tau_i}\right] \geq t\cdot\Pr\left(\tau_{i+1}-\tau_i \geq t\right) \geq t\cdot\Pr\left(\tau_{i+1}-\tau_i\geq\beta\right) = t\Pr\left(L=i|L\geq i\right),$$

we have

$$\begin{aligned}
\Pr\left(L\geq i\right)\mathbb{E}\left[\int_{\tau_i}^{\min(\tau_i+\beta,\tau_{i+1})} Q(t)\mathrm{d}t\middle| L\geq i\right] &\geq \Pr\left(L\geq i\right)\mathbb{E}\left[\int_{\tau_i}^{\tau_i+\beta}\tilde{Q}\left(\min\left(t,\tau_{i+1}\right)\right)\mathrm{d}t\middle| L\geq i\right] \\
&\geq \int_{\tau_i}^{\tau_i+\beta} \mu i t\Pr\left(L=i\right) \\
&= \mu\frac{\beta^2}{2} i\Pr\left(L=i\right)
\end{aligned}$$

Summing across all i, we obtain (5.2).

Proof Sketch for (5.3), bound on $\mathbb{E}\left[L\right]$.

We defer the full proof of this to Section A.9, and for now give a proof sketch.

We prove (5.3) by first showing that

$$\Pr\left(L>j|L\geq j\right) \geq \left(1-\frac{j}{R}\right)\left(1-\frac{b_1}{\sqrt{\mu\beta R}}\right),$$

where $b_1 = \frac{2}{\sqrt{\pi}}$. Next, we show that this implies that, for any $\delta \in (0, 1)$ and any $j < \delta R$,

$$\Pr(L > j) \geq \left(1 - \frac{b_1}{\sqrt{\mu\beta R}}\right)^{j+1} e^{-\frac{j(j+1)}{2R}\frac{1}{1-\delta}}.$$

From here, we use the sum of tails formula $\mathbb{E}[L] = \sum_{j=0}^{\infty} \Pr(L > j)$ to show

$$\mathbb{E}[L] \geq \left(1 - \frac{b_1}{\sqrt{\mu\beta R}}\right) \left(\left[\sqrt{\frac{\pi}{2}}(1-\delta) - \frac{1.15(1-\delta)}{\sqrt{\mu\beta}}\right]\sqrt{R} - \frac{1}{2} - \frac{2(1-\delta)}{\delta} e^{-R\frac{\delta^2}{1-\delta}}\right).$$

Choosing $\delta = \frac{2}{\sqrt{R}}$ then noting that $\mu\beta \geq 100$ and $R \geq 100$ gives the result.

5.3.5 Lower Bound on Integral after $\tau_L + \beta$.

To finish our lower bound on the integral, we now show the following claim.

Claim 5.2. *Let L be the index of the first long epoch. Then,*

$$\mathbb{E}\left[\int_{\tau_L+\beta}^{X} Q(t)dt\right] \geq I^{\text{busy}}\left(\left[\mu\beta L_1\sqrt{R} - (k-R)\right]^+, k-R\right) \tag{5.4}$$

where L_1 is some absolute constant.

Claim 5.2: Proof Strategy. First, we show that the remaining integral is bounded by

$$\mathbb{E}\left[\int_{\tau_L+\beta}^{X} Q(t)\mathrm{d}t\right] \geq I^{\text{busy}}\left([\mathbb{E}[N(\tau_L+\beta)] - k]^+, k-R\right). \tag{5.5}$$

Then, we use martingales again to show that

$$\mathbb{E}[N(\tau_L+\beta)] \geq R + \mu\beta\mathbb{E}[L]. \tag{5.6}$$

Applying (5.3), our bound on $\mathbb{E}[L]$, we obtain the result.

Proof of (5.5), Bound in terms of $\mathbb{E}[N(\cdot)]$.

To prove (5.5), we make a simple coupling argument. Let $\eta_k \triangleq \min\{t \geq \tau_L + \beta : N(t) \leq k\}$. Since the draining phase starts at $T_A \geq \tau_L + \beta$ and the end of the cycle $X = \min\{t \geq T_A : N(t) \leq R\}$, we know that $X \geq \eta_k$. Moreover, we know the number of busy servers $Z(t) \leq k$; it follows by Claim A.1 that we can define $\tilde{N}(t)$ as

$$\tilde{N}(t) \triangleq N(\tau_L+\beta) + A(\tau_L+\beta, t) - \mathcal{D}[k]((\tau_L+\beta, t))$$

and have $\tilde{N}(t) \leq N(t)$ for any $t > \tau_L + \beta$. Even further, we can defined a coupled hitting time $\tilde{\eta}_k \triangleq \min\left\{t > \tau_L + \beta : \tilde{N}(t) \leq k\right\}$ which must happen before η_k. In other words,

$$\begin{aligned}\int_{\tau_L+\beta}^{X} Q(t)\mathrm{d}t &\geq \int_{\tau_L+\beta}^{\eta_k} Q(t)\mathrm{d}t \\ &\geq \int_{\tau_L+\beta}^{\eta_k} \left[N(t) - k\right]\mathrm{d}t \\ &\geq \int_{\tau_L+\beta}^{\tilde{\eta}_k} \left[\tilde{N}(t) - k\right]\mathrm{d}t.\end{aligned}$$

This final term is just the time integral of the number of jobs in system over a M/M/1 busy period started by $\left[N\left(\tau_L+\beta\right) - k\right]^+$ jobs, where jobs arrive at rate $k\lambda$ and depart at rate $k\mu$. Accordingly, we have

$$\begin{aligned}\mathbb{E}\left[\int_{\tau_L+\beta}^{X} Q(t)\mathrm{d}t\right] &\geq \mathbb{E}\left[I^{\text{busy}}\left(\left[N\left(\tau_L+\beta\right)-k\right]^+, k-R\right)\right] \\ &\geq I^{\text{busy}}\left(\mathbb{E}\left[N\left(\tau_L+\beta\right)-k\right]^+, k-R\right) \\ &\geq I^{\text{busy}}\left(\mathbb{E}\left[\left[N\left(\tau_L+\beta\right)-R-(k-R)\right]^+\right], k-R\right),\end{aligned}$$

where in the last two lines we have applied Jensen's inequality. □

Proof of (5.6), Bound on $\mathbb{E}\left[N(\cdot)\right]$.

To bound $\mathbb{E}\left[N(\tau_L+\beta)\right]$, we condition on the value of L, then make a martingale argument.

$$\begin{aligned}\mathbb{E}\left[N(\tau_L+\beta)\right] &= \sum_{i=0}^{R} \mathbb{E}\left[N(\tau_i+\beta)\mathbf{1}_{L=i}\right] \\ &= \sum_{i=0}^{R} \mathbb{E}\left[N(\tau_i+\beta)\mathbf{1}_{L=i,L\geq i}\right] \\ &\geq \sum_{i=0}^{R} \Pr\left(L\geq i\right)\mathbb{E}\left[N(\tau_i+\beta)\mathbf{1}_{L=i}|L\geq i\right].\end{aligned}$$

Continuing with this conditional expectation,

$$\begin{aligned}&\mathbb{E}\left[N(\tau_i+\beta)\mathbf{1}_{L=i}|\mathcal{F}_{\tau_i}\right] \\ &= \mathbb{E}\left[N(\tau_i+\beta)\mathbf{1}_{\tau_i+\beta<\tau_{i+1}}\middle|\mathcal{F}_{\tau_i}\right] \\ &= \mathbb{E}\left[\left[N(\tau_i+\beta)-(R-(i+1))\right]\mathbf{1}_{\tau_i+\beta<\tau_{i+1}}\middle|\mathcal{F}_{\tau_i}\right] + (R-i-1)\Pr\left(\tau_i+\beta<\tau_{i+1}|\mathcal{F}_{\tau_i}\right) \\ &= \mathbb{E}\left[N(\min\left(\tau_i+\beta,\tau_{i+1}\right)) - (R-(i+1))|\mathcal{F}_{\tau_i}\right] + +(R-i-1)\Pr\left(\tau_i+\beta<\tau_{i+1}|\mathcal{F}_{\tau_i}\right) \\ &= 1 + \mu i\mathbb{E}\left[\min\left(\beta,\tau_{i+1}-\tau_i\right)\right] + (R-i-1)\Pr\left(\tau_i+\beta<\tau_{i+1}\right) \\ &\geq 1 + \mu i\beta\Pr\left(\tau_{i+1}-\tau_i\geq\beta\right) + (R-i-1)\Pr\left(\tau_i+\beta<\tau_{i+1}\right) \\ &= 1 + \mu i\beta\Pr\left(L=i|L\geq i\right) + (R-i-1)\Pr\left(L=i|L\geq i\right).\end{aligned} \tag{5.7}$$

Summing across i, we find

$$\begin{aligned}\mathbb{E}\left[N(\tau_L+\beta)\right] &= \sum_{i=0}^{R} \Pr\left(L \geq i\right)((5.7)) \\ &= (1+\mathbb{E}\left[L\right]) + (\mu\beta\mathbb{E}\left[L\right]) + (R - \mathbb{E}\left[L\right] - 1) \\ &= R + \mu\beta\mathbb{E}\left[L\right],\end{aligned}$$

as desired. □

Combining Claims 5.1 and 5.2, we obtain a lower bound on $\mathbb{E}\left[\int_0^X Q(t)\mathrm{d}t\right]$, proving Lemma 5.1.

5.3.6 Proof of Lemma 5.2: Upper Bound on the Accumulation Time $\mathbb{E}\left[T_A\right]$

Defining a coupling. To prove Lemma 5.2, we first note that, during the accumulation phase, we have two bounds on the number of busy servers $Z(t)$: it must be less than the total number of jobs $N(t)$ and it must be less than R; the former because every busy server must be working on a job, and the latter because otherwise the accumulation phase would be over. Thus, we can define a coupled $M/M/R$ system for which the number of jobs $\tilde{N}(t)$ in the coupled system is always at least the number of jobs $N(t)$ in the original system.

How we use the coupling. To use this coupled process to bound $\mathbb{E}\left[T_A\right]$, recall that the accumulation point T_A is the first time the $(R+1)$-th server turns on. Accordingly, one can also think of this as the first time that there has been at least $R+1$ jobs in the system for β time. Thus, if we define a coupled accumulation point $\tilde{T}_A \triangleq \min\left\{t \geq \beta : \min_{s\in[t-\beta,t)} \tilde{N}(t) \geq R+1\right\}$, then we know $\tilde{T}_A \geq T_A$. In other words, it suffices to bound $\mathbb{E}\left[\tilde{T}_A\right]$.

General Strategy. We bound $\mathbb{E}\left[\tilde{T}_A\right]$ using the MIST method of Lemma 4.1. As such, we define a few stopping times, then list the preconditions/claims that we will satisfy to complete the proof of Lemma 5.2.

Definition of γ and α. Let the initial cycle-downcrossing occur at $\alpha_0 \triangleq 0$ and iteratively define the upcrossings γ and downcrossings α as

$$\gamma_i \triangleq \min\left\{t \geq \alpha_i : \tilde{N}(t) \geq R+1\right\}$$

and

$$\alpha_{i+1} \triangleq \min\left\{t \geq \gamma_i : \tilde{N}(t) \geq R+1\right\}.$$

Application of Lemma 4.1, the IST Lemma. Applying Lemma 4.1 using $0 = \alpha_0$ as our starting point, the coupled accumulation point $\tilde{T}_A$ as our ending point, our test function as $Y_t = 1$, and the cycle-upcrossings (α_i) as our intervening stopping times, we now must prove that

$$\mathbb{E}\left[\gamma_i - \alpha_i | n_\alpha \geq i\right] \leq \frac{1}{\mu} e^{\frac{1}{12R}} \sqrt{1 + \frac{1}{R}} \frac{\sqrt{2\pi}}{\sqrt{R}} \leq \frac{c_3}{\mu\sqrt{R}}, \tag{5.8}$$

$$\mathbb{E}\left[\min\left(\tilde{T}_A, \alpha_{i+1}\right) - \gamma_i \middle| n_\alpha \geq i\right] \leq b_1 \sqrt{\frac{\beta}{\mu R}} + \frac{6}{\mu R}, \tag{5.9}$$

and

$$\Pr\left(n_\alpha \geq i+1 | n_\alpha \geq i\right) \leq 1 - \frac{b_1}{\sqrt{2}} e^{-\frac{1}{3(\mu 2R\beta - 1)}} \frac{1}{\sqrt{\mu 2R\beta + 2}} \leq 1 - \frac{b_1 c_4}{\sqrt{\mu R\beta}}, \tag{5.10}$$

where $b_1 \triangleq \sqrt{\frac{2}{\pi}}$, $c_3 = 1.001\sqrt{2\pi}$, and $c_4 = 0.499$.

Completion of Proof, assuming (5.8), (5.9), and (5.10). Applying Lemma 4.1, one finds that

$$\begin{aligned}
\mathbb{E}\left[\tilde{T}_A\right] &= \sum_{i=0}^{\infty} \mathbb{E}\left[\min\left(\tilde{T}_A, \alpha_{i+1}\right) - \alpha_i \middle| n_\alpha \geq i\right] \Pr\left(n_\alpha \geq i\right) \\
&\leq \left[\frac{c_3}{\mu\sqrt{R}} + \frac{b_1\sqrt{\beta}}{\sqrt{\mu R}} + \frac{6}{\mu R}\right] \frac{\sqrt{\mu R\beta}}{b_1 c_4} \\
&= \frac{1}{\mu}\left[\frac{c_3}{b_1 c_4}\sqrt{\mu\beta} + \frac{1}{c_4}\beta + \frac{6}{b_1 c_4}\sqrt{\frac{\mu\beta}{R}}\right].
\end{aligned}$$

Proof of (5.8): Upper bound on initial up-crossing time.

To prove (5.8), we note that, since our coupled system is an $M/M/R$, the expected time $\mathbb{E}\left[\gamma_i - \alpha_i | n_\alpha \geq i\right]$ is simply the expected passage time from state R to $(R+1)$ in an $M/M/R$ (and equivalently an

$M/M/R/(R+1)$, an $M/M/R$ which can contain only $R+1$ jobs. Solving, one finds that

$$\begin{aligned}
\mathbb{E}\left[T_{R\to(R+1)}\right] &\leq \mathbb{E}\left[T_{(R+1)\to(R+1)}\right] \\
&= \frac{1}{\mu(R+1)}\frac{1}{\pi_{R+1}} \\
&= \frac{1}{\mu(R+1)}\frac{\sum_{i=0}^{R+1}\frac{R^i}{i!}}{\frac{R^{R+1}}{(R+1)!}} \\
&\leq \frac{1}{\mu(R+1)}e^R\frac{(R+1)!}{R^{R+1}} \\
&\leq \frac{1}{\mu(R+1)}e^R\frac{e^{\frac{1}{12(R+1}}\sqrt{2\pi(R+1)}(R+1)^{R+1}e^{-(R+1)}}{R^{R+1}} \\
&= e^{\frac{1}{12R}}\frac{1}{\mu}\sqrt{2\pi}\frac{\sqrt{R+1}}{R}\left(1+\frac{1}{R}\right)^R e^{-1} \\
&\leq \frac{1}{\mu}e^{\frac{1}{12R}}\sqrt{1+\frac{1}{R}}\frac{\sqrt{2\pi}}{\sqrt{R}} \\
&\leq \frac{1}{\mu}1.006\frac{\sqrt{2\pi}}{\sqrt{R}} \\
&\triangleq \frac{c_3}{\mu\sqrt{R}}
\end{aligned}$$

Proof of (5.9)**: Bound on time between up-crossings.**

To bound the expected time $\mathbb{E}\left[\min\left(\tilde{T}_A, \alpha_{i+1}\right) - \gamma_i \middle| n_\alpha \geq i\right]$, we first note that, if $\gamma_i+\beta \leq \alpha_{i+1}$, then $\tilde{T}_A = \gamma_i + \beta$. Likewise, if $\gamma_i + \beta > \alpha_{i+1}$, then $\tilde{T}_A > \alpha_{i+1}$. It follows that, given that $n_\alpha \geq i$, the time $\min\left(\tilde{T}_A, \alpha_{i+1}\right) = \min\left(\beta+\gamma_i, \alpha_{i+1}\right)$. Thus, we have that

$$\mathbb{E}\left[\min\left(\tilde{T}_A, \alpha_{i+1}\right) - \gamma_i \middle| n_\alpha \geq i\right] = \mathbb{E}\left[\min\left(\beta, \alpha_{i+1}-\gamma_i\right)|n_\alpha \geq i\right] = \int_0^\beta \Pr\left(\alpha_{i+1}-\gamma_i > s|n_\alpha \geq i\right)\mathrm{d}s.$$

We continue by bounding this tail probability. To begin, note that, while $\tilde{N}(t)$ stays above $R+1$, the dynamics of $\tilde{N}$ are precisely that of a critically-loaded $M/M/1$ queue with arrival rate and departure rate equal to $k\lambda$. The tail probability we are interested in bounding is precisely the probability that a busy period (started with 1 job) in such a system lasts longer than s time. Applying Claim A.8, one finds that, for any $t \geq \frac{3}{\mu 2R}$,

$$\Pr\left(\alpha_{i+1}-\gamma_i > s|n_\alpha \geq i\right) \leq b_1\left(\frac{1}{\sqrt{\mu 2Rs}} + \frac{b_2}{(\mu 2Rs)^{3/2}}\right).$$

Integrating, we find that

$$\begin{aligned}\int_0^\beta \Pr\left(\alpha_{i+1}-\gamma_i > s | n_\alpha \geq i\right) \mathrm{d}s &\leq \frac{3}{\mu 2R} + \frac{b_1}{\sqrt{2}} \int_{\frac{3}{\mu 2R}}^{\beta} \frac{1}{\sqrt{\mu 2Rs}} + \frac{b_2}{(\mu 2Rs)^{3/2}} \mathrm{d}s \\ &\leq \frac{3}{\mu 2R} + \frac{b_1}{\sqrt{2}} \left[\sqrt{\frac{2\beta}{\mu R}} + b_2 \sqrt{\frac{2}{3}} \frac{1}{\mu R}\right] \\ &= b_1 \sqrt{\frac{\beta}{\mu R}} + \left(\frac{3}{2} + (b_1 + 2.5)\sqrt{\frac{2}{3}}\right) \frac{1}{\mu R} \\ &\leq \frac{2}{\sqrt{\pi}} \sqrt{\frac{\beta}{\mu R}} + \frac{6}{\mu R}.\end{aligned}$$

□

Proof of (5.10): Bound on probability of another γ up-crossing.

To prove (5.10), it suffices to note that, upon conditioning on the filtration at γ_i, the probability $\Pr\left(n_\alpha \geq i+1 | n_\alpha \geq i\right)$ is simply the probability that a busy period in a critically-loaded M/M/1, with arrival and departure rate equal to μR, ends before β time has passed. Applying Claim A.8, one finds that this is

$$\Pr\left(n_\alpha \geq i+1 | n_\alpha \geq i\right) \geq 1 - \frac{b_1}{\sqrt{2}} e^{-\frac{1}{3(\mu 2R\beta - 1)}} \frac{1}{\sqrt{\mu 2R\beta + 2}}.$$

□

5.3.7 Proof of Lemma 5.3: Upper Bound on the Cycle Time $\mathbb{E}\left[X - T_A\right]$.

We now prove the upper bound on $\mathbb{E}\left[X - T_A\right]$. We make use of the "wait-busy" idea from Section 6.2.2 as well as our main tool, Lemma 4.1. As such, we begin by defining some stopping times.

Definition of $\zeta_i^{(\mathbf{d})}$ and $\zeta_i^{(\mathbf{u})}$. Recall that the draining phase begins at time T_A. Let $M_L \triangleq \min\left(k - R, \max\left(\frac{\sqrt{R}}{D_1\sqrt{\beta}}, 1\right)\right)$ be a specially-set analysis threshold. Let the stopping time $\zeta_1^{(\mathrm{d})} \triangleq \min\left\{t \geq T_A : N(t) < R + M_L\right\}$ be the first time the number of jobs $N(t)$ drops below $R + M_L$, and recursively define

$$\zeta_i^{(\mathrm{u})} \triangleq \min\left\{t \geq \zeta_i^{(\mathrm{d})} : N(t) \geq R + M_L\right\}$$

and

$$\zeta_{i+1}^{(\mathrm{d})} \triangleq \min\left\{t \geq \zeta_i^{(\mathrm{u})} : N(t) < R + M_L\right\}.$$

Specification Step. Now, we apply Lemma 4.1 using the accumulation point T_A as our initial point, the cycle end X as our ending point, the constant function $Y_t = 1$ as our test function, and the draining-downcrossing points $\left(\zeta_i^{(\mathrm{d})}\right)$ as our intervening points; we use n_ζ to count the number of intervening points. To complete the proof, we must show that the following claims:

$$\mathbb{E}\left[\zeta_1^{(\mathrm{d})} - T_A\right] \leq \beta + \frac{1}{\mu}\frac{F_1\beta\sqrt{R}}{k-R} + \frac{1}{\mu}\frac{3}{2}\ln(\beta) + \frac{1}{\mu}\ln(F_1 D_1), \tag{5.11}$$

$$\mathbb{E}\left[\min\left(X, \zeta_{i+1}^{(\mathrm{d})}\right) - \zeta_i^{(\mathrm{d})} \middle| n_\zeta \geq i\right] \leq \frac{D_2}{\mu\sqrt{R}} + \frac{D_3}{\mu R} + \frac{2}{\mu M_L} \tag{5.12}$$

$$\Pr\left(n_\zeta \geq i+1 | n_\zeta \geq i\right) \leq \frac{1}{M_L}. \tag{5.13}$$

Completion of Proof assuming (5.11)**,** (5.12)**, and** (5.13)**.** Before proving the claims, we now prove the lemma. It suffices to give a bound on $\mathbb{E}\left[X - \zeta_1^{(\mathrm{d})}\right]$; applying Lemma 4.1 gives

$$\begin{aligned}\mathbb{E}\left[X - \zeta_1^{(\mathrm{d})}\right] &\leq M_L\left[\frac{D_2}{\mu\sqrt{R}} + \frac{D_3}{\mu R} + \frac{2}{\mu M_L}\right] \\ &= \frac{2}{\mu} + \left[D_2 + \frac{D_3}{\sqrt{R}}\right]\max\left(\frac{1}{D_1\sqrt{\mu\beta}}, \frac{1}{\sqrt{R}}\right).\end{aligned}$$

Proof of (5.11)**.**

To bound $\mathbb{E}\left[\zeta_1^{(\mathrm{d})} - T_A\right]$, we make a coupling argument then apply basic results on M/M/1 busy periods. Moreover, instead of proving (5.11) directly, we first show a more general claim.

Claim 5.3. *For $M_L \leq j \leq N(T_A) - R$, define η_j as the first time after T_A that $N(t) \leq R + j$. Note that this means that $\eta_{N(T_A)} = T_A$ and $\eta_{M_L} = \zeta_1^{(\mathrm{d})}$. Then we have the following bound:*

$$\mathbb{E}\left[\eta_{M_L} - \eta_j \middle| \mathcal{F}_{\eta_j}\right] \leq Y_{R+j}(\eta_j) + \frac{1}{\mu}\sum_{i=M_L}^{j}\frac{1}{\min(i, k-R)}.$$

Afterwards, we complete the proof by noting that $[N(T_A) - k]^+ \leq [N(T_A) - R]$, taking expectations, applying Jensen's inequality to the minimum function and the $\ln(\cdot)$ (which is concave), using the bound on $\mathbb{E}[N(T_A) - R]$ from Claim 6.7, then letting $h = M_L$.

Proof of Claim 5.3. We prove Claim 5.3 by induction. In the base case, suppose that $j = M_L + 1$. Note that at time η_{M_L+1}, the numbers of jobs $N(\eta_{M_L+1} = R + M_L + 1$ and the remaining time until the $(R + M_L + 1)$-th server turns on is $Y_{R+M_L+1}(\eta_{M_L+1})$. As such, we can simply wait until either that server turns on, in which case we can analyze the system as an M/M/1 busy period with departure rate $\mu\min(R + M_L + 1, k)$, or the number of jobs $N(t)$ drops below $R + M_L + 1$ on its own. In other words, (using j here to save space)

$$\mathbb{E}\left[\eta_{j-1} - \eta_j \middle| \mathcal{F}_{\eta_j}\right] \leq Y_{R+j}(\eta_j) + \frac{\mathbb{E}\left[\left[N\left(\eta_j + Y_{R+j}(\eta_j)\right) - (R + (j-1))\right]\mathbf{1}_{\eta_{j-1} > \eta_j + Y_{R+j}(\eta_j)} \middle| \mathcal{F}_{\eta_j}\right]}{\mu\min(j, k-R)}.$$

Now, we reframe the expectation as an expectation up to a stopping time. We note that, if $\eta_{j-1} > \eta_j + Y_{R+j}(\eta_j)$, then we have that

$$N(\eta_j + Y_{R+j}(\eta_j)) = N(\min(\eta_j + Y_{R+j}(\eta_j), \eta_{j-1})).$$

Likewise, if $\eta_{j-1} \leq \eta_j + Y_{R+j}(\eta_j)$, then

$$R + j - 1 = N(\eta_{j-1}) = N(\min(\eta_j + Y_{R+j}(\eta_j), \eta_{j-1})).$$

Using this and applying a simple coupling argument, one sees that

$$\begin{aligned}
&\mathbb{E}\left[\left[N(\eta_j + Y_{R+j}(\eta_j)) - (R + (j-1))\right]\mathbf{1}_{\eta_{j-1} > \eta_j + Y_{R+j}(\eta_j)} \middle| \mathcal{F}_{\eta_j}\right] \\
&= \mathbb{E}\left[N(\min(\eta_j + Y_{R+j}(\eta_j), \eta_{j-1})) - (R + j - 1) \middle| \mathcal{F}_{\eta_j}\right] \\
&\leq N(\eta_j) - (R + j - 1) = 1.
\end{aligned}$$

Thus, we find that

$$\mathbb{E}\left[\eta_{j-1} - \eta_j \middle| \mathcal{F}_{\eta_j}\right] \leq Y_{R+j}(\eta_j) + \frac{1}{\mu \min(j, k-R)}.$$

Inductive case. The inductive case proceeds in much the same way, except now, if $N(t)$ does drop below $R+j$ "early", then we can factor in the time that has elapsed in the value of $Y_{R+j}(\eta_j)$. In particular, note that, since the $(R+j)$-th server would have already turned on,

$$\mathbb{E}\left[\eta_{M_L} - \eta_j \middle| \mathcal{F}_{\eta_j}\right]\mathbf{1}_{\eta_j \geq \eta_{j+1} + Y_{R+j+1}(\eta_{j+1})} \leq \frac{1}{\mu}\sum_{i=M_L}^{j} \frac{1}{\mu \min(i, k-R)}\mathbf{1}_{\eta_j \geq \eta_{j+1} + Y_{R+j+1}(j+1)}.$$

It follows that

$$\mathbb{E}\left[\eta_{M_L} - \eta_j \middle| \mathcal{F}_{\eta_j}\right] \leq Y_{R+j}(\eta_j)\mathbf{1}_{\eta_j < \eta_{j+1} + Y_{R+j+1}(\eta_{j+1})} + \frac{1}{\mu}\sum_{i=M_L}^{j} \frac{1}{\mu \min(i, k-R)}.$$

Now, we note that

$$\begin{aligned}
Y_{R+j}(\eta_j)\mathbf{1}_{\eta_j < \eta_{j+1} + Y_{R+1+j}(\eta_{j+1})} &= \left[Y_{R+j}(\eta_j) + \eta_j - \eta_j\right]\mathbf{1}_{\eta_j < \eta_{j+1} + Y_{R+1+j}(\eta_{j+1})} \\
&= \left[Y_{R+j}(\eta_{j+1}) + \eta_{j+1} - \eta_j\right]\mathbf{1}_{\eta_j < \eta_{j+1} + Y_{R+1+j}(\eta_{j+1})} \\
&\leq \left[Y_{R+j+1}(\eta_{j+1}) + \eta_{j+1} - \eta_j\right]\mathbf{1}_{\eta_j < \eta_{j+1} + Y_{R+1+j}(\eta_{j+1})} \\
&= \left[Y_{R+j+1}(\eta_{j+1}) + \eta_{j+1} - \eta_j\right]^+,
\end{aligned}$$

so that we find

$$\mathbb{E}\left[\eta_{M_L} - \eta_j \middle| \mathcal{F}_{\eta_j}\right] \leq \left[Y_{R+j+1}(\eta_{j+1}) + \eta_{j+1} - \eta_j\right]^+ + \frac{1}{\mu}\sum_{i=M_L}^{j} \frac{1}{\mu \min(i, k-R)}.$$

Finally, we note that

$$\mathbb{E}\left[\eta_j - \eta_{j+1} \middle| \mathcal{F}_{\eta_{j+1}}\right] \leq \mathbb{E}\left[\min(\eta_j - \eta_{j+1}, Y_{R+j+1}(\eta_{j+1})) \middle| \mathcal{F}_{\eta_{j+1}}\right] + \frac{1}{\mu \min(j+1, k-R)}.$$

Summing these final two expressions gives the inductive result, proving Claim 5.3.

Using Claim 5.3. Thus, we obtain that, using H_i to denote the i-th harmonic number,

$$\begin{aligned}\mathbb{E}\left[\zeta_1^{(\mathrm{d})} - T_A \middle| \mathcal{F}_{T_A}\right] &\leq \beta + \frac{1}{\mu}\frac{[N(T_A)-k]^+}{k-R} + \frac{1}{\mu}\left[H_{\min(N(T_A)-R,k-R)} - H_{M_L}\right] \\ &\leq \beta + \frac{1}{\mu}\frac{[N(T_A)-R]^+}{k-R} + \frac{1}{\mu}\ln\left(\frac{\min\left(N(T_A)-R,k-R\right)}{M_L}\right).\end{aligned}$$

Taking expectations and applying Jensen's inequality twice, we find

$$\begin{aligned}\mathbb{E}\left[\zeta_1^{(\mathrm{d})} - T_A \middle| \mathcal{F}_{T_A}\right] &\leq \beta + \frac{1}{\mu}\frac{F_1\mu\beta\sqrt{R}}{k-R} + \frac{1}{\mu}\ln\left(\frac{F_1\mu\beta\sqrt{R}}{M_L}\right) \\ &\leq \beta + \frac{1}{\mu}\frac{F_1\mu\beta\sqrt{R}}{k-R} + \frac{1}{\mu}\ln\left(\frac{\min\left(F_1\mu\beta\sqrt{R},k-R\right)}{\min\left(\max\left(1,\frac{\sqrt{R}}{D_1\sqrt{\beta}}\right),k-R\right)}\right) \\ &\leq \beta + \frac{1}{\mu}\frac{F_1\mu\beta\sqrt{R}}{k-R} + \frac{1}{\mu}\ln\left(F_1D_1\beta^{3/2}\right) \\ &= \beta + \frac{1}{\mu}\frac{F_1\mu\beta\sqrt{R}}{k-R} + \frac{1}{\mu}\frac{3}{2}\ln(\beta) + \frac{1}{\mu}\ln\left(F_1D_1\right).\square\end{aligned}$$

Proof of (5.12)**.**

To bound the expectation $\mathbb{E}\left[\min\left(\zeta_{i+1}^{(\mathrm{d})},X\right) - \zeta_i^{(\mathrm{d})}\middle|\mathcal{F}_{\zeta_i^{(\mathrm{d})}}\right]$, we split the interval into two parts, $\left[\min\left(\zeta_i^{(\mathrm{u})},X\right) - \zeta_i^{(\mathrm{d})}\right]$ and $\left[\zeta_{i+1}^{(\mathrm{d})} - \zeta_i^{(\mathrm{d})}\right]$.

To bound the expectation of the first quantity, it suffices to note that, if we couple the system to an M/M/∞, the coupled number of jobs $\tilde{N}(t)$ will reach $R+M_L$ only after the original system. Using Claim A.10 to bound this passage time, we thus know that

$$\begin{aligned}\mathbb{E}\left[\min\left(\zeta_i^{(\mathrm{u})},X\right) - \zeta_i^{(\mathrm{d})}\middle|\mathcal{F}_{\zeta_i^{(\mathrm{d})}}\right] &\leq \mathbb{E}\left[\min\left(T^{M/M/\infty}_{(R+M_L-1)\to(R+M_L)} + \zeta_i^{(\mathrm{d})},X\right) - \zeta_i^{(\mathrm{d})}\middle|\mathcal{F}_{\zeta_i^{(\mathrm{d})}}\right] \\ &\leq \mathbb{E}\left[T^{M/M/\infty}_{(R+M_L-1)\to(R+M_L)}\right] \\ &\leq \frac{D_2}{\sqrt{R}}.\end{aligned}$$

To bound the expectation of the second quantity, we provide two bounds. First, we again make use of the "wait-busy" idea; as we argued in the proof of (5.11),

$$\mathbb{E}\left[\zeta_{i+1}^{(\mathrm{d})} - \zeta_i^{(\mathrm{u})}\middle|\mathcal{F}_{\zeta_i^{(\mathrm{u})}}\right] \leq \mathbb{E}\left[\min\left(\zeta_{i+1}^{(\mathrm{d})} - \zeta_i^{(\mathrm{u})},\beta\right)\middle|\mathcal{F}_{\zeta_i^{(\mathrm{u})}}\right] + \frac{1}{\mu M_L}.$$

From here, we note, by coupling to an M/M/1 with arrival rate and departure rate both equal to $k\lambda$, we can bound $\mathbb{E}\left[\min\left(\zeta_{i+1}^{(\mathrm{d})} - \zeta_i^{(\mathrm{u})},\beta\right)\middle|\zeta_i^{(\mathrm{u})} < X\right]$ by the expected minimum between β and the length of a single-job busy period in that system. Applying Claim A.9, we can complete the proof, finding that

$$\mathbb{E}\left[\min\left(\zeta_{i+1}^{(\mathrm{d})} - \zeta_i^{(\mathrm{u})}, \beta\right) \Big| \mathcal{F}_{\zeta_i^{(\mathrm{u})}}\right] \le D_1 \frac{\sqrt{\beta}}{\sqrt{\mu R}} + \frac{6}{\mu R}.$$

For the second bound, we simply note that, during the draining phase, the number of busy servers $Z(t) \ge R+1$. It follows from a simple coupling argument that

$$\mathbb{E}\left[\min\left(\zeta_{i+1}^{(\mathrm{d})} - \zeta_i^{(\mathrm{u})}, \beta\right) \Big| \mathcal{F}_{\zeta_i^{(\mathrm{u})}}\right] \le \frac{1}{\mu}.$$

Combining the bounds pessimistically, we find that

$$\begin{aligned}\mathbb{E}\left[\min\left(\zeta_{i+1}^{(\mathrm{d})}, X\right) - \zeta_i^{(\mathrm{d})} \Big| \mathcal{F}_{\zeta_i^{(\mathrm{d})}}\right] &\le \frac{D_2}{\mu\sqrt{R}} + \frac{D_3}{\mu R} + \frac{1}{\mu M_L} + \min\left(D_1 \frac{\sqrt{\beta}}{\sqrt{\mu R}}, \frac{1}{\mu}\right) \\ &\le \frac{D_2}{\mu\sqrt{R}} + \frac{D_3}{\mu R} + \frac{2}{\mu M_L}\end{aligned}$$

Proof of (5.13)**.**

To bound the probability of an additional downcrossing, we again make a coupling argument. In particular, we couple again to the system which only has R servers busy, which gives an upper bound on the number of jobs in the system $N(t)$. If, in our coupled system, we reach $\tilde{N}(t) = R + M_L$ before we reach $\tilde{N}(t) = R$, then another upcrossing *must* have previously occurred in the original system, and thus another downcrossing must also occur. But, of course, we know classically that the probability that this happens is just $\frac{1}{M_L}$; this is precisely what is asserted by (5.13). □

5.4 The Lower Bounds: Review of Findings

In this chapter, we proved two lower bounds on the average waiting time in the M/M/k/Setup-Deterministic. The first lower bound, Theorem 5.1, was the first-ever explicit result for the average waiting time in this model. The second lower bound, Theorem 5.2, is a considerable strengthening of Theorem 5.1, and also was far easier to prove once we made use of the MIST method.

Chapter 6

The Upper Bound

In this chapter, we present our upper bound on the average waiting time in the M/M/k/Setup-Deterministic.

6.1 Why we need an upper bound.

From a provisioning standpoint, an upper bound tells us what system parameters **sufficient** to achieve a certain average waiting time. By combining this bound with our lower bound, we find out what is necessary and sufficient for good performance. Theoretically-speaking, having the two bounds allows us to fully characterize how the average waiting time in the M/M/k/Setup-Deterministic scales with its system parameters, modulo some constant multiplicative factors.

6.2 The Upper Bound

We now state and prove the upper bound.

Theorem 6.1 (Upper Bound on Average Queue Length)**.** *For an M/M/k/Setup-Deterministic with an offered load $R \triangleq k\rho \geq 100$ and a setup time $\beta \geq 1000\frac{1}{\mu}$, the expected number of jobs in queue in steady state is upper-bounded as*

$$\mathbb{E}\left[Q(\infty)\right] \leq A_1\sqrt{\mu\beta R}+A_2\frac{R}{M}+\frac{A_3\beta^2\mu\sqrt{R}+I^{\text{busy}}\left(B_5\sqrt{\mu\beta R}+B_6\mu\beta\sqrt{R},M\right)+A_4I^{\text{busy}}\left(M,M\right)}{\beta+T^{\text{busy}}\left(D_1\beta\mu\sqrt{R},k-R\right)},$$

where A_1, A_2, A_3, A_4, B_5,B_6, and D_1 are constants independent of system parameters, and

$$M \triangleq \min\left(C_1\sqrt{\mu\beta R}, k-R\right)$$

for some constant C_1 independent of system parameters.

We now describe the full proof of Theorem 6.1. As discussed in Chapter 4, it suffices to prove the three following lemmas.

Lemma 6.1 (Accumulation Period Upper Bound)**.** *Suppose the system begins at time* 0 *with* R *jobs in service and no jobs in the queue (and thus no servers in setup), and define the accumulation time*

$$T_A \triangleq \min\{t \geq 0 : Z(t) = R+1\}$$

to be the moment the $(R+1)$*-th server turns on.*

Then,

$$\mathbb{E}\left[\int_0^{T_A} [N(t) - R]dt\right] \leq B_1\sqrt{\mu\beta R} \cdot \mathbb{E}\left[T_A\right] + B_2\beta^2\mu\sqrt{R},$$

where $B_1 = 3.6$ *and* $B_2 = 1.04$.

Lemma 6.2 (Draining Period Upper Bound)**.** *Recall that accumulation time* T_A *is the first (and only) time the* $(R+1)$*-th server turns on during a renewal cycle, and that the next renewal point* $X = \min\{t > T_A : Z(t) = R\}$ *is simply the next time the* $(R+1)$*-th server turns off. Then,*

$$\begin{aligned}
&\mathbb{E}\left[\int_{T_A}^{X} [N(t) - R]^+ dt\right] \\
&\leq \left(B_5\sqrt{\mu\beta R} + B_6\mu\beta\sqrt{R}\right)\cdot\beta + I^{\text{busy}}\left(B_5\sqrt{\mu\beta R} + B_6\mu\beta\sqrt{R}, M\right) + B_7\frac{\beta R}{M} \\
&+ \left[2M + 2\frac{R}{M}\right]\mathbb{E}\left[X - T_A\right] + \frac{1}{1-p_2} I^{\text{busy}}\left(M, M\right),
\end{aligned}$$

where all of these quantities are defined in Chapter 3 and Section A.6.

Lemma 6.3 (Cycle Length Lower Bound)**.** *Suppose the system begins at time* 0 *with* R *jobs in service and no jobs in the queue (and thus no servers in setup), and let*

$$X \triangleq \min\left\{t > 0 : Z(t^-) = R+1, Z(t) = R\right\}$$

be the next time the $(R+1)$*-th server turns off.*

Then,

$$\mathbb{E}\left[X\right] \geq \beta + T^{\text{busy}}\left(D_1\beta\mu\sqrt{R}, k - R\right),$$

where D_1 *is a constant independent of system parameters.*

After proving these lemmas, the result follows by a bit of algebra. First, note that, by summing the two integral bounds, one obtains

$$\begin{aligned}
&\mathbb{E}\left[\int_0^X [N(t)-R]\mathrm{d}t\right] \\
&\quad \leq B_1\sqrt{\mu\beta R}\cdot\mathbb{E}\left[T_A\right] + B_2\beta^2\mu\sqrt{R} + \left(B_5\sqrt{\mu\beta R} + B_6\mu\beta\sqrt{R}\right)\cdot\beta \\
&\quad + I^{\text{busy}}\left(B_5\sqrt{\mu\beta R} + B_6\mu\beta\sqrt{R}, M\right) + B_7\frac{\beta R}{M} + \left[2M + 2\frac{R}{M}\right]\mathbb{E}\left[X - T_A\right] + \frac{1}{1-p_2}I^{\text{busy}}\left(M, M\right) \\
&\quad \leq B_1\sqrt{\mu\beta R}\cdot\mathbb{E}\left[T_A\right] + B_2\beta^2\mu\sqrt{R} + \left(B_5\sqrt{\mu\beta R}\right)\mathbb{E}\left[T_A\right] + B_6\mu\beta^2\sqrt{R} \\
&\quad + I^{\text{busy}}\left(B_5\sqrt{\mu\beta R} + B_6\mu\beta\sqrt{R}, M\right) + B_7\frac{R}{M}\mathbb{E}\left[T_A\right] + \left[2M + 2\frac{R}{M}\right]\mathbb{E}\left[X - T_A\right] + \frac{1}{1-p_2}I^{\text{busy}}\left(M, M\right) \\
&\quad \leq \max\left(2M + 2\frac{R}{M}, (B_1 + B_5)\sqrt{\mu\beta R} + B_7\frac{R}{M}\right)\mathbb{E}\left[X\right] + (B_2 + B_6)\beta^2\mu\sqrt{R} \\
&\quad + I^{\text{busy}}\left(B_5\sqrt{\mu\beta R} + B_6\mu\beta\sqrt{R}, M\right) + \frac{1}{1-p_2}I^{\text{busy}}\left(M, M\right) \\
&= \left(A_1\sqrt{\mu\beta R} + A_2\frac{R}{M}\right)\mathbb{E}\left[X\right] + A_3\beta^2\mu\sqrt{R} + I^{\text{busy}}\left(B_5\sqrt{\mu\beta R} + B_6\mu\beta\sqrt{R}, M\right) + A_4 I^{\text{busy}}\left(M, M\right),
\end{aligned}$$

where we have taken the constant $A_1 \triangleq \max\left(B_1 + B_5, C_3\right)$, the constant $A_2 \triangleq B_2 + B_3$, the constant $A_3 \triangleq B_2 + B_6$, and the constant $A_4 = \frac{1}{1-p_2}$. Upon dividing the reward integral by the cycle length, we obtain that

$$\begin{aligned}
\mathbb{E}\left[Q(\infty)\right] &= \frac{\mathbb{E}\left[\int_0^X [N(t)-R]\mathrm{d}t\right]}{\mathbb{E}\left[X\right]} \\
&\leq \frac{\left(A_1\sqrt{\mu\beta R} + A_2\frac{R}{M}\right)\mathbb{E}\left[X\right] + A_3\beta^2\mu\sqrt{R} + I^{\text{busy}}\left(B_5\sqrt{\mu\beta R} + B_6\mu\beta\sqrt{R}, M\right) + A_4 I^{\text{busy}}\left(M, M\right)}{\mathbb{E}\left[X\right]} \\
&= A_1\sqrt{\mu\beta R} + A_2\frac{R}{M} + \frac{A_3\beta^2\mu\sqrt{R} + I^{\text{busy}}\left(B_5\sqrt{\mu\beta R} + B_6\mu\beta\sqrt{R}, M\right) + A_4 I^{\text{busy}}\left(M, M\right)}{\mathbb{E}\left[X\right]} \\
&\leq A_1\sqrt{\mu\beta R} + A_2\frac{R}{M} + \frac{A_3\beta^2\mu\sqrt{R} + I^{\text{busy}}\left(B_5\sqrt{\mu\beta R} + B_6\mu\beta\sqrt{R}, M\right) + A_4 I^{\text{busy}}\left(M, M\right)}{\beta + T^{\text{busy}}\left(D_1\beta\mu\sqrt{R}, k - R\right)},
\end{aligned}$$

which is the upper bound stated in Theorem 6.1.

6.2.1 Proof of Lemma 6.1, Upper Bound on the Acc. Phase Integral

We prove this result via two applications of Lemma 4.1. To apply this decomposition lemma, there are two broad steps. First, we must specify a starting stopping time (T_0), an ending stopping time (P), a series of intervening stopping times (T_i), the process (Y_t), and an counting variable (F). Second, we must prove that the three preconditions of the lemma hold, given these specifications.

First Application of Lemma 4.1, Epoch-Level.

Definition of (τ_j). We define the sequence of stopping times $(\tau_j : j = 0, 1, \ldots, R)$ as

$$\tau_j \triangleq \min\left\{t > 0 : N(t) \leq R - j\right\},$$

i.e., τ_j is the first time there are only $R - j$ jobs within the system. Note that, by definition, $\tau_0 = 0$. We call the period $\left[\tau_j, \min\left(\tau_{j+1}, T_A\right)\right)$ the j*-th epoch*, and say epoch j occurs whenever $\tau_j < T_A$. We then let the random variable n_e denote the number of epochs which occur in a given renewal cycle.

Specification Step. Since we are interested in bounding $\mathbb{E}\left[\int_0^{T_A}[N(t) - R]\mathrm{d}t\right]$, we let our starting stopping time be $T_0 = 0$, our ending stopping time be $P = T_A$, our intervening stopping times be $T_j = \tau_j$, the process of interest $Y_t = N(t) - R$ and our counting variable be $F = n_e$. Let

$$p_{\text{rise}}^{(j)} \triangleq \Pr\left(\max_{t \in [\tau_j, \min(\tau_{j+1}, T_A)} N(t) \geq R + C_3\sqrt{\mu\beta R}\middle| n_e \geq j\right)$$

be the probability that the total number of jobs $N(t)$ exceeds $R + C_3\sqrt{\mu\beta R}$ during epoch j.

Required Claims. From here, we can apply Lemma 4.1 after showing the following claims:

Claim 6.1 (Accumulation Phase, Integral Bound)**.** *Let* $\tau_j \triangleq \min\left\{t \geq 0 : N(t) \leq R - j\right\}$*,* $T_A \triangleq \min\left\{t \geq 0 : Z(t) = R + 1\right\}$*, and let* $n_e \triangleq \max\left\{i \in \mathbb{Z}^+ : \tau_i < T_A\right\}$*. Then,*

$$\mathbb{E}\left[\int_{\tau_j}^{\min(\tau_{j+1}, T_A)} [N(t) - R]dt \middle| n_e \geq j\right] \leq B_1\sqrt{\mu\beta R}\cdot\mathbb{E}\left[\min\left(\tau_{j+1}, T_A\right) - \tau_j | n_e \geq j\right] + C_2\beta^2\mu j p_{rise}^{(j)},$$

where $B_1 = 3.6$ *and* $C_2 = \frac{1}{2 \cdot 0.98} < 0.511$*.*

Claim 6.2 (Accumulation Phase, Continuing Probability Bound)**.** *Recall that the total number of epochs* $n_e \triangleq \max\left\{j \in \mathbb{Z}^+ : \tau_j < T_A\right\}$*. Then,*

$$\Pr\left(n_e \geq j + 1 | n_e \geq j\right) \leq 1 - C_4 p_{rise}^{(j)},$$

where $C_4 = 0.98$*.*

Proof of Lemma 6.1 assuming Claims 6.1 and 6.2.

Before going further, we show how to complete the proof of Lemma 6.1, assuming the two prior claims. Applying Lemma 4.1, we find that

$$\begin{aligned}
\mathbb{E}\left[\int_0^{T_A}[N(t)-R]\mathrm{d}t\right] &\leq B_1\sqrt{\mu\beta R}\cdot\mathbb{E}\left[T_A\right]+C_2\beta^2\mu\sum_{j=1}^{R}jp_{\text{rise}}^{(j)}\prod_{i=1}^{j-1}\left(1-C_4p_{\text{rise}}^{(j)}\right)\\
&\leq B_1\sqrt{\mu\beta R}\cdot\mathbb{E}\left[T_A\right]+\frac{C_2}{C_4}\beta^2\mu\left[\sum_{j=1}^{R}jC_4p_{\text{rise}}^{(j)}\prod_{i=1}^{j-1}\left(1-C_4p_{\text{rise}}^{(j)}\right)\right]\\
&= B_1\sqrt{\mu\beta R}\cdot\mathbb{E}\left[T_A\right]+\frac{C_2}{C_4}\beta^2\mu\left[\sum_{j=1}^{R}\prod_{i=1}^{j}\left(1-C_4p_{\text{rise}}^{(j)}\right)\right],
\end{aligned}$$

where we have made use of the "expectation as a sum of tails" trick in the final line. From here, we make use of the following claim:

Claim 6.3. *Let* $p_{rise}^{(j)} \triangleq \Pr\left(\max_{t\in[\tau_j,\min(\tau_{j+1},T_A)}N(t)\geq R+C_3\sqrt{\mu\beta R}\middle|n_e\geq j\right)$ *be the probability that the total number of jobs* $N(t)$ *exceeds* $R+C_3\sqrt{\mu\beta R}$ *during epoch* j*. Then, for* $j\geq A_5\sqrt{R}$*,*

$$p_{rise}^{(j)}\geq 0.99\frac{A_5}{\sqrt{R}}.$$

We defer the proof of Claim 6.3 to Section A.3. Applying the claim's result, we find that

$$\begin{aligned}
\sum_{j=1}^{R}\prod_{i=1}^{j}\left(1-C_4p_{\text{rise}}^{(j)}\right) &\leq \sum_{j=1}^{R}\left(1-\frac{0.99C_4A_5}{\sqrt{R}}\right)^{[j-A_5\sqrt{R}]^+}\\
&\leq \sum_{j=1}^{\infty}\left(1-\frac{0.99C_4A_5}{\sqrt{R}}\right)^{[j-A_5\sqrt{R}]^+}\\
&= A_5\sqrt{R}+\sum_{j=0}^{\infty}\left(1-\frac{0.99C_4A_5}{\sqrt{R}}\right)^{j}\\
&= A_5\sqrt{R}+\frac{1}{0.99C_4A_5}\sqrt{R}
\end{aligned}$$

Returning to our original inequality, we obtain that

$$\mathbb{E}\left[\int_0^{T_A}[N(t)-R]\mathrm{d}t\right]\leq B_1\sqrt{\mu\beta R}\cdot\mathbb{E}\left[T_A\right]+\frac{C_2}{C_4}\left(A_5+\frac{1}{0.99C_4A_5}\right)\beta^2\mu\sqrt{R}.$$

Noting that $A_5=1$ and taking $B_2\triangleq 1.04>\frac{C_2}{C_4}\left(A_5+\frac{1}{0.99C_4A_5}\right)$, the proof is complete. □

And so, assuming the preconditions of Lemma 4.1 (Claim 6.1 and Claim 6.2) as well as a few helper claims, we have proven Lemma 6.1. Thus, it now suffices to prove those preconditions.

Proof of Claim 6.2, Upper Bound on Epoch-Continuation Probability.

Proving Claim 6.1 will be more involved; we deal with Claim 6.2 first. Note that

$$\begin{aligned}\Pr\left(n_e \geq j+1 | n_e \geq j\right) &= 1 - \Pr\left(n_e = j | n_e \geq j\right) \\ &= 1 - \Pr\left(T_A < \tau_{j+1} | n_e \geq j\right).\end{aligned}$$

Thus, it suffices to show that

$$\Pr\left(T_A < \tau_{j+1} | n_e \geq j\right) \geq C_4 p_{\text{rise}}^{(j)}.$$

To do this, we note a particular sequence of events which results in $T_A < \tau_{j+1}$. In particular, we define the up-crossing time $u = \min\left\{t > \tau_j : N(t) \geq R + C_3\sqrt{\mu\beta R}\right\}$ and the down-crossing time $d = \min\left\{t > u : N(t) \leq R\right\}$, then let $\mathcal{E}_1$ be the event that $\gamma < \tau_{j+1}$ and let $\mathcal{E}_2$ be the event that $T_A < d$. Note also that, since jobs depart one at a time, the down-crossing time $d < \tau_{j+1}$. Thus,

$$\Pr\left(T_A < \tau_{j+1} | n_e \geq j\right) \geq \Pr\left(\mathcal{E}_1 | n_e \geq j\right) \Pr\left(\mathcal{E}_2 | E_1\right) = p_{\text{rise}}^{(j)} \Pr\left(\mathcal{E}_2 | E_1\right).$$

To bound $\Pr\left(\mathcal{E}_2 | E_1\right)$, we couple our original number of jobs $N(t)$ to a coupled number of jobs $\tilde{N}(t)$. We construct $\tilde{N}(t)$ as

$$\tilde{N}(t) \triangleq N(u) + \Pi_A\left((u,t]\right) - \mathcal{D}\left[R\right]\left((u,t]\right),$$

where $t \in \left[u, \min\left(d, T_A\right)\right)$. Since $Z(t) \leq R$ for all $t < T_A$, we have that $\tilde{N}(t) \leq N(t)$ for all t in our interval. Let $\tilde{d} \triangleq \min\left\{t > u : \tilde{N}(t) \geq R\right\}$ be the analogous down-crossing time in the coupled system. Clearly, $\tilde{d} \leq d$. Thus,

$$\Pr\left(T_A < \tau_{j+1} | \mathcal{E}_1\right) \geq \Pr\left(T_A < d | \mathcal{E}_1\right) \geq \Pr\left(T_A < \tilde{d} \middle| \mathcal{E}_1\right).$$

Finally, note that, if $\tilde{d} > u + \beta$, then $N(t)$ must have stayed above R for an entire setup time, i.e. $\tilde{d} < T_A$. Thus, it suffices to show that

$$\Pr\left(\tilde{d} - u > \beta\right) \geq C_4.$$

We prove this with a straightforward application of Claim A.3. In particular, since we know the number of busy servers $Z(t) \leq R$ for any time before T_A, we can apply the second case and see that

$$\Pr\left(\tilde{d} - u > \beta\right) \geq 1 - 2\Phi\left(-\frac{C_3}{\sqrt{2}}\right) - \frac{2}{3\sqrt{\mu\beta R}} \geq 0.98$$

Taking this to be C_4, we have completed our proof. □

Proof of Claim 6.1, Bound on Epoch Integral

With the probability claim proven, we return to the proving Claim 6.1, the upper bound on the time integral over an epoch. As discussed previously, we do this via another application of Lemma 4.1; we specify the stopping times being used, then prove that the preconditions hold. Before the specification step, though, we must define a paired sequence of stopping times, which we call up-crossings and down-crossings.

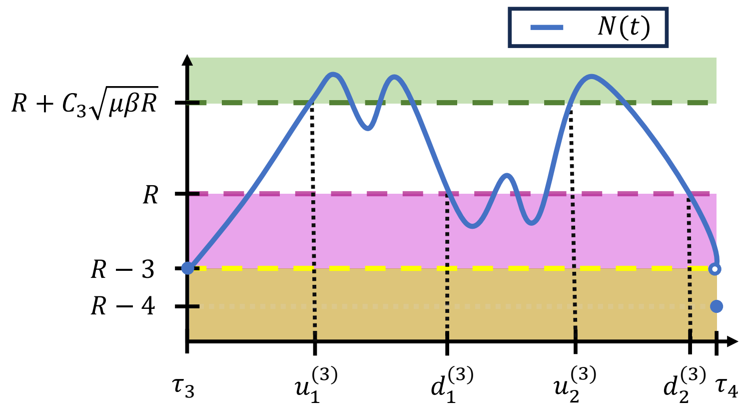

Figure 6.1: A depiction of the up-crossings and down-crossings defined in Section 6.2.1. In this example, we see that the number of up-crossings in epoch 3 is $n_e^{(3)} = 2$ and that, in this case, epoch 3 ends when epoch 4 begins (i.e. at time τ_4).

Definition of up-crossings and down-crossings. Let the 0-th down-crossing time in epoch j occur at time τ_j, i.e.

$$d_0^{(j)} \triangleq \tau_j.$$

Next, define the first up-crossing in epoch j as the first time during epoch j that the total number of jobs $N(t)$ exceeds $R + C_3\sqrt{\mu\beta R}$, i.e.

$$u_1^{(j)} \triangleq \min\left\{t > \tau_j : N(t) \geq R + C_3\sqrt{\mu\beta R}\right\}.$$

From here, define i-th down-crossing in epoch j and the $i+1$-th up-crossing in epoch j as

$$d_i^{(j)} \triangleq \min\left\{t \geq u_i^{(j)} : N(t) \leq R\right\}$$

and

$$u_{i+1}^{(j)} \triangleq \min\left\{t \geq d_i^{(j)} : N(t) \geq R + C_3\sqrt{\mu\beta R}\right\},$$

respectively; we visualize these definitions in Figure 6.1. We call the interval $\left[d_i^{(j)}, \min\left(u_i^{(j)}, \min\left(T_A, \tau_{j+1}\right)\right)\right)$ the i-th rise, and the interval $\left[u_i^{(j)}, \min\left(d_i^{(j)}, \min\left(T_A, \tau_{j+1}\right)\right)\right)$ the i-th fall. We say the i-th up-crossing occurs if $u_i^{(j)} < \min\left(T_A, \tau_{j+1}\right)$ and let

$$n_u \triangleq \max\left\{i \geq 0 : u_i^{(j)}\right\} < \min\left(\tau_{j+1}, T_A\right)$$

be the random number of up-crossings which occur in epoch j. Note that, if the i-th up-crossing occurs, then, by definition, $d_i < \tau_{j+1}$; this means that the i-th fall can always be written as $\left[u_i, \min\left(T_A, d_i\right)\right)$. For readability, we fix our epoch of interest and omit the superscript j on our up-crossings and down-crossings.

Specification Step. With up-crossings and down-crossings defined, we are now ready to specify our application of Lemma 4.1. We let our starting stopping time be $T_0 = \tau_j = d_0$, our ending stopping time be $P = \min(T_A, \tau_{j+1})$, our intervening sequence be $(u_i)_{i=1}^{\infty}$, our process of interest be $[N(t) - R]$, and our counting variable be $F = n_e$.

Required Claims. From here, in order to apply Lemma 4.1, we must show the following three claims:

Claim 6.4 (Within Epoch, Initial Integral Bound)**.** *One can bound the initial integral by*

$$\mathbb{E}\left[\int_{d_0}^{\min(u_1,\min(T_A,\tau_{j+1}))} [N(t)-R]dt \middle| n_e \geq j\right] \leq B_1\sqrt{\mu\beta R}\cdot\mathbb{E}\left[\min\left(u_1, \min\left(T_A, \tau_{j+1}\right)\right) - \tau_j | n_e \geq j\right],$$

where $B_1 = 3.6$*.*

Claim 6.5 (Within Epoch, Continuing Integral Bound)**.** *One can bound the intervening sequence integrals by*

$$\mathbb{E}\left[\int_{u_i}^{\min(u_{i+1},\min(T_A,\tau_{j+1}))} [N(t)-R]dt \middle| n_u \geq i\right] \leq B_1\sqrt{\mu\beta R}\cdot\mathbb{E}\left[\min\left(u_{i+1}, \min\left(T_A, \tau_{j+1}\right)\right) - u_i | n_u \geq i\right] + r_1\beta^2\mu j,$$

where $r_1 = \frac{1}{2}$ *is an absolute constant independent of system parameters.*

Claim 6.6 (Within Epoch, Continuing Probability Bound)**.** *Recall that* $p_{rise}^{(j)}$ *is the probability that the total number of jobs* $N(t)$ *ever exceeds* $R + C_3\sqrt{\mu\beta R}$ *during epoch* j*, given that epoch* j *occurs. Then,*

$$\Pr\left(n_u > 0\right) = p_{rise}^{(j)}$$

and, for all $i \geq 1$*,*

$$\Pr\left(n_u \geq i+1 | n_u \geq i\right) \leq 0.02 = (1 - p_2),$$

where p_2 *is an absolute constant independent of system parameters.*

Proof of Claim 6.1, assuming Claims 6.4, 6.5, and 6.6. Once again, before we move on to the proofs of these claims, we show that they indeed suffice to prove our goal, Claim 6.1. By Lemma 4.1,

$$\begin{aligned}\mathbb{E}\left[\int_{\tau_j}^{\min(T_A,\tau_{j+1})} [N(t)-R]\mathrm{d}t \middle| n_e \geq j\right] &\leq B_1\sqrt{\mu\beta R}\cdot\mathbb{E}\left[\min\left(T_A, \tau_{j+1}\right) - \tau_j | n_e \geq j\right] \\ &\quad + p_{\text{rise}}^{(j)} r_1\beta^2\mu j \sum_{i=1}^{\infty}\left(1-p_2\right)^{i-1} \\ &= B_1\sqrt{\mu\beta R}\cdot\mathbb{E}\left[\min\left(T_A, \tau_{j+1}\right) - \tau_j | n_e \geq j\right] + p_{\text{rise}}^{(j)}\frac{r_1}{p_2}\beta^2\mu j,\end{aligned}$$

setting $C_2 \triangleq \frac{r_1}{p_2}$, we obtain Claim 6.1. ☐

All that remains in our proof of Lemma 6.1 is to show the three aforementioned claims.

Proof of Claim 6.4: Within Epoch, Initial Integral Bound. Proving this claim is quite simple. In fact, we now prove a far more general claim, that

$$\int_{d_i}^{\min(u_i,\min(T_A,\tau_{j+1}))} [N(t)-R]\mathrm{d}t \le C_3\sqrt{\mu\beta R}\cdot\left[\min\left(T_A,\tau_{j+1}\right)-d_i\right]. \tag{6.1}$$

To see this, it suffices to note that, at any point between a down-crossing and up-crossing, the total number of jobs $N(t)$ must be strictly less than $R+C_3\sqrt{\mu\beta R}$. Apply this to the 0-th down-crossing and we have the claim. □

Proof of Claim 6.5: Within Epoch, Continuing Integral Bound. This proof is a bit more involved. We separate the interval $\Big[u_i, \min\left(u_{i+1}, \min\left(T_A,\tau_{j+1}\right)\right)\Big)$ into the i-th fall and the i-th rise, as discussed previously. For the rising portion, we can simply apply (6.1). For the falling portion, we apply the integral coupling claim, Claim A.2. In particular, note that $Z(t)\ge R-j$ until time τ_{j+1} and that the interval $[u_i,\min\left(d_i,T_A\right))$ is equivalent to the interval $[u_i,\min\left(d_i,u_i+Y_{R+1}(u_i)\right))$. Applying Claim A.2, we find that

$$\begin{aligned}\mathbb{E}\left[\int_{u_i}^{\min(d_i,T_A)}[N(t)-R]\mathrm{d}t\middle|S(u_i)\right] &\le \frac{1}{2}\beta^2\mu j+[N(u_i)-R]\cdot Y_{R+1}(u_i)\\ &= \frac{1}{2}\beta^2\mu j + C_3\sqrt{\mu\beta R}\cdot Y_{R+1}(u_i).\end{aligned}$$

From here, we note that $Y_{R+1}(u_i)\le\min\left(d_i,T_A\right)$ with probability at least

$$\min_{S(u_i)}\Pr\left(d_i<T_A|n_u\ge i,S(u_i)\right)\ge p_2.$$

Thus,

$$Y_{R+1}(u_i)\le\frac{1}{p_2}\mathbb{E}\left[\min\left(d_i,T_A\right)-u_i|S(u_i)\right].$$

After combining our bound on the i-th fall with our bound on the i-th rise, then taking $B_1=\frac{C_3}{p_2}$ and $r_1=\frac{1}{2}$, we have the claim. □

Proof of Claim 6.6: Within Epoch, Continuing Probability Bound. We now proceed to our final claim, a statement concerning the probability $\Pr\left(n_u>0\right)$ and the conditional continuation probability $\Pr\left(n_u\ge i+1|n_u\ge i\right)$.

To begin, we first note that, since the first up-crossing occurs once $N(t)$ exceeds $R+C_3\sqrt{\mu\beta R}$ during epoch j, one has

$$p_{\text{rise}}^{(j)}\triangleq\Pr\left(u_1<\min\left(T_A,\tau_{j+1}\right)\right)=\Pr\left(n_u>0\right).$$

Next, we note that there can only be another up-crossing after the i-th up-crossing in an epoch if the i-th down-crossing before time T_A, i.e.

$$\Pr\left(n_u\ge i+1|n_u\ge i\right)\le\Pr\left(d_i<T_A|n_u\ge i\right).$$

To bound this down-crossing probability, we condition on the filtration at time u_i and make a worst-case bound based on the state at that time. First, we show that the event $\{d_i < T_A\} = \{d_i < u_i + Y_{R+1}(u_i)\}$, given that the i-th up-crossing has occurred. To see this, note that at time u_i, if another down-crossing *doesn't* occur before the $(R+1)$-th server turns on, we know that $T_A = u_i + Y_{R+1}(u_i)$. Likewise, if another down-crossing *does* occur before the $(R+1)$-th server turns on, then

$$T_A > d_i + \beta > u_i + \beta > u_i + Y_{R+1}(u_i).$$

Thus, the event $\{d_i < T_A\} = \{d_i < u_i + Y_{R+1}(u_i)\}$, given that the i-th up-crossing has occurred.

To bound this probability, we appeal to Claim A.2. Since we are in the first part of our renewal cycle, we know that the $(R+1)$-th server has not yet turned on, i.e. that $Z(t) \leq R$. If we let $\gamma = d_i - u_i$, then

$$\gamma = \min\{t > 0 : N(t + u_i) \leq R\},$$

and thus

$$\Pr(d_i < u_i + Y_{u_i} | n_u \geq i) = \Pr(\gamma < Y_{u_i} | n_u \geq i).$$

Applying the second case of Claim A.2 and noting that $N(u_i) - R = C_3\sqrt{\mu\beta R}$, we obtain that

$$\Pr(\gamma < Y_{u_i} | n_u \geq i) \leq 2\Phi\left(-\frac{C_3}{\sqrt{2}}\right) + \frac{1}{100} < 0.02$$

Taking this final quantity to be $(1 - p_2)$, we have the claim. □

6.2.2 Proof of Lemma 6.2, the Upper Bound on Draining Phase Integral.

To prove this lemma, we again make use of Lemma 4.1. We proceed through the usual two-step process, first defining the stopping time sequence we will analyze over, then proving the three preconditions of the lemma.

Definition of the upcrossings $\upsilon_i^{(\mathbf{up})}$ and downcrossings $\upsilon_i^{(\mathbf{down})}$. Recall that the draining phase begins at time T_A. Let $M_B \triangleq \min\left(k - R, \max\left(\frac{\sqrt{R}}{D_1\sqrt{\beta}}, 1\right)\right)$ be a specially-set analysis threshold. Let the stopping time $\zeta_1^{(\mathrm{d})} \triangleq \min\{t \geq T_A : N(t) < R + M_B\}$ be the first time the number of jobs $N(t)$ drops below $R + M_B$, and recursively define

$$\zeta_i^{(\mathrm{u})} \triangleq \min\left\{t \geq \zeta_i^{(\mathrm{d})} : N(t) \geq R + M_B\right\}$$

and

$$\zeta_{i+1}^{(\mathrm{d})} \triangleq \min\left\{t \geq \zeta_i^{(\mathrm{u})} : N(t) < R + M_B\right\}.$$

Application of Lemma 4.1. Applying Lemma 4.1, we take our initial stopping time to be T_A, our final stopping time to be X, our intervening stopping times to be $\zeta_i^{(\mathrm{d})}$, our process of interest to be $[N(t) - R]$ and our counting index to be n_b.

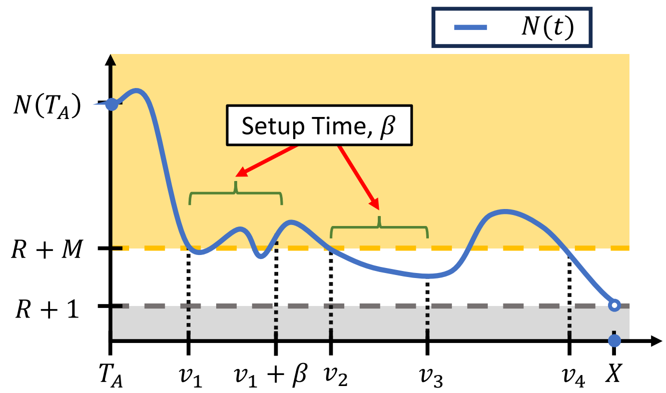

Figure 6.2: A depiction of the separated visits defined in Section 6.2.2. In this example, we find that the number of visits is $n_v = 4$. Notice that the second separated visit only happens when *both* a setup time β has passed *and* the number of jobs $N(t)$ is currently $\leq R + M$. For the third visit, these events happen simultaneously.

Required Claims. From here, all we need to show is the usual three claims: a bound on the initial integral

$$\mathbb{E}\left[\int_{T_A}^{\zeta_1^{(\mathrm{d})}} [N(t) - R]\,\mathrm{d}t \middle| \mathcal{F}_{T_A}\right] \leq [N(T_A) - R] \cdot \left[\frac{1}{\mu} + \beta + \frac{1}{2\mu M_B} + \frac{R}{\mu M_B^2}\right] + [N(T_A) - R]^2 \frac{1}{2\mu M_B} + \frac{R\beta}{M_B^2}, \tag{6.2}$$

a bound on the continuing integral

$$\mathbb{E}\left[\int_{\zeta_i^{(\mathrm{d})}}^{\zeta_{i+1}^{(\mathrm{d})} [N(t)-R]\mathrm{d}t} \middle| \mathcal{F}_{\zeta_i^{(\mathrm{d})}}\right] \leq \frac{1}{\mu} + \beta + \frac{1}{\mu M_B}\left[1 + b_1\sqrt{\beta R} + \frac{R}{M_B}\right], \tag{6.3}$$

and

$$\Pr\left(n_b \geq i + 1 | n_b \geq i\right) \leq \frac{1}{M_B}. \tag{6.4}$$

Proof of Lemma 6.2 assuming (6.2),(6.3),(6.4)**.** From these, we have

$$\mathbb{E}\left[\int_{T_A}^{\zeta_1^{(\mathrm{d})}} [N(t) - R]\,\mathrm{d}t\right] \leq \mathbb{E}\left[N(T_A) - R\right] \cdot \left[\frac{1}{\mu} + \beta + \frac{1}{2\mu M_B} + \frac{R}{\mu M_B^2}\right] + \mathbb{E}\left[(N(T_A) - R)^2\right] \frac{1}{2\mu M_B} + \frac{R\beta}{M_B^2},$$

and also that

$$\mathbb{E}\left[\int_{\zeta_1^{(\mathrm{d})}}^{X} [N(t) - R]\,\mathrm{d}t\right] \leq M_B^2 \frac{1}{\mu}\left[\frac{b_1\sqrt{\beta}}{\sqrt{R}} + \frac{6}{R} + \frac{b_2}{\sqrt{R}}\right] + \frac{M_B}{2\mu} + M_B\beta + \frac{1}{\mu}\left[\frac{R}{M_B} + \frac{3}{2}M_B + \frac{1}{2}\right] + D_1\sqrt{\mu R\beta}.$$

To finish off the proof, we use the following claim, whose proof we defer until Section A.6:

Claim 6.7 (Upper Bound on $\mathbb{E}[N(T_A)]$). *Recall that* $T_A \triangleq \min\{t > 0 : Z(t) = R + 1\}$. *Then,*

$$\mathbb{E}[N(T_A) - R] \le F_1 \mu\beta\sqrt{R}\left(1 + \frac{F_2}{\sqrt{\mu\beta}}\right)$$

and

$$\mathbb{E}\left[(N(T_A) - R)^2\right] \le F_1^2(\mu\beta)^2 R\left(1 + \frac{F_2}{\sqrt{\mu\beta}}\right)^2 + 2\mu\beta R,$$

where $F_1 = 2.12$ *and* $F_2 = 3.645$.

Continuing our proof of Lemma 6.2, we apply Claim 6.7 to find that

$$\begin{aligned}\mathbb{E}\left[\int_{T_A}^{\zeta_1^{(d)}} [N(t) - R]\,\mathrm{d}t\right] &\le \mathbb{E}[N(T_A) - R]\cdot\left[\frac{1}{\mu} + \beta + \frac{1}{2\mu M_B} + \frac{R}{\mu M_B^2}\right] \\ &\quad + \mathbb{E}\left[(N(T_A) - R)^2\right]\frac{1}{2\mu M_B} + \frac{R\beta}{M_B^2} \\ &\le F_1\mu\beta\sqrt{R}\left(1 + \frac{F_2}{\sqrt{\mu\beta}}\right)\left(1 + \frac{3}{2\mu\beta}\right)\end{aligned}$$

As such, to complete our proof it suffices to show (6.2),(6.3), and (6.4).

Precursor: The "Wait-Then-Busy" Period Idea. To prove these inequalities, we make use of the following useful claim.

Claim 6.8 (Wait-Busy Period Bound). *Let τ be some stopping time. Let the next down-crossing d_{gen}, defined as*

$$d_{\text{gen}} \triangleq \min\{t \ge 0 : N(t+\tau) \le R + h\},$$

be the length of time until the number of jobs $N(t) \le R + h$, where h is some positive integer which is smaller than $k - R$. Suppose that, at time τ, we know that $Z(t+\tau) \ge R$ for any $t \in [0, d_{\text{gen}}]$. Let $\tilde{d_{\text{gen}}}$ be the relative downcrossing time in a coupled system with exactly R busy servers. Then we have the following bound for the time integral of $N(t) - R$ from τ to $\tau + d_{\text{gen}}$:

$$\begin{aligned}\mathbb{E}\left[\int_{\tau}^{\tau+d_{\text{gen}}} [N(t) - R]dt\,\middle|\,\mathcal{F}_\tau\right] &\le [N(\tau) - (R+h)]\cdot\left[\frac{1}{\mu} + \beta + \frac{1}{2\mu h} + \frac{R}{\mu h^2}\right] \\ &\quad + \left([N(\tau) - (R+h)]^+\right)^2\frac{1}{2\mu h} \\ &\quad + \mathbb{E}\left[\min\left(\beta, \tilde{d_{\text{gen}}}\right)\right]\cdot\left[h + \frac{R}{h^2}\right]\end{aligned}$$

We defer the proof of Claim 6.8 until Section 6.2.2. For now, we give a brief intuition for how the bound is derived and how we use it in our proof. Essentially, we can consider performing the following procedure at time τ: First, watch the system for β time. If the number of jobs ever dips below $R + h$ during this watching period, we can end our integral immediately. If the number of jobs $N(t)$ never dips below $R+h$ during this watching period, then we know for sure that we have at least $R+h$ servers on at time $\tau+\beta$, since the number of busy server $Z(t)$ satisfies

$$Z(t)=\min\left(k,\min_{s\in[t-\beta,t]}N(s)\right).$$

Moreover, by the same justification, those servers will *stay* on until the number of jobs $N(t)$ dips below $R+h$; in other words, they will stay on until time d_{gen}. The proof of the claim follows along essentially the same lines, formalizing things and performing computations using coupling and martingales.

Proof of (6.2): Bound on Integral until First Visit.

To prove (6.2), we simply apply Claim 6.8 with starting time $\tau=T_A$ and threshold $h=M_B$. From there, we note $N(t)-(R+h)\le N(t)-R$ and that $\mathbb{E}\left[\min\left(\beta,\zeta_1^{(\text{d})}\right)\right]\le\beta$; this gives the claim directly.

Proof of (6.3): Bound on Integral Between Visits.

To prove (6.3), we break the integral into two parts; from the downcrossing $\zeta_i^{(\text{d})}$ to the upcrossing $\zeta_i^{(\text{u})}$ and from the upcrossing $\zeta_i^{(\text{u})}$ to the downcrossing $\zeta_{i+1}^{(\text{d})}$.

First, we note that

$$\int_{\zeta_i^{(\text{d})}}^{\min\left(\zeta_i^{(\text{u})},X\right)}[N(t)-R]\,\mathrm{d}t\le\left[\min\left(\zeta_i^{(\text{u})},X\right)-\zeta_i^{(\text{d})}\right]\cdot M_B, \tag{6.5}$$

since $\zeta_i^{(\text{u})}$ is the next time $N(t)\ge R+M_B$. To bound $\mathbb{E}\left[\min\left(\zeta_i^{(\text{u})},X\right)-\zeta_i^{(\text{d})}\middle|\mathcal{F}_{\zeta_i^{(\text{d})}}\right]$, we couple the system to an $M/M/\infty$ at time $\zeta_i^{(\text{d})}$, and note that the coupled up-crossing time

$$\tilde{T}_{(R+M_B-1)\to(R+M_B)}+\zeta_i^{(\text{d})}\ge\zeta_i^{(\text{u})}\ge\min\left(\zeta_i^{(\text{u})},X\right).$$

From standard results on the $M/M/\infty$ (reproduced in Section 5.3.7 for completeness), we have that

From $\zeta_i^{(\text{u})}$ onwards, we use the "wait-busy" bound. Applying Claim 6.8 with $h=M_B$, we obtain

$$\begin{aligned}\mathbb{E}\left[\int_{\zeta_i^{(\text{u})}}^{\zeta_{i+1}^{(\text{d})}}[N(t)-R]\,\mathrm{d}t\middle|\mathcal{F}_{\zeta_i^{(\text{u})}}\right]&\le\left[\frac{1}{\mu}+\beta+\frac{1}{2\mu M_B}+\frac{R}{\mu M_B^2}\right]+\frac{1}{2\mu M_B}\\&\quad+\mathbb{E}\left[\min\left(\beta,\zeta_{i+1}^{(\text{d})}-\zeta_i^{(\text{u})}\right)\middle|\mathcal{F}_{\zeta_i^{(\text{u})}}\right]\cdot\left[M_B+\frac{R}{M_B}\right].\end{aligned}$$

To bound the conditional expectation of $\mathbb{E}\left[\min\left(\beta,\zeta_{i+1}^{(\text{d})}-\zeta_i^{(\text{u})}\right)\middle|\mathcal{F}_{\zeta_i^{(\text{u})}}\right]$, it suffices to bound $\mathbb{E}\left[\min\left(\tilde{d_{\text{gen}}},\beta\right)\right]$, for the coupled relative down-crossing time $\tilde{d_{\text{gen}}}$. But note that $\tilde{d_{\text{gen}}}$ is exactly the busy period length of a critically-loaded M/M/1 queue with arrival rate and departure rate $k\lambda$. From standard results (reproduced in Claim A.9 for completeness), we have

$\mathbb{E}\left[\min\left(\tilde{d_{\text{gen}}}, \beta\right)\right] \leq \frac{b_1\sqrt{\beta}}{\sqrt{R}} + \frac{6}{R}$ and thus that

$$\mathbb{E}\left[\int_{\zeta_i^{(u)}}^{\zeta_{i+1}^{(d)}} [N(t) - R]\, \mathrm{d}t \middle| \mathcal{F}_{\zeta_i^{(u)}}\right] \leq \frac{1}{\mu} + \beta + \frac{1}{\mu M_B}\left[1 + b_1\sqrt{\beta R} + \frac{R}{M_B}\right] + M_B \cdot \left[\frac{b_1\sqrt{\beta}}{\sqrt{R}} + \frac{6}{R}\right].$$

Proof of (6.4): Bound on the probability of another visit.

To see (6.4), we first note that, if there is another upcrossing, then there must be another downcrossing. As such, it suffices to upper bound $\Pr\left(\zeta_i^{(u)} < X \middle| \zeta_i^{(d)}\right)$. To do this, we note that the number of busy servers $Z(t) \geq R$. From Claim A.1, it thus suffices to bound the corresponding probability in the coupled system with exactly R busy servers. But this is simply the probability that, in a simple random walk started at $W(0) = M_B - 1$, the walk process $W(t)$ hits $W(t) = M_B$ before it hits $W(t) = 0$. Classically, this probability is $\frac{1}{M_B}$. □

6.2.3 Proof of Lemma 6.3: Lower Bound on the Cycle Length.

Preliminaries. The proof of this lemma is much simpler than the others. Before describing our strategy, we first state some preliminaries. Recall the definition of the start of the j-th epoch, τ_j:

$$\tau_j \triangleq \min\left\{t \geq 0 : N(t) \leq R - j\right\}.$$

We call the period $\left[\tau_j, \min\left(\tau_{j+1}, T_A\right)\right)$ the j-th epoch, and say epoch j occurs if $\tau_j < T_A$. Say an epoch j is long if it lasts longer than a setup time β; note that such an epoch must exist, since servers can only turn on during long epochs, and a server must turn on before the accumulation phase is over. Let L be the index of the *first* long epoch, i.e.

$$L \triangleq \min\left\{j \in \{0, 1, 2, \ldots, R\} : \min\left(\tau_{j+1}, T_A\right) - \tau_j > \beta\right\}.$$

Note that, although the random time τ_L is *not* a stopping time (we do not know how long an epoch will last when the epoch starts), the first moment we can identify epoch L, the random time $\tau_L + \beta$, *is* a stopping time. Moreover, we know that $\tau_L + \beta < T_A$. From here, one sees that

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[\tau_L + \beta] + \mathbb{E}[X - (\tau_L + \beta)] \\ &\geq \beta + \mathbb{E}[X - (\tau_L + \beta)].\end{aligned}$$

To complete the proof, it suffices to show via a simple coupling argument that

$$\mathbb{E}[X - (\tau_L + \beta)] \geq T^{\text{busy}}\left(\mathbb{E}\left[N\left(\tau_L + \beta\right) - R\right], k - R\right). \tag{6.6}$$

From there, we can apply (5.6), the bound on $\mathbb{E}[N(\tau_L + \beta)]$ in terms $\mathbb{E}[L]$, and (5.3), the bound on $\mathbb{E}[L]$, to obtain Lemma 6.3.

Proof of (6.6)**: Lower Bound on the Cycle Length.** To show (6.6), we make a simple coupling argument. Consider a coupled system with the same arrival process but the maximum possible departure rate, $k\mu$. In other words, define for all time $t \geq \tau_L + \beta$

$$\tilde{N}(t) \triangleq N(\tau_L + \beta) + A((\tau_L + \beta, t]) - \mathcal{D}[k]((\tau_L + \beta, t]).$$

Since $Z(t) \leq k$, it follows that $N(t) \leq \tilde{N}(t)$ for all $t \geq \tau_L + \beta$. Naturally then, the coupled down-crossing time $\tilde{X} \triangleq \min\left\{t \geq \tau_L + \beta : \tilde{N}(t) \leq R\right\}$ must be smaller than the end of our renewal cycle, since the cycle ends once a similar down-crossing occurs *after* the $(R+1)$-th server has turned on, i.e.

$$X \triangleq \min\left\{t > 0 : Z(t^-) = R + 1, Z(t) = R\right\} = \min\left\{t > T_A : N(t) \leq R\right\}.$$

Thus we have

$$\begin{aligned}
\mathbb{E}\left[X - (\tau_L + \beta) | \mathcal{F}_{\tau_L + \beta}\right] &\geq \mathbb{E}\left[\tilde{X} - (\tau_L + \beta) \middle| \mathcal{F}_{\tau_L + \beta}\right] \\
&= T^{\text{busy}}\left([N(\tau_L + \beta) - R]^+, k - R\right) \\
&\geq T^{\text{busy}}\left(N(\tau_L + \beta) - R, k - R\right).
\end{aligned}$$

Since the length of a busy period scales linearly with the number of jobs which start it, we have (6.6) via linearity of expectation. □

6.3 The Upper Bound: Review of Findings

In this chapter, we proved an upper bound on the average waiting time in the M/M/k/Setup-Deterministic. We proved this bound via a number of applications of the MIST Lemma, Lemma 4.1. In fact, to bound the accumulation phase integral, we needed to use the MIST Lemma in a nested way: First, we used it to break the accumulation phase into epochs, and then we used it to break each epoch into "rises" and "falls," periods of time punctuated by up-crossings and down-crossings. Compared to the lower bound of Chapter 5, the upper bound proven here truly highlights the utility of the MIST method.

Chapter 7

The Approximation

In Chapter 7, we present our approximation for the average waiting time in the M/M/k/Setup-Deterministic. We begin by discussing why we need such an approximation, then state the approximation, then give a short justification for its form.

7.1 Why we need an approximation

Despite our success in analyzing the M/M/k/Setup-Deterministic, our upper and lower bounds *alone* are not suitable for practical use in predicting the value of the average waiting time $\mathbb{E}[T_Q]$.

There are two reasons for this. First, our bounds are off by constant factors. Although we can prove that the true value of the average waiting time lies within our two bounds, it's not *a priori* obvious whether the true value of $\mathbb{E}[T_Q]$ will get closer to one bound or the other as we vary the system parameters. Although the true value does not *seem* to ever get closer to a particular bound (and so we could conceivably just scale our lower bound to serve as a predictor), it would be better to have a more concrete theoretical justification for our prediction.

The second reason why our bounds are unsuitable for practical use is their complexity. Although both the upper and lower bounds are far more straightforward to compute than, for example, the average waiting time in the M/M/k/Setup-Exponential, both bounds incorporate a large number of terms and are thus somewhat difficult to reason about on the fly. As such, it would be better to have a predictor which incorporates only a few, easy-to-remember terms.

7.2 The approximation

To this end, we introduce the following approximation; the justification for the approximation follows. An empirical evaluation of this approximation is left to Section 8.2; it is extremely accurate.

Approximation 1 (Approximation to the average queue length.)**.** *In the M/M/k/Setup-Deterministic,* *for offered loads $R \triangleq k\rho > 2$,*

$$\mathbb{E}\left[Q(\infty)\right] \approx Q_{\text{apx}} \triangleq \frac{\frac{1}{2}\beta^2 C_{\text{apx}}\sqrt{R} + \frac{\beta C_{\text{apx}}\sqrt{R}}{\mu k(1-\rho)}\left[\frac{\beta C_{\text{apx}}\sqrt{R}+1}{2} + \frac{1}{1-\rho}\right]}{\beta + \frac{\beta C_{\text{apx}}\sqrt{R}}{\mu k(1-\rho)}}, \tag{7.1}$$

where $C_{\text{apx}} \triangleq \sqrt{\frac{\pi}{2}}$.

7.3 Justification

We arrive at this bound via a straightforward combination of our results from Chapters 5 and 6, along with a few modifications. We follow our renewal-reward analysis, separately approximating the expected time integral over our renewal cycle and the expected length of that renewal cycle, the numerator and denominator of 7.1, respectively.

7.3.1 Justification of Numerator

We first approximate the numerator of our expression, the expected time integral over our chosen renewal cycle. We begin by recalling the lower bound on the time integral, Lemma 5.1, which states

$$\mathbb{E}\left[\int_0^X Q(t)\mathrm{d}t\right] \geq L_1\beta^2\sqrt{R} + I^{\text{busy}}\left(\left[L_1\beta\sqrt{R} - (k-R)\right]^+, k-R\right),$$

where

$$I^{\text{busy}}(x,z) \triangleq \frac{x}{\mu z}\left[\frac{x+1}{2} + \frac{1}{1-\frac{k\lambda}{k\lambda+\mu z}}\right]$$

represents the time integral of the queue length a certain M/M/1 queue over a busy period started by x jobs.

To obtain the appropriate constant C_{apx}, we next note that, although our theorem states L_1 as an absolute constant, as the setup time β and the offered load R grow, the best possible constant will become $C_{\text{apx}} = \sqrt{\frac{\pi}{2}}$. Under the hood, this convergence stems from the fact that

$$\sum_{j=1}^{R}\prod_{i=1}^{j}\left(1 - \frac{j}{R}\right) \approx \int_0^\infty e^{\frac{-j^2}{2R}}\mathrm{d}j = \frac{1}{2}\sqrt{2\pi R};$$

see the proof of Lemma 5.1 for more details.

To complete the bound, it suffices to remove the subtraction of $(k-R)$ in the busy period term, which we anticipate is an artifact of our analysis. Removing it, we obtain the desired approximation

$$\mathbb{E}\left[\int_0^X Q(t)\mathrm{d}t\right] \approx \frac{1}{2}\beta^2 C_{\text{apx}}\sqrt{R} + \frac{\beta C_{\text{apx}}\sqrt{R}}{\mu k(1-\rho)}\left[\frac{\beta C_{\text{apx}}\sqrt{R}+1}{2} + \frac{1}{1-\rho}\right]. \tag{7.2}$$

7.3.2 Justification of Denominator

We next approximate the denominator of our expression, the expected length of our chosen renewal cycle. To do so, we again make use of the lower bound on the expected cycle length $\mathbb{E}[X]$ from Lemma 6.3, which states

$$\mathbb{E}[X] \geq \beta + \frac{L_1 \beta \sqrt{R}}{\mu k(1-\rho)}.$$

By making the same convergence argument for L_1, i.e. that $L_1 \to C_{\text{apx}}$ for large setup times β and large offered loads R, we obtain the denominator, completing both parts of our bound.

Chapter 8

Evaluation

In this chapter, we review and discuss the practical takeaways of this thesis. In the previous chapters, we derived upper and lower bounds on the average waiting time in the M/M/k/Setup-Deterministic, and, from these upper and lower bounds, we constructed a new approximation for the average waiting time. Here, we highlight the utility of our results by investigating three practical questions concerning the M/M/k/Setup:

1. Does setup distribution matter?
2. How does our approximation's accuracy change as we vary our system parameters?
3. How do our results affect provisioning?

8.1 Does Setup Distribution Matter?

The choice of setup distribution in the study of the M/M/k/Setup makes a tremendous difference in its queueing behavior. In particular, when setup times are Deterministic, modeling them as Exponential, or, worse yet, not modeling them at all, can lead to a dramatic underestimation of the delay caused by setup time. We illustrate this point in Figure 8.1, where we vary the number of servers k in the system, while holding fixed the load ρ and the setup time β.

We make two observations. First, we observe that there is an almost unfathomable difference in the average waiting time behavior for a no-setup system as compared to either an Exponential or a Deterministic setup system. Given that this is a log-log plot, we see that the average waiting time appears to decay polynomially in the number of servers k, while we know from Erlang that the average waiting time in a no-setup M/M/k queue decays exponentially as k grows.

Second, we observe that the Deterministic setup system seems to have a substantially different scaling with the number of servers k when compared to the Exponential setup system. Returning to our log-log plot Figure 8.1, while the average waiting time in the M/M/k/Setup-Deterministic seems to decay exactly as $k^{-\frac{1}{2}}$, the waiting time behavior of the M/M/k/Setup-Exponential does not seem to decay with any obvious trend. Moreover, the average waiting time in the Exponential model decays much more quickly than in the Deterministic model.

These experimental observations provide an experimental confirmation of exactly what we asserted in the beginning of this thesis: that, when modeling multiserver systems, **setup time, as well as its distribution, cannot be ignored.**

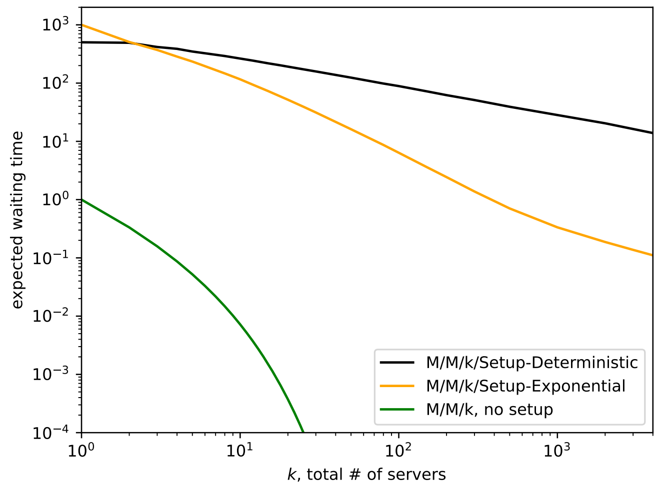

Figure 8.1: Simulation results for the M/M/k/Setup-Deterministic, M/M/k/Setup-Exponential, M/M/k (no setup), with $\mu = 1$, setup time $\beta = 1000$, and load kept at a constant $\rho = 0.5$. Note the high separation between the Exponential setup and Deterministic setup models at large scales, as well as the large separation between the models with and without setup.

8.2 How Accurate Is Our Approximation?

We have found our approximation (1) to be extremely accurate across a wide range of system parameters, so long as the offered load $R > 2$. In our analysis, we assume that the offered load $R \gg \sqrt{R}$ and often make considerable use of this fact. In that sense, it is a testament to the strength of our approach that our resulting approximation remains accurate all the way up to an offered load of $R = 2$. Offered loads smaller than this are of limited interest in practical settings. That said, when the offered load is that small, we anticipate that the multiserver system exhibits a "single-server" bottleneck effect; preliminary investigations seem to confirm this.

8.3 How Do Our Results Change Provisioning?

A common, but sometimes complex problem which arises in many systems is that of designing the system such that the average waiting time of a customer is below some target waiting time. Historically, we have studied this provisioning problem a great deal [5, 7, 16, 21, 26], and understand the problem well for systems without setup times, e.g. there's a straightforward formula for the average waiting time in the M/M/k without setup.

Unfortunately, our understanding of this problem is still quite poor for many modern systems, since their average waiting times are affected by setup times. In particular, many modern systems dynamically control the number of servers that they keep on, periodically turning servers off in order to save energy. As mentioned previously, previous results on understanding the relationship between setup times and the average waiting time leave much to be desired. Our new results expand on the state-of-the-art Exponential model in two important ways: 1) obtaining the pre-

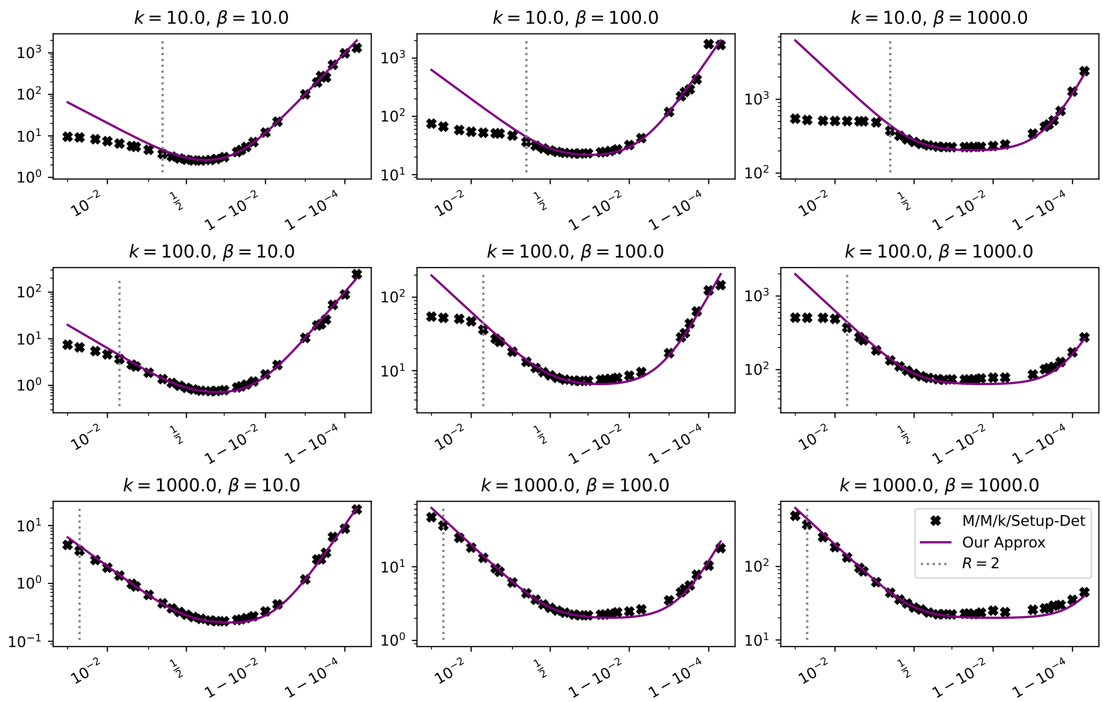

Figure 8.2: An example highlighting the excellent accuracy of our simple approximation (1) to the average waiting time in the M/M/k/Setup-Deterministic. For each of these 9 plots, we plot the behavior of the average waiting time as one varies the load ρ from 0 to 1, holding fixed the total number of servers k as well as the setup time β. In each row, we hold the number of servers k constant while testing increasing values of the setup times β. In each column, we hold the setup time β constant while increasing the number of servers. We also include, as a reference, a dotted line illustrating the point at which the offered load $R \triangleq k\rho = 2$.

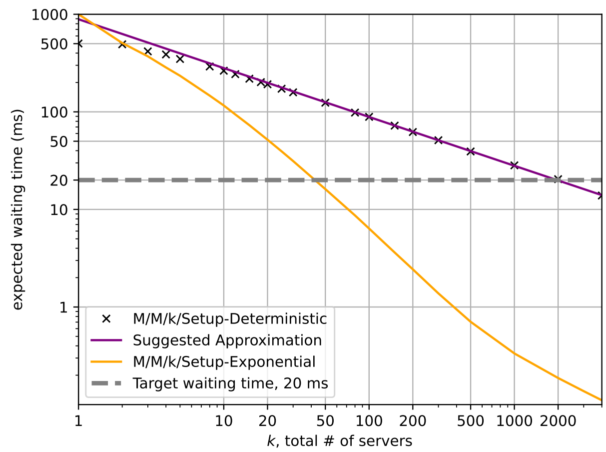

Figure 8.3: An example highlighting the differences between M/M/k/Setup-Deterministic and M/M/k/Setup-Exponential, varying the number of servers k while keeping the mean service time $\frac{1}{\mu} = 1$ ms, the mean setup time $\beta = 1000$ ms, and the load $\rho = 0.5$ fixed. Note the ability of our approximation to accurately predict the behavior of the M/M/k/Setup-Deterministic. To achieve a target waiting time of 20 ms, our approximation accurately predicts it will take $k \approx 2000$, while the Exponential model only predicts that a k of around 50 is needed.

dicted average waiting time is much easier computationally, and 2) the quality of the prediction is much better.

Easier predictions. Compared to the Exponential model, our new Deterministic approximation greatly simplifies the design process. In particular, when using the Exponential model, one must solve a system of $O(k^2)$ quadratic equations to find the average waiting time $\mathbb{E}[T_Q]$. Two practical issues arise from this fact. First, the equations change depending on the number of servers k, meaning that the computation must be repeated every time one wishes to test a new number of servers. Second, the opacity of the process makes it difficult to draw intuition about how the average waiting time is changed when one alters the system parameters. In contrast, our approximation is a relatively simple function of the relevant parameters. The simplicity of our approximation has, likewise, two benefits:1) computing the average waiting time becomes easy, and 2) the form of our approximation makes it clear how and why the system's behavior changes in response to certain parameters.

Higher quality predictions. Moreover, when compared to the predictions of the Exponential model, the predictions we obtain using our Deterministic approximation are of a much higher quality. This difference in quality is perhaps best illustrated by looking at a simple example. In Figure 8.3, we compare the prediction from the Exponential model to the prediction from our approximation, plotting how the predicted average waiting time changes as one increases the number of servers k while keeping fixed the load $\rho = 0.5$, the average setup time $\beta = 1000$ ms, and the average service time $\frac{1}{\mu} = 1$ ms. Our goal is to determine how large the number of servers

k needs to be before we reach our target waiting time $T_{\text{target}} = 20$ ms. In both models, the average waiting time decreases as the system grows larger and larger. However, the Exponential model predicts that average waiting time will be small enough once $k = 50$. On the other hand, as captured by our approximation, the Deterministic setup system will only reach the target waiting time once the number of servers $k \approx 2000$, a full 40 times larger than what the Exponential system predicts! For even a modestly-large number of servers, the Exponential system predicts waiting times which are orders of magnitude smaller than what really occur.

Chapter 9

Conclusion

In this chapter, we summarize the thesis, discuss some broader impacts of this thesis, and state some related open problems.

9.1 Summary and Takeaways

In this thesis, we studied the effect of setup times on the queueing behavior of multiserver systems. In particular, we studied how the average waiting time $\mathbb{E}[T_Q]$ in the M/M/k/Setup depends on the system parameters like the number of servers k, the average setup time β, and the load ρ. In Chapter 1, we first noted that the fundamental difficulty in analyzing setup in multiserver systems was the fact that multiple servers can be *in setup* at the same time. We then that all prior theoretical work made that simplifying assumption that setup times were distributed i.i.d. Exponential, even though, practically-speaking, setup times are much closer to *Deterministic*; see Chapter 2 for more details. Furthermore, we found in simulation that this distributional assumption has a large impact on the behavior of the system: systems with Deterministic setup times have very different behavior from systems with Exponential setup times.

Accordingly, we narrowed our focus to studying the average waiting time in the M/M/k/Setup-Deterministic (defined in Chapter 3), deriving the first-ever lower and upper bounds on this quantity in Chapters 5 and 6, respectively. Next, in Chapter 7, we described how to take the tightest parts of our bounds and combine them to make an approximation which is extremely accurate. Finally, in Chapter 8, we reviewed and discussed the practical takeaways of our work:

- that the *average waiting time in the M/M/k/Setup-Exponential is drastically smaller* than the average waiting time in the corresponding M/M/k/Setup-Deterministic (Section 8.1);
- that *our approximation is highly accurate* in predicting the average waiting time in the M/M/k/Setup-Deterministic (Section 8.2);
- and that the simplicity and accuracy of *our approximation radically simplifies capacity provisioning* for dynamically-scaled systems (Section 8.3).

9.2 Broader Impacts

This thesis has the potential to impact a large number of different fields, since setup times arise in so many different settings.

9.2.1 Computer Science

In computer science, setup times arise most directly when performing dynamic-scaling in the cloud. There, booting up another container (or virtual machine) might take a few seconds while the actual runtime of a specific task might only take a few milliseconds. Because we do not fully understand how to manage systems with setup times, these servers could be burning 5 percent more energy than necessary. *A priori* that might not sound like much, but consider the following: energy waste does not solely affect these companies' profitability, but also has affects our climate as well. Given that one percent of *all power globally* is spent running these datacenters, if we can save two or three percent more energy in their operations, that would be a significant gain for the entire world.

9.2.2 Operations/Management

From an operations/management perspective, the effect of setup times is well-illustrated in employee turnover. When hiring, it might take months to fully onboard a new team member, whereas a typical task might be completed in a day; on the other hand, many employees can be laid off more-or-less instantly. The way in which a firm goes about hiring people, migrating them between different teams, and deciding to lay them off is a great example of the human side of dynamic scaling. Effective management is timelessly relevant, and a setup-time-oriented perspective could provide insights and tools in the same vein as the Pollaczek–Khinchine formula or the Erlang-C model.

9.2.3 Healthcare

Setup times also occur in the medical setting, e.g. when managing on-call doctors. Because patient need (i.e. service demand) is unpredictable, some doctors are often kept "on-call" for up to 36 hours at a time. While on-call, although a physician may not always have work to do, if their service is requested, then they are expected to respond within, say, 30 minutes (which includes travel time to the hospital, if required). For context, most requests can be handled in a very short amount of time, e.g. under a minute. Because these physicians must stay ready-to-respond for multiple days, the current on-call system can lead to extreme sleep deprivation and, accordingly, a poor standard of care for patients. Along the lines of this thesis, further research on dynamically allocating physicians might someday lead us to a new, more sustainable on-call system, with both better care quality and better physician well-being.

9.3 Open Problems

9.3.1 Standby States

Within this thesis, we assume that servers have two persistent states: *on* and *off*; for some systems, this assumption turns out to be wrong. In reality, many servers possess intermediate *standby* states. A server on *standby* takes a shorter amount of time to get ready than a server that is completely *off*, but it also burns more energy. Since the setup process itself takes energy, by using these *standby* states cleverly, we might be able to both improve performance and improve energy efficiency within these systems. Given this, one might ask: **"When should we put a server on *standby* versus turning it completely *off*? What are the benefits of using the standby state?"**

9.3.2 Analyzing Tail Performance in the M/M/k/Setup.

Another important open problem lies in analyzing tail performance in the M/M/k/Setup. For context, when customers purchase cloud hosting, an ubiquitous component in their purchase agreements is some kind of "tail/deadline constraint" on their job delay. For example, the agreement will stipulate that "95% of submitted jobs must complete service within one second of their arrival," with some sort of financial penalty if this constraint is not honored.

Tail constraints in queueing pose a number of technical challenges. In even the single-server case, we do not yet understand how to schedule jobs to optimally meet these constraints. In the multiserver case, though, we have another perspective from which we can analyze the problem: that of *dynamic-scaling*. Instead of thinking about how to schedule these tail-constrained jobs, we can instead think about how we should dynamically-scale our system so that we *guarantee* that our tail constraints are met. This scaling perspective provides a natural way of thinking about the different costs involved. With enough servers, we should be able to ensure that our tail-constraint is met. As such, we can now ask: **"How and when should a system use additional servers to satisfy a given tail constraint?"**

The above question is challenging, and worth considering even in systems without setup times. However, as we have made clear throughout this thesis, setup times often have an enormous impact on the queueing behavior of a dynamically-scaled system. Although there exists extensive study of the performance of dynamic staffing [6, 26], especially in the time-varying arrival rate case [7, 21], much of that work has yet to be extended to the setup time case. As such, we should also ask a more fundamental question: **"How does setup time impact the *distribution* of waiting time in the M/M/k/Setup?"**

Appendix A

Miscellaneous Claims

In this Appendix, we prove some miscellaneous claims.

A.1 Coupling Claims

A.1.1 Three Coupling Claims

We now describe three useful claims applied throughout the proof. The first, we will state and prove immediately. The latter two, we prove later, in Section A.1.2.

Claim A.1. *Suppose that we have two processes N_1 and N_2 with an initial relation*

$$N_1(a) \le N_2(a),$$

where the behavior of each process is governed, for all times s from a up to some stopping time τ, by the equation

$$N_j(s) \triangleq N_j(a) + \Pi_A((a,s]) - \mathcal{D}[Z_j(x)]((a,s]), \textit{ for } j \in \{1,2\}.$$

Furthermore, suppose that the function $Z_1(x)$ dominates the function $Z_2(x)$ on the interval $[a,\tau]$, i.e.

$$Z_1(s) \ge Z_2(s), \quad \forall s \in [a,\tau].$$

Then, for all $s \in [a,\tau]$,

$$N_1(s) \le N_2(s).$$

Proof. Proof. It suffices to show that $N_2(s) - N_1(s) \ge 0$, for all $s \in [a,\tau]$. Applying the definitions of N_1 and N_2, we find

$$\begin{aligned} N_2(s) - N_1(s) &= N_2(a) - N_1(a) + \left[\mathcal{D}[Z_1(x)]((a,s]) - \mathcal{D}[Z_2(x)]((a,s])\right] \\ &\ge \left[\mathcal{D}[Z_1(x)]((a,s]) - \mathcal{D}[Z_2(x)]((a,s])\right] \\ &= \sum_{i=1}^{k} \int_a^s \mathbf{1}\{Z_1(x) \ge i\}\,\mathrm{d}\Pi_i(x) - \sum_{i=1}^{k} \int_a^s \mathbf{1}\{Z_2(x) \ge i\}\,\mathrm{d}\Pi_i(x) \\ &= \sum_{i=1}^{k} \int_a^s \Big[\mathbf{1}\{Z_1(x) \ge i\} - \mathbf{1}\{Z_2(x) \ge i\}\Big]\mathrm{d}\Pi_i(x). \end{aligned}$$

Since $Z_1(x) \geq Z_2(x)$ for all $x \in [a, s]$, the integrand $\left[\mathbf{1}\left\{Z_1(x) \geq i\right\} - \mathbf{1}\left\{Z_2(x) \geq i\right\}\right]$ must be non-negative; the claim follows. □ □

This claim leads nicely into a couple more claims. The first claim, Claim A.2, uses a coupling argument to bound the expected integral of $N(t)$ from some arbitrary time τ until $N(t)$ drops below some pre-defined threshold h, provided that one has a lower bound on the number of busy servers $Z(t)$ over that period. The second claim, Claim A.3, uses a related argument to bound the probability that $N(t)$ drops below some threshold h within some amount of time ℓ, given that one has bounds on $Z(t)$ over the relevant period.

Claim A.2 (Coupling Bound: Time Integral)**.** *Let τ be some stopping time. Let the next down-crossing time d_{gen}, defined as*

$$d_{\text{gen}} \triangleq \min\left\{t \geq 0 : N(t+\tau) \leq h\right\},$$

be the length of time that passes until the number of jobs $N(t)$ is less than some fixed threshold h. Suppose that, at time τ, we have a lower bound on the number of busy servers over a period, i.e.

$$Z(t) \geq R - j,$$

for all $t \in [\tau, \tau + \min(\ell, d_{\text{gen}})]$ and for some non-negative j. Then we have the following bound on the integral over this time period:

$$\mathbb{E}\left[\int_{\tau}^{\tau+\min(d_{\text{gen}},\ell)} [N(t) - h]\, dt \,\middle|\, \mathcal{F}_\tau\right] \leq \ell \cdot [N(\tau) - h]^+ + \frac{1}{2}\mu j \ell^2.$$

Claim A.3 (Coupling Bound: Probability)**.** *Let τ be some stopping time. Let the next down-crossing time d_{gen}, defined as*

$$d_{\text{gen}} \triangleq \min\left\{t \geq 0 : N(t+\tau) \leq h\right\},$$

be the length of time until the number of jobs $N(t)$ is less than some fixed threshold h. We consider two cases.

In the first case, suppose that we have a lower bound on the number of busy servers $Z(t)$ over some length ℓ interval starting at time τ, i.e.

$$Z(t) \geq R - j,$$

for all $t \in [\tau, \tau + \min(\ell, d_{\text{gen}})]$ and for some non-negative j. Then, we can bound the threshold-crossing probability by

$$\Pr\left(d_{\text{gen}} < \ell | \mathcal{F}_\tau\right) \geq 2\Phi\left(-\left[\frac{N(\tau) - h + \mu j \ell}{\sqrt{\ell(2k\lambda - \mu j)}}\right]\right) - \frac{2}{3\sqrt{\ell(2k\lambda - \mu j)}}.$$

In particular, if $N(\tau) - h = c\sqrt{\mu\beta R}$, then

$$\Pr\left(d_{\text{gen}} < \ell | \mathcal{F}_\tau\right) \geq 2\Phi\left(-\frac{c_1}{\sqrt{2}}\right) - \frac{1}{100}.$$

In the second case, suppose that we instead have an upper bound on $Z(t) \leq R$ during this interval instead. Then,

$$\Pr\left(d_{\text{gen}} < \ell | \mathcal{F}_\tau\right) \leq 2\Phi\left(-\left[\frac{N(\tau)-h}{\sqrt{2\ell k\lambda}}\right]\right) - \frac{2}{3\sqrt{2k\lambda\ell}}.$$

As before, if $N(\tau) - h = c\sqrt{\mu\beta R}$, then

$$\Pr\left(d_{\text{gen}} < \ell | \mathcal{F}_\tau\right) \leq 2\Phi\left(-\frac{c}{\sqrt{2}}\right) + \frac{1}{100}.$$

A.1.2 Proof of Claim A.2: Coupled Time Integral Bound

Claim A.2 (Coupling Bound: Time Integral)**.** *Let τ be some stopping time. Let the next down-crossing time d_{gen}, defined as*

$$d_{\text{gen}} \triangleq \min\left\{t \geq 0 : N(t+\tau) \leq h\right\},$$

be the length of time that passes until the number of jobs $N(t)$ is less than some fixed threshold h. Suppose that, at time τ, we have a lower bound on the number of busy servers over a period, i.e.

$$Z(t) \geq R - j,$$

for all $t \in [\tau, \tau + \min(\ell, d_{\text{gen}})]$ and for some non-negative j. Then we have the following bound on the integral over this time period:

$$\mathbb{E}\left[\int_\tau^{\tau+\min(d_{\text{gen}},\ell)} [N(t)-h]\, dt \middle| \mathcal{F}_\tau\right] \leq \ell \cdot [N(\tau)-h]^+ + \frac{1}{2}\mu j \ell^2.$$

Proof. We prove this claim in three parts. First, we construct a coupled process $\tilde{N}(t) \geq N(t)$ on the interval of interest. Then, we give an upper bound on $\mathbb{E}\left[\int_\tau^{\tau+\min(d_{\text{gen}},\ell)} \tilde{N}(t)\mathrm{d}t \middle| \mathcal{F}_\tau\right]$. Define $\tilde{N}(t)$ as

$$\tilde{N}(t) \triangleq N(\tau) + A(\tau,t) - \mathcal{D}\left[R-j\right]((\tau,t)).$$

Then, by Claim A.1, we have that

$$\tilde{N}(t) \geq N(t).$$

on the interval of interest. To develop the integral, we first move the minimum from the bounds of integration into the integrand. In particular, we note that the quantity $N(d_{\text{gen}}) - h = 0$, and thus, for any $t > \tau + d_{\text{gen}}$, the quantity $N(\min(\tau + d_{\text{gen}}, t)) - h = 0$. On the other hand, for any

$t < \tau + d_{\text{gen}}$, the quantity $N\left(\min\left(\tau + d_{\text{gen}}, t\right)\right) = N\left(t\right)$. It follows that

$$\begin{aligned}
\int_{\tau}^{\tau+\min(d_{\text{gen}},\ell)} \left[N(t) - h\right] \mathrm{d}t &= \int_{\tau}^{\tau+\min(d_{\text{gen}},\ell)} \left[N(\min\left(t, \tau + d_{\text{gen}}\right)) - h\right] \mathrm{d}t \\
&= \int_{\tau}^{\tau+\min(d_{\text{gen}},\ell)} \left[N(\min\left(t, \tau + d_{\text{gen}}\right)) - h\right] \mathrm{d}t \\
&\quad + \int_{\tau+\min(d_{\text{gen}},\ell)}^{\tau+\ell} \left[N(\min\left(t, \tau + d_{\text{gen}}\right)) - h\right] \mathrm{d}t \\
&= \int_{\tau}^{\tau+\ell} \left[N(\min\left(t, \tau + d_{\text{gen}}\right)) - h\right] \mathrm{d}t \\
&\leq \int_{\tau}^{\tau+\ell} \left[\tilde{N}(\min\left(t, \tau + d_{\text{gen}}\right)) - h\right] \mathrm{d}t.
\end{aligned}$$

Defining $\tilde{d_{\text{gen}}} \triangleq \min\left\{t > 0 : \tilde{N}(\tau + t) \leq h\right\}$, since $\tilde{N}(t) \geq N(t)$, we know both that $\tilde{d_{\text{gen}}} \geq d_{\text{gen}}$ and that, for any $t \in [\tau + d_{\text{gen}}, \tau + \tilde{d_{\text{gen}}}]$,

$$\tilde{N}(t) - h \geq 0.$$

Moreover, the process $V(t)$ defined as

$$V(t) \triangleq \tilde{N}(t) - \mu j t$$

is a martingale. Thus, we have

$$\int_{\tau}^{\tau+\ell} \left[\tilde{N}(\min\left(t, \tau + d_{\text{gen}}\right)) - h\right] \mathrm{d}t \leq \int_{\tau}^{\tau+\ell} \left[\tilde{N}(\min\left(t, \tau + \tilde{d_{\text{gen}}}\right)) - h\right] \mathrm{d}t.$$

Taking the expectation, we find that

$$\begin{aligned}
\mathbb{E}\left[\int_{\tau}^{\tau+\ell} \left[\tilde{N}(\min\left(t, \tau + \tilde{d_{\text{gen}}}\right)) - h\right] \mathrm{d}t \Big| \mathcal{F}_{\tau}\right] &= \int_{\tau}^{\tau+\ell} \mathbb{E}\left[\tilde{N}(\min\left(t, \tau + \tilde{d_{\text{gen}}}\right)) - h \Big| \mathcal{F}_{\tau}\right] \mathrm{d}t \\
&= \int_{\tau}^{\tau+\ell} \mathbb{E}\left[V(\min\left(t, \tau + \tilde{d_{\text{gen}}}\right)) + \mu j \left(\min\left(\tau + \tilde{d_{\text{gen}}}, t\right)\right) - h \Big| \right. \\
&= \int_{\tau}^{\tau+\ell} \mathbb{E}\left[V(\tau) + \mu j \left(\min\left(\tau + \tilde{d_{\text{gen}}}, t\right)\right) - h \Big| \mathcal{F}_{\tau}\right] \mathrm{d}t \qquad \text{(A.1)} \\
&\leq \int_{\tau}^{\tau+\ell} \mathbb{E}\left[V(\tau) + \mu j t - h | \mathcal{F}_{\tau}\right] \mathrm{d}t \\
&= \int_{\tau}^{\tau+\ell} \mathbb{E}\left[\tilde{N}(\tau) - \mu j \tau + \mu j t - h \Big| \mathcal{F}_{\tau}\right] \mathrm{d}t \\
&= \left[\tilde{N}(\tau) - h\right] \ell + \frac{1}{2} \mu j \ell^{2},
\end{aligned}$$

where (A.1) is an application of Doob's Optimal Stopping Theorem. □

A.1.3 Proof of Claim A.3: Coupled Probability Bound

We now prove Claim A.3, restated here for the reader's convenience.

Claim A.3 (Coupling Bound: Probability)**.** *Let τ be some stopping time. Let the next down-crossing time d_{gen}, defined as*

$$d_{\text{gen}} \triangleq \min\left\{t \geq 0 : N(t+\tau) \leq h\right\},$$

be the length of time until the number of jobs $N(t)$ is less than some fixed threshold h. We consider two cases.

In the first case, suppose that we have a lower bound on the number of busy servers $Z(t)$ over some length ℓ interval starting at time τ, i.e.

$$Z(t) \geq R - j,$$

for all $t \in [\tau, \tau + \min(\ell, d_{\text{gen}})]$ and for some non-negative j. Then, we can bound the threshold-crossing probability by

$$\Pr\left(d_{\text{gen}} < \ell | \mathcal{F}_\tau\right) \geq 2\Phi\left(-\left[\frac{N(\tau) - h + \mu j \ell}{\sqrt{\ell(2k\lambda - \mu j)}}\right]\right) - \frac{2}{3\sqrt{\ell(2k\lambda - \mu j)}}.$$

In particular, if $N(\tau) - h = c\sqrt{\mu\beta R}$, then

$$\Pr\left(d_{\text{gen}} < \ell | \mathcal{F}_\tau\right) \geq 2\Phi\left(-\frac{c_1}{\sqrt{2}}\right) - \frac{1}{100}.$$

In the second case, suppose that we instead have an upper bound on $Z(t) \leq R$ during this interval instead. Then,

$$\Pr\left(d_{\text{gen}} < \ell | \mathcal{F}_\tau\right) \leq 2\Phi\left(-\left[\frac{N(\tau) - h}{\sqrt{2\ell k\lambda}}\right]\right) - \frac{2}{3\sqrt{2k\lambda\ell}}.$$

As before, if $N(\tau) - h = c\sqrt{\mu\beta R}$, then

$$\Pr\left(d_{\text{gen}} < \ell | \mathcal{F}_\tau\right) \leq 2\Phi\left(-\frac{c}{\sqrt{2}}\right) + \frac{1}{100}.$$

Proof. We prove this result in three parts. First, we use Claim A.1 to construct a process $\tilde{N}(t) \geq N(t)$ on the interval of interest. Afterwards, we analyze the down-crossing probability of this coupled process. In particular, we use a reflection argument to show that

$$\Pr\left(d_{\text{gen}} < \ell\right) \geq 2\Pr\left(\tilde{N}(\tau + \ell) \leq h\right),$$

then use a Berry-Esseen bound to bound this final probability. In what follows, we focus on the lower-bound; the upper bound follows in precisely the same way.

To construct our coupled process, we note that, by assumption, the number of busy servers $Z(t) \geq R - j$ for any $t \in [\tau, \tau + \min(\ell, d_{\text{gen}})]$. Thus, by Claim A.1, the process $\tilde{N}(t)$ defined as

$$\tilde{N}(t) \triangleq N(\tau) + A(\tau, \tau + t) + \mathcal{D}[R - j]([\tau, \tau + t])$$

is an upper bound for $N(t + \tau)$, i.e.

$$\tilde{N}(t) \geq N(\tau + t)$$

for any $t \in [0, \min(\ell, d_{\text{gen}})]$. By definition, we have that

$$\begin{aligned}
\Pr(d_{\text{gen}} < \ell) &= \Pr\left(\inf_{t \in [0,\ell)} N(\tau + t) \leq h\right) \\
&\geq \Pr\left(\inf_{t \in [0,\ell)} \tilde{N}(t) \leq h\right).
\end{aligned}$$

From a reflection argument, since $\tilde{N}$ is upwards-biased,

$$\begin{aligned}
\Pr\left(\inf_{t \in [0,\ell)} \tilde{N}(t) \leq h\right) &= \Pr\left(\inf_{t \in [0,\ell)} \tilde{N}(t) \leq h, \tilde{N}(\ell) < h\right) + \Pr\left(\inf_{t \in [0,\ell)} \tilde{N}(t) \leq h, \tilde{N}(\ell) \geq h\right) \\
&\geq 2\Pr\left(\inf_{t \in [0,\ell)} \tilde{N}(t) \leq h, \tilde{N}(\ell) < h\right) \\
&= 2\Pr\left(\tilde{N}(\ell) < h\right).
\end{aligned}$$

Let $\sigma \triangleq \sqrt{\ell(2k\lambda - \mu j)}$. Now, assume that, for any x,

$$\left|\Pr\left(\tilde{N}(\ell) < \tilde{N}(0) + \mu j \ell + x\sigma\right) - \Phi(x)\right| \leq \frac{0.3328}{\sigma}, \tag{A.2}$$

we have

$$\begin{aligned}
\Pr\left(\tilde{N}(\ell) < h\right) &= \Pr\left(\tilde{N}(\ell) < \tilde{N}(0) + \mu j \ell + \frac{h - \mu j \ell - \tilde{N}(0)}{\sigma} \cdot \sigma\right) \\
&\geq \Phi\left(\frac{h - \mu j \ell - \tilde{N}(0)}{\sigma}\right) - \frac{1}{3\sigma} \\
&= \Phi\left(-\frac{[N(\tau) - h + \mu j \ell]}{\sigma}\right) - \frac{1}{3\sigma}.
\end{aligned}$$

Putting this all together, we find

$$\Pr(d_{\text{gen}} < \ell | \mathcal{F}_\tau) \geq 2\Phi\left(-\frac{[N(\tau) - h + \mu j \ell]}{\sigma}\right) - \frac{2}{3\sigma}.$$

From here, then, it suffices to show (A.2). To begin, note that, if we choose some arbitrarily large n and define

$$X_i \triangleq \Pi_i'\left(\frac{k\lambda\ell}{n}\right) - \Pi_i''\left(\frac{\mu(R - j)\ell}{n}\right) - \frac{\mu j \ell}{n},$$

where each $\Pi(y)$ is an independent Poisson random variable with mean y, then

$$\tilde{N}(\ell) =_d \sum_{i=1}^{n} X_i + \mu j\ell + \tilde{N}(0).$$

To compute the moments of X_i, note that one can define centered Poisson random variables $A_i = \Pi\left(\frac{k\lambda\ell}{n}\right) - \frac{k\lambda\ell}{n}$ and $B_i = \Pi\left(\frac{\mu(R-j)\ell}{n}\right) - \frac{\mu(R-j)\ell}{n}$, and then take $X_i = A_i - B_i$. Doing this, one finds that

$$\mathbb{E}\left[X_i^2\right] = \mathbb{E}\left[(A_i - B_i)^2\right] = \frac{k\lambda\ell}{n} + \frac{\mu(R-j)\ell}{n} = \frac{\mu(2R-j)\ell}{n}$$

and, using the triangle inequality, that

$$\mathbb{E}\left[|X_i|^3\right] = \mathbb{E}\left[|A_i - B_i|^3\right] \le \mathbb{E}\left[|A_i|^3\right] + \mathbb{E}\left[\left|B_i^3\right|\right] = \frac{\mu(2R-j)\ell}{n} + o\left(\frac{1}{n^2}\right).$$

We now apply the main result of [29]. Let $\sigma_n \triangleq \sqrt{\mathbb{E}\left[X_i^2\right]} = \sqrt{\frac{\mu(2R-j)\ell}{n}} = \frac{\sigma}{\sqrt{n}}$ and note that $\rho_n = \mathbb{E}\left[|X_i|^3\right] < \sigma_n + o\left(\frac{1}{n^2}\right)$ (from [4]). Then, noting that $\rho_n \ge 1.286\sigma_n^3$ for sufficiently large n, we have

$$\max_x \left|\Pr\left(\frac{\sum X_i}{\sqrt{n}\sigma_n} < x\right) - \Phi(x)\right| \le \frac{0.3328\rho_n + 0.429\sigma_n^3}{\sigma_n^3\sqrt{n}} = \frac{0.3328}{\sqrt{\mu(2R-j)\ell}} + o\left(\frac{1}{n}\right).$$

Now noting that

$$\frac{\sum_{i=1}^n X_i}{\sqrt{n}\sigma} = \frac{\tilde{N}(\ell) - \tilde{N}(0) - \mu j\ell}{\sigma}$$

and taking $n \to \infty$, we have our result. □

A.1.4 Proof of Claim A.4: Coupled Expectation Bound

Claim A.4 (Coupling Bound: Expected Value)**.** *Let τ be some stopping time. Let the next down-crossing time d_{gen}, defined as*

$$d_{\text{gen}} \triangleq \min\left\{t \ge 0 : N(t+\tau) \le h\right\},$$

be the length of time that passes until the number of jobs $N(t)$ is less than some fixed threshold h. Suppose that, at time τ, we have a lower bound on the number of busy servers over a period, i.e.

$$Z(t) \ge R - j,$$

for all $t \in [\tau, \tau + \min(\ell, d_{\text{gen}})]$ and for some $j \in [R]$. Then,

$$\mathbb{E}\left[[N(\tau+\ell) - h]\,\mathbf{1}_{d_{\text{gen}}>\ell} \middle| \mathcal{F}_\tau\right] \le [N(\tau) - h] + \mu j\ell. \tag{A.3}$$

and

$$\mathbb{E}\left[[N(\tau+\ell) - h]\,\mathbf{1}_{d_{\text{gen}}>\ell} \middle| \mathcal{F}_\tau\right] \le [N(\tau) - h + \mu j\ell]^2 + 2\mu R\ell. \tag{A.4}$$

Proof.

The proof is essentially an application of Doob's Optional Stopping Theorem to an appropriately selected martingale. To begin, we define a coupled process $\tilde{N}(t)$ with

$$\tilde{N}(t-\tau) \triangleq N(\tau) + A[\tau,t] - \mathcal{D}\left[R-j\right]\left([\tau,t]\right);$$

by Claim A.1, we know that $\tilde{N}(t-\tau) \geq N(t)$ for any $t \in [\tau, \tau + \min\left(d_{\text{gen}}, \ell\right)]$, and that the coupled hitting time $\tilde{d_{\text{gen}}} \triangleq \min\left\{t > 0 : \tilde{N}(t) \leq h\right\}$ can not be smaller than the original hitting time d_{gen}. It follows that

$$N\left(\tau+\ell\right)\mathbf{1}_{d_{\text{gen}}>\ell} \leq \tilde{N}\left(\ell\right)\mathbf{1}_{\tilde{d_{\text{gen}}}>\ell}.$$

Thus, it suffices to bound coupled versions of (A.3) and (A.4).

Construction of martingales. We now construct our martingales and set up the language of optional stopping. Note that, for any process $\tilde{N}(t)$ with independent, stationary increments, both functions V_1 and V_2, defined as

$$V_1(t) \triangleq \left[\tilde{N}(t) - h\right] - \mathbb{E}\left[\tilde{N}(t) - \tilde{N}(0)\right]$$

and

$$\begin{aligned} V_2(t) &\triangleq \left[\tilde{N}(t) - h - \mathbb{E}\left[\tilde{N}(t) - \tilde{N}(0)\right]\right]^2 - \mathbb{E}\left[\left[\tilde{N}(t) - h - \mathbb{E}\left[\tilde{N}(t) - \tilde{N}(0)\right]\right]^2\right] \\ &= \left(\tilde{N}(t) - h - \mu j t\right)^2 - \mu\left(2R-j\right)t \end{aligned}$$

are martingales [23]. Moreover, one has that

$$\begin{aligned} \left[\tilde{N}(\ell) - h\right]\mathbf{1}_{\tilde{d_{\text{gen}}}>\ell} &= \left[\tilde{N}\left(\min\left(\tilde{d_{\text{gen}}}, \ell\right)\right) - h\right]\mathbf{1}_{\tilde{d_{\text{gen}}}>\ell} \\ &= \left[\tilde{N}\left(\min\left(\tilde{d_{\text{gen}}}, \ell\right)\right) - h\right]\mathbf{1}_{\tilde{d_{\text{gen}}}>\ell} + \left[\tilde{N}\left(\min\left(\tilde{d_{\text{gen}}}, \ell\right)\right) - h\right]\mathbf{1}_{\ell \leq \tilde{d_{\text{gen}}}} \\ &= \left[\tilde{N}\left(\min\left(\tilde{d_{\text{gen}}}, \ell\right)\right) - h\right]. \end{aligned}$$

Proof of (A.3). Combining these facts allows us to prove our desired result. Applying Doob's Optional Stopping Theorem along with our previous deductions, we obtain

$$\begin{aligned} \mathbb{E}\left[\left[N\left(\tau+\ell\right) - h\right]\mathbf{1}_{d_{\text{gen}}>\ell}\middle|\mathcal{F}_\tau\right] &\leq \mathbb{E}\left[\left[\tilde{N}\left(\ell\right) - h\right]\mathbf{1}_{\tilde{d_{\text{gen}}}>\ell}\right] \\ &= \mathbb{E}\left[\tilde{N}\left(\min\left(\tilde{d_{\text{gen}}}, \ell\right)\right) - h\right] \\ &= \mathbb{E}\left[V_1\left(\min\left(\tilde{d_{\text{gen}}}, \ell\right)\right)\right] + \mu j \mathbb{E}\left[\min\left(\tilde{d_{\text{gen}}}, \ell\right)\right] \\ &= \mathbb{E}\left[V_1\left(0\right)\right] + \mu j \mathbb{E}\left[\min\left(\tilde{d_{\text{gen}}}, \ell\right)\right] \\ &= \left[\tilde{N}(0) - h\right] + \mu j \mathbb{E}\left[\min\left(\tilde{d_{\text{gen}}}, \ell\right)\right] \\ &\leq \left[\tilde{N}(0) - h\right] + \mu j \ell \\ &= \left[N(\tau) - h\right] + \mu j \ell. \end{aligned}$$

Proof of (A.4)**.** To do the same for the squared martingale $V_2(t)$, we must first note, via some algebra, that

$$\left(\tilde{N}(t) - h\right)^2 = V_2(t) + \left(\tilde{N}(t) - h\right)\mu j t - \mu j^2 t^2 + \mu\left(2R - j\right)t.$$

Now, applying the same deductions we made previously,

$$\begin{aligned}
&\mathbb{E}\left[\left[N\left(\tau+\ell\right) - h\right]\mathbf{1}_{d_{\text{gen}}>\ell}\middle|\mathcal{F}_\tau\right] \\
&\leq \mathbb{E}\left[\left[\tilde{N}\left(\ell\right) - h\right]^2\mathbf{1}_{\tilde{d_{\text{gen}}}>\ell}\right] \\
&= \mathbb{E}\left[\left(\tilde{N}\left(\min\left(\tilde{d_{\text{gen}}},\ell\right)\right) - h\right)^2\right] \\
&= \mathbb{E}\left[V_2\left(\min\left(\tilde{d_{\text{gen}}},\ell\right)\right)\right] + \mathbb{E}\left[\left(\tilde{N}(\min\left(\tilde{d_{\text{gen}}},\ell\right)) - h\right)\mu j \min\left(\tilde{d_{\text{gen}}},\ell\right)\right] - \mu j^2\left(\min\left(\tilde{d_{\text{gen}}},\ell\right)\right)^2 \\
&\quad + \mu\left(2R-j\right)\mathbb{E}\left[\min\left(\tilde{d_{\text{gen}}},\ell\right)\right] \\
&\leq \mathbb{E}\left[V_2\left(\min\left(\tilde{d_{\text{gen}}},\ell\right)\right)\right] + \mathbb{E}\left[\left(\tilde{N}(\min\left(\tilde{d_{\text{gen}}},\ell\right)) - h\right)\right]\mu j\ell + \mu\left(2R\right)\ell \\
&\leq \mathbb{E}\left[V_2\left(0\right)\right] + \left[\tilde{N}(0) - h + \mu j\ell\right]\mu j\ell + \mu\left(2R\right)\ell \\
&= \left[\tilde{N}(0) - h\right]^2 + \left[\tilde{N}(0) - h\right]\mu j\ell + \left(\mu j\ell\right)^2 + \mu\left(2R\right)\ell \\
&= \left[\tilde{N}(0) - h + \mu j\ell\right]^2 - \left[\tilde{N}(0) - h\right]\mu j\ell + \mu 2R\ell \\
&\leq \left[\tilde{N}(0) - h + \mu j\ell\right]^2 + 2\mu R\ell \\
&= \left[N(\tau) - h + \mu j\ell\right]^2 + 2\mu R\ell.
\end{aligned}$$

□

A.2 Proof of Claim A.5: Bound on the integral over a busy period.

Claim A.5 (Busy Period Integral Bound)**.** *Suppose that, at time τ, we can guarantee that $N(\tau) \geq Z(\tau) \geq R + j$. Let $\eta_i \triangleq \min\{t > 0 : N(t) \leq R + i\}$, for $i \in \{j, j+1, \ldots, [N(\tau) - R]\}$. Then,*

$$\mathbb{E}\left[\int_\tau^{\eta_j}[N(t) - R]dt\middle|\mathcal{F}_\tau\right] \leq \left(N(\tau) - (R+j)\right)\left[\frac{3}{2\mu j} + \frac{1}{\mu} + \frac{R}{\mu j^2}\right] + \frac{\left(N(\tau) - (R+j)\right)^2}{2\mu j} \triangleq I^{\text{busy}}\left([N(\tau) - R],\right.$$

Proof. We prove this claim via an appeal to conventional M/M/1 busy period analysis. In particular, we first note that

$$\int_\tau^{\eta_j}\left[N(t) - R\right]\mathrm{d}t = \sum_{i=j+1}^{N(\tau)-R}\int_{\eta_i}^{\eta_{i-1}}\left[N(t) - R\right]\mathrm{d}t,$$

meaning we need only bound the integrals between the η_i's. To bound that process, we define a coupled process $\tilde{N}(t)$ and bound the integrals over that process.

To do so, note that, until time η_j, the number of busy servers $Z(t) \geq R + j$. By Claim A.1, we can define, for each index i, the i-th coupled process $\tilde{N}_i(t)$ as

$$\tilde{N}_i(t) = N(\eta_{i+1}) + A(\eta_{i+1}, t) - \mathcal{D}\left[R + j\right]\left([\eta_{i+1}, t)\right),$$

and have $\tilde{N}(t) \geq N(t)$ on the interval $[\eta_{i+1}, \eta_i]$. Furthermore, we can extend our integral of interest from the interval $[\eta_{i+1}, \eta_i)$ to the interval $[\eta_{i+1}, \tilde{\eta}_i)$, where $\tilde{\eta}_i \triangleq \min\left\{t > 0 : N(t) \leq R + i\right\}$. Now, we note that

$$\mathbb{E}\left[\int_{\eta_{i+1}}^{\tilde{\eta}_i} \left[\tilde{N}_i(t) - R\right] \mathrm{d}t \middle| \mathcal{F}_\tau\right] = \mathbb{E}\left[\int_{\eta_{i+1}}^{\tilde{\eta}_i} \left[\tilde{N}_i(t) - (R + i)\right] \mathrm{d}t \middle| \mathcal{F}_\tau\right] + i\mathbb{E}\left[\eta_{i+1} - \tilde{\eta}_i | \mathcal{F}_\tau\right].$$

The first term on the right is simply the expected time integral of the number of jobs in an M/M/1 queue over a busy period, with arrival rate $k\lambda$ and departure rate $\mu(R + j)$. The second term is simply the quantity i multipled by the expected length of that M/M/1 busy period. Let $\rho_j = \frac{k\lambda}{\mu(R+j)}$. Then, from standard results on the M/M/1 busy period,

$$\mathbb{E}\left[\int_{\eta_{i+1}}^{\tilde{\eta}_i} \left[\tilde{N}_i(t) - (R + i)\right] \mathrm{d}t \middle| \mathcal{F}_\tau\right] = \frac{1}{\mu j}\left[\frac{1}{1 - \rho_j}\right] = \frac{1}{\mu j}\left[\frac{R}{j} + 1\right] = \frac{1}{\mu j} + \frac{R}{\mu j^2}.$$

Summing over all values of i, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\int_{\tau}^{\eta_j} \left[N(t) - R\right] \mathrm{d}t \middle| \mathcal{F}_\tau\right] \\
&\leq \sum_{i=j+1}^{N(\tau)-R} \mathbb{E}\left[\int_{\eta_i}^{\tilde{\eta}_{i-1}} \left[\tilde{N}_i(t) - R\right] \mathrm{d}t \middle| \mathcal{F}_\tau\right] \\
&= \sum_{i=j+1}^{N(\tau)-R} \left[\frac{1}{\mu j} + \frac{R}{\mu j^2}\right] + i\frac{1}{\mu j} \\
&= \left(N(\tau) - (R + j)\right)\left[\frac{1}{\mu j} + \frac{R}{\mu j^2}\right] + \left(N(\tau) - (R + j)\right)\frac{1}{\mu} \\
&\quad + \frac{1}{\mu j}\left[\frac{\left(N(\tau) - (R + j)\right)\left(N(\tau) - (R + j) + 1\right)}{2}\right] \\
&= \left(N(\tau) - (R + j)\right)\left[\frac{3}{2\mu j} + \frac{1}{\mu} + \frac{R}{\mu j^2}\right] + \frac{\left(N(\tau) - (R + j)\right)^2}{2\mu j},
\end{aligned}$$

as desired. □ □

A.3 Proof of Claim 6.3: Bound on the probability of an upcrossing.

We now prove Claim 6.3, restated here.

Claim 6.3. *Let* $p_{rise}^{(j)} \triangleq \Pr\left(\max_{t\in[\tau_j,\min(\tau_{j+1},T_A)} N(t) \geq R + C_3\sqrt{\mu\beta R}\middle| n_e \geq j\right)$ *be the probability that the total number of jobs* $N(t)$ *exceeds* $R + C_3\sqrt{\mu\beta R}$ *during epoch* j*. Then, for* $j \geq A_5\sqrt{R}$*,*

$$p_{rise}^{(j)} \geq 0.99\frac{A_5}{\sqrt{R}}.$$

We show a more general claim: that, for $j \geq A_5\sqrt{R}$,

$$p_{\text{rise}}^{(j)} \geq 0.99\frac{j}{R}. \tag{A.5}$$

Proof of (A.5)

We begin with a simple probability manipulation:

$$\begin{aligned} p_{\text{rise}}^{(j)} &\triangleq \Pr\left(N(t) \geq C_3\sqrt{\mu\beta R} \text{ at some point during epoch } j\middle|\mathcal{F}_{\tau_j}\right) \\ &\geq \Pr\left(N(t) \geq C_3\sqrt{\mu\beta R} \text{ during the interval } [\tau_j, \min(\tau_j+\beta, \tau_{j+1})]\middle|\mathcal{F}_{\tau_j}\right). \end{aligned}$$

From here, we make with a useful observation: since there are no server *in setup* at the beginning of an epoch (as we have just turned off a server), no servers can complete setup in the first β time of an epoch. Thus, the number of busy servers $Z(t) \leq R-j$ during this time, and, by Claim A.1, the coupled process

$$\tilde{N}(t) \triangleq N(\tau_j) + A(\tau_j, t) - \mathcal{D}[R-j]((\tau_j, t))$$

must be a lower bound on $N(t)$, during the interval $[\tau_j, \tau_j+\beta]$. Moreover, the number of busy servers $Z(t)$ can not be smaller than $R-j$ until the beginning of epoch $j+1$ either. Thus, we find that the behavior of $N(t)$ corresponds *exactly* with the behavior of $\tilde{N}(t)$ during the interval $[\tau_j, \min(\tau_{j+1}, \tau_j+\beta)]$.

We now use this coupled process to analyze our original probability. Define the up-crossing time τ_{up} as

$$\tau_{\text{up}} \triangleq \min\left\{t > 0 : \tilde{N}(\tau_j + t) \geq R + C_3\sqrt{\mu\beta R}\right\}.$$

Likewise, define the down-crossing time τ_{down} as

$$\tau_{\text{down}} \triangleq \min\left\{t > 0 : \tilde{N}(\tau_j + t) \leq R - (j+1)\right\}.$$

It follows that

$$\begin{aligned} &\Pr\left(\text{reach } N(t) \leq R-(j+1) \text{ during the interval } [\tau_j, \min(\tau_j+\beta, \tau_{j+1})]\middle|\mathcal{F}_{\tau_j}\right) \\ &= \Pr\left(\text{reach } \tilde{N}(t-\tau_j) \leq R-(j+1) \text{ during the interval } [\tau_j, \min(\tau_j+\beta, \tau_{j+1})]\middle|\mathcal{F}_{\tau_j}\right) \\ &= \Pr\left(\tau_{\text{up}} \leq \beta, \tau_{\text{up}} < \tau_{\text{down}}\right) \\ &= \Pr\left(\tau_{\text{up}} \leq \beta\right) - \Pr\left(\tau_{\text{up}} \leq \beta, \tau_{\text{up}} \geq \tau_{\text{down}}\right) \\ &= \Pr\left(\tau_{\text{up}} \leq \beta\right) - \Pr\left(\tau_{\text{up}} \leq \beta | \tau_{\text{up}} \geq \tau_{\text{down}}\right)\Pr\left(\tau_{\text{up}} \geq \tau_{\text{down}}\right). \end{aligned}$$

We now observe that

$$\Pr\left(\tau_{\text{up}} \leq \beta | \tau_{\text{up}} \geq \tau_{\text{down}}\right) \leq \Pr\left(\tau_{\text{up}} \leq \beta\right), \tag{A.6}$$

since the process has farther to go, less time to do so, and the process's behavior is translation-invariant (this last point is why we needed to analyze the coupled process instead).

Continuing from where we left off, we find that

$$\begin{aligned}
p_{\text{rise}}^{(j)} &= \Pr\left(\tau_{\text{up}} \leq \beta\right) - \Pr\left(\tau_{\text{up}} \leq \beta | \tau_{\text{up}} \geq \tau_{\text{down}}\right) \Pr\left(\tau_{\text{up}} \geq \tau_{\text{down}}\right) \\
&\geq \Pr\left(\tau_{\text{up}} \leq \beta\right) - \Pr\left(\tau_{\text{up}} \leq \beta\right) \Pr\left(\tau_{\text{up}} \geq \tau_{\text{down}}\right) \\
&= \Pr\left(\tau_{\text{down}} > \tau_{\text{up}}\right) \Pr\left(\tau_{\text{up}} \leq \beta\right) \\
&\geq \Pr\left(\tau_{\text{down}} > \infty\right) \Pr\left(\tau_{\text{up}} \leq \beta\right) \\
&= \frac{j}{R} \Pr\left(\tau_{\text{up}} \leq \beta\right),
\end{aligned}$$

where the last equality is a classical result on upwards-biased discrete random walks (one can think of $\tilde{N}$ as a discrete random walk driven by a Poisson process of rate $(k\lambda + \mu(R-j)$, where the probability that $\tilde{N}$ increases at a Poisson event is $\frac{k\lambda}{k\lambda+\mu(R-j)} = \frac{R}{2R-j}$).

From here, it suffices to lower bound $\Pr\left(\tau_{\text{up}} \leq \beta\right)$. To begin, note

$$\begin{aligned}
\Pr\left(\tau_{\text{up}} \leq \beta\right) &= \Pr\left(\sup_{t\in[0,\beta)} \tilde{N}(t) \geq R + C_3\sqrt{\mu\beta R}\right) \\
&\geq \Pr\left(\tilde{N}(\beta) \geq R + C_3\sqrt{\mu\beta R}\right) \\
&= \Pr\left(A(\tau_j, \tau_j + \beta) - \mathcal{D}\left[R-j\right](\tau_j, \tau_j+\beta) \geq j + C_3\sqrt{\mu\beta R}\right).
\end{aligned}$$

Noting that the number of arrivals $A(\tau_j, \tau_j+\beta)$ and the number of departures $\mathcal{D}\left[R-j\right]([\tau_j, \tau_j+\beta])$ are independent Poisson r.v.'s, we can apply the Berry-Esseen bound of Claim A.6 to find

$$\begin{aligned}
&= 1 - \Phi\left(\frac{\mu\beta j - j - C_3\sqrt{\mu\beta R}}{\sqrt{\mu\beta(2R-j)}}\right) - \frac{1}{3\sqrt{\mu\beta(2R-j)}} \\
&\geq 1 - \Phi\left(\frac{0.99\mu\beta j - C_3\sqrt{\mu\beta R}}{\sqrt{2\mu\beta R}}\right) - \frac{1}{3\sqrt{\mu\beta R}} \\
&= 1 - \Phi\left(-0.99\frac{j}{\sqrt{R}}\sqrt{\mu\beta} + \frac{C_3}{\sqrt{2}}\right) - \frac{1}{3\sqrt{\mu\beta R}} \\
&\geq 1 - \Phi\left(-9.9A_5 + \frac{C_3}{\sqrt{2}}\right) - \frac{1}{300}.
\end{aligned}$$

To complete the proof, we set the constant A_5 such that the final probability is ≥ 0.99. In particular, we need

$$\Phi\left(-9.9A_5 + \frac{C_3}{\sqrt{2}}\right) \leq \frac{2}{300},$$

which is achieved when $A_5 > \frac{C_3}{9.9\sqrt{2}} + 0.25$; choosing $A_5 = 1$ gives the result. □

A.4 Proof of Claim A.6: Berry-Esseen bound for the Skellam distribution.

Claim A.6 (Berry-Esseen bound for the Skellam distribution)**.** *Given two independent random variables $Y_1 \sim$ Poisson(μ_1) and $Y_2 \sim$ Poisson(μ_2), as well as a constant C with $\mu_1 > \mu_2 + C$, one has*

$$\Pr\left(Y_1 - Y_2 \geq C\right) \geq 1 - \Phi\left(-\left[\frac{\mu_1 - \mu_2 - C}{\mu_1 + \mu_2}\right]\right) - \frac{1}{3\sqrt{\mu_1 + \mu_2}}.$$

A.4.1 Proof.

This follows directly from the Poisson Berry-Esseen bound of [4], applied twice. □

A.5 Proof of Claim 6.8: Wait-Busy Period Bound.

Claim 6.8 (Wait-Busy Period Bound)**.** *Let τ be some stopping time. Let the next down-crossing d_{gen}, defined as*

$$d_{\text{gen}} \triangleq \min\left\{t \geq 0 : N(t+\tau) \leq R + h\right\},$$

be the length of time until the number of jobs $N(t) \leq R + h$, where h is some positive integer which is smaller than $k - R$. Suppose that, at time τ, we know that $Z(t+\tau) \geq R$ for any $t \in [0, d_{\text{gen}}]$. Let $\tilde{d_{\text{gen}}}$ be the relative downcrossing time in a coupled system with exactly R busy servers. Then we have the following bound for the time integral of $N(t) - R$ from τ to $\tau + d_{\text{gen}}$:

$$\begin{aligned}\mathbb{E}\left[\int_{\tau}^{\tau+d_{\text{gen}}} [N(t) - R]dt \middle| \mathcal{F}_\tau\right] &\leq [N(\tau) - (R+h)] \cdot \left[\frac{1}{\mu} + \beta + \frac{1}{2\mu h} + \frac{R}{\mu h^2}\right] \\ &+ \left([N(\tau) - (R+h)]^+\right)^2 \frac{1}{2\mu h} \\ &+ \mathbb{E}\left[\min\left(\beta, \tilde{d_{\text{gen}}}\right)\right] \cdot \left[h + \frac{R}{h^2}\right]\end{aligned}$$

This proof is a straightforward application of some martingale theory and some basic queueing results on the M/M/1 busy period.

A.5.1 Proof.

We split the integral in question into two parts:

$$\int_{\tau}^{d_{\text{gen}}} [N(t) - R]\mathrm{d}t = \int_{\tau}^{\min(d_{\text{gen}}, \tau+\beta)} [N(t) - R]\mathrm{d}t + \mathbf{1}_{d_{\text{gen}} > \tau + d_{\text{gen}}} \int_{\tau+\beta}^{d_{\text{gen}}} [N(t) - R]\mathrm{d}t.$$

To find the expectation of the first term, we use our assumption that the number of busy servers $Z(t) \geq R$ and apply Claim A.2, finding

$$\mathbb{E}\left[\int_{\tau}^{\min(d_{\text{gen}}, \tau+\beta)} [N(t) - R]\mathrm{d}t \middle| \mathcal{F}_\tau\right] \leq \beta \cdot [N(\tau) - R].$$

To find the expectation of the second term, we first investigate a conditional version of the expectation. We note that, if, by time $\tau + \beta$, the number of jobs $N(t)$ has not dipped below $R + M$, then we must have at least $R + M$ servers on at time $\tau + \beta$. Moreover, none of these servers can turn off until time d_{gen}. In other words, for the interval $[\tau + \beta, d_{\text{gen}}]$, the number of busy servers $Z(t) \geq R + M$. Thus, we can apply Claim A.5 (a bound on the integral over a busy period with at least $R + M$ servers), and find that

$$\begin{aligned}
&\mathbb{E}\left[\int_{\tau+\beta}^{d_{\text{gen}}} [N(t) - R]\mathrm{d}t \middle| \mathcal{F}_{\tau+\beta}, \tau + \beta < d_{\text{gen}}\right] \\
&\quad \leq [N(\tau + \beta) - (R + M)]^+ \left[\frac{3}{2\mu M} + \frac{1}{\mu} + \frac{R}{\mu M^2}\right] + \frac{\left([N(\tau+\beta) - (R+M)]^+\right)^2}{2\mu M}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}_{d_{\text{gen}} > \tau+\beta} \int_{\tau+\beta}^{d_{\text{gen}}} [N(t) - R]\mathrm{d}t \middle| \mathcal{F}_\tau\right] \\
&= \mathbb{E}\left[\mathbf{1}_{d_{\text{gen}} > \tau+\beta} \mathbb{E}\left[\int_{\tau+\beta}^{d_{\text{gen}}} [N(t) - R]\mathrm{d}t \middle| \mathcal{F}_{\tau+\beta}\right] \middle| \mathcal{F}_\tau\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{d_{\text{gen}} > \tau+\beta}\left[(N(\tau+\beta) - (R+M))\left[\frac{3}{2\mu M} + \frac{1}{\mu} + \frac{R}{\mu M^2}\right] + \frac{(N(\tau+\beta) - (R+M))^2}{2\mu M}\right] \middle| \mathcal{F}_\tau\right]\right] \\
&= \left[\frac{3}{2\mu M} + \frac{1}{\mu} + \frac{R}{\mu M^2}\right] \mathbb{E}\left[\mathbf{1}_{d_{\text{gen}} > \tau+\beta} (N(\tau+\beta) - (R+M)) \middle| \mathcal{F}_\tau\right] \\
&\quad + \left[\frac{1}{2\mu M}\right] \mathbb{E}\left[\mathbf{1}_{d_{\text{gen}} > \tau+\beta} (N(\tau+\beta) - (R+M))^2 \middle| \mathcal{F}_\tau\right].
\end{aligned}$$

Applying Claim A.4, we obtain

$$\begin{aligned}
&\leq \left[\frac{3}{2\mu M} + \frac{1}{\mu} + \frac{R}{\mu M^2}\right][N(\tau) - (R+M)] + \left[\frac{1}{2\mu M}\right]\left([N(\tau) - (R+M)]^2 + 2\mu R\beta\right) \\
&= \left[\frac{3}{2\mu M} + \frac{1}{\mu} + \frac{R}{\mu M^2}\right][N(\tau) - (R+M)] + \left[\frac{1}{2\mu M}\right][N(\tau) - (R+M)]^2 + \frac{R\beta}{M} \\
&= I^{\text{busy}}\left([N(\tau) - (R+M)], M\right) + \frac{R\beta}{M}.
\end{aligned}$$

□

A.6 Proof of Claim 6.7: Bound on $\mathbb{E}[N(T_A)]$.

Claim 6.7 (Upper Bound on $\mathbb{E}[N(T_A)]$)**.** *Recall that* $T_A \triangleq \min\{t > 0 : Z(t) = R + 1\}$. *Then,*

$$\mathbb{E}[N(T_A) - R] \leq F_1 \mu\beta\sqrt{R}\left(1 + \frac{F_2}{\sqrt{\mu\beta}}\right)$$

and
$$\mathbb{E}\left[(N(T_A)-R)^2\right] \le F_1^2(\mu\beta)^2 R\left(1+\frac{F_2}{\sqrt{\mu\beta}}\right)^2 + 2\mu\beta R,$$
where $F_1 = 2.12$ *and* $F_2 = 3.645$.

To prove this claim, we need only bound $\mathbb{E}\left[N(T_A)-R\right]$ and $\mathbb{E}\left[(N(T_A)-R)^2\right]$, since

$$\mathbb{E}\left[I^{\text{busy}}\left(\left[N\left(T_A\right)-R\right],M\right)\right] = \left[\frac{3}{2\mu M}+\frac{1}{\mu}+\frac{R}{\mu M^2}\right]\mathbb{E}\left[N\left(T_A\right)-R\right] + \left[\frac{1}{2\mu M}\right]\mathbb{E}\left[\left[N\left(T_A\right)-R\right]^2\right].$$

To be particular, we show that

$$\mathbb{E}\left[N\left(T_A\right)-R\right] \le F_1\mu\sqrt{R}\beta + F_2\sqrt{\mu R\beta}$$

and
$$\mathbb{E}\left[\left[N\left(T_A\right)-R\right]^2\right] \le F_3\mu^2 R\beta^2 + F_4$$

Proof.

The beginning of the proof will be the same for both of these inequalities. Using the up-crossing and down-crossing decomposition of Section 6.2.1, we know that time T_A occurs either during a *rise* or during a *fall*. Since the number of jobs $N(t) \le R + C_3\sqrt{\mu\beta R}$ during a rise,

$$\left[N\left(T_A\right)-R\right]\mathbf{1}_{T_A \text{ during a rise}} \le C_3\sqrt{\mu\beta R}\mathbf{1}_{T_A \text{ during a rise}}.$$

If T_A occurs during a fall, we need a more nuanced bound. Writing out the event $\{T_A \text{ during a fall}\}$ in terms of disjoint events, we find

$$\{T_A \text{ during a fall}\} = \bigcup_{j=0}^{R}\bigcup_{i=1}^{\infty}\left\{u_i^{(j)} \le T_A < d_i^{(j)}\right\},$$

so that, for $c \in \{1,2\}$,

$$\begin{aligned}\mathbb{E}\left[\left[N\left(T_A\right)-R\right]^c\mathbf{1}_{T_A \text{ during a fall}}\right] &= \sum_{j=0}^{R-1}\sum_{i=1}^{\infty}\mathbb{E}\left[\left[N\left(T_A\right)-R\right]^c\mathbf{1}_{u_i^{(j)}\le T_A<d_i^{(j)}}\right]\\ &= \sum_{j=0}^{R-1}\sum_{i=1}^{\infty}\mathbb{E}\left[\left[N\left(T_A\right)-R\right]^c\mathbf{1}_{T_A<d_i^{(j)}}\middle|\mathcal{F}_{u_i^{(j)}}, n_u^{(j)}\ge i\right]\Pr\left(n_u^j \ge i\right).\end{aligned}$$

To bound this conditional expectation, we apply Claim A.2. Notice that $N(u_i^{(j)}) - R = C_3\sqrt{\mu\beta R}$, the $(R+1)$-th server starts up at time $T_A = u_i^{(j)} + Y_{R+1}(u_i^{(j)})$ if $T_A < d_i^{(j)}$, the time $d_i^{(j)}$ is a hitting time, and that $Z(t) \ge R - j$ until time $\tau_{j+1} \ge d_i^{(j)}$. It follows that

$$\mathbb{E}\left[\left[N\left(T_A\right)-R\right]\mathbf{1}_{T_A<d_i^{(j)}}\middle|\mathcal{F}_{u_i^{(j)}}, n_u^{(j)}\ge i\right] \le C_3\sqrt{\mu\beta R} + \mu j Y_{R+1}(u_i^{(j)}) \le C_3\sqrt{\mu\beta R} + \mu j\beta,$$

and that

$$\mathbb{E}\left[\left[N\left(T_A\right)-R\right]^2\mathbf{1}_{T_A<d_i^{(j)}}\middle|\mathcal{F}_{u_i^{(j)}},n_u^{(j)}\geq i\right]\leq\left(C_3\sqrt{\mu\beta R}+\mu jY_{R+1}\left(u_i^{(j)}\right)\right)^2+\mu 2RY_{R+1}\left(u_i^{(j)}\right)$$
$$\leq\left(C_3\sqrt{\mu\beta R}+\mu j\beta\right)^2+\mu 2R\beta$$
$$=C_3^2\mu\beta R+2C_3\sqrt{\mu\beta R}\mu\beta j+(\mu\beta)^2j^2+2\mu R\beta.$$

It now suffices to bound $\sum_j\sum_i j^c\Pr\left(n_u^{(j)}\geq i\right)$, where $c\in\{0,1,2\}$. We do this via the same method used in Section 6.2.1:

$$\sum_{j=1}^{R}\sum_{i=1}^{\infty}j^c\Pr\left(n_u^{(j)}\geq i\right)\leq\sum_{j=1}^{R}\sum_{i=1}^{\infty}j^c\Pr\left(n_e\geq j\right)p_{\text{rise}}^{(j)}(1-p_2)^{i-1}$$
$$=\frac{1}{p_2}\sum_{j=1}^{R}j^cp_{\text{rise}}^{(j)}\Pr\left(n_e\geq j\right)$$
$$\leq\frac{1}{C_4p_2}\sum_{j=1}^{R}j^cC_4p_{\text{rise}}^{(j)}\prod_{\ell=0}^{j-1}\left(1-C_4p_{\text{rise}}^{(\ell)}\right).$$

This is simply the expectation of a time-varying geometric random variable G, with $\Pr\left(G=j|G\geq j\right)=C_4p_{\text{rise}}^{(j)}$. It follows that if one lower-bounds $p_{\text{rise}}^{(j)}$, then an upper bound on the desired expectation is obtained. Applying Claim 6.3, we note that we are essentially bounding G using $Y\sim$ Geometric $\left(\frac{0.99C_4A_5}{\sqrt{R}}\right)$ and saying $Y+A_5\sqrt{R}$ stochastically-dominates G. It follows that

$$\mathbb{E}\left[G\right]\leq A_5\sqrt{R}+\frac{1}{0.99C_4A_5}\sqrt{R}$$

and that, for any b,

$$\mathbb{E}\left[(G+b)^2\right]\leq\mathbb{E}\left[(Y+A_5\sqrt{R}+b)^2\right]=\mathbb{E}\left[Y^2\right]+2(A_5\sqrt{R}+b)\mathbb{E}\left[Y\right]+(A_5\sqrt{R}+b)^2$$
$$=2\mathbb{E}\left[Y\right]^2-\mathbb{E}\left[Y\right]+2\left(A_5\sqrt{R}+b\right)\mathbb{E}\left[Y\right]+\left(A_5\sqrt{R}+b\right)^2$$
$$\leq\left(\mathbb{E}\left[Y\right]+A_5\sqrt{R}+b\right)^2+\mathbb{E}\left[Y\right]^2.$$

Defining $B_5\triangleq\frac{C_3}{C_4p_2}$, $B_6\triangleq\frac{1}{C_4p_2}\left(\frac{1}{0.99C_4A_5}+A_5\right)$, and $B_7\triangleq\frac{1}{2C_4p_2}\left[\frac{1}{(0.99C_4A_5)^2}+2\right]$, it follows that

$$
\begin{aligned}
&\mathbb{E}\left[[N(T_A)-R]^2\,\mathbf{1}_{T_A \text{ during a fall}}\right] \\
&\quad\le \frac{1}{C_4p_2}\sum_{j=1}^{R} C_4p_{\text{rise}}^{(j)}\left[\left(C_3\sqrt{\mu\beta R}+\mu j\beta\right)^2+\mu 2R\beta\right]\prod_{\ell=0}^{j-1}\left(1-C_4p_{\text{rise}}^{(\ell)}\right) \\
&\quad= \frac{1}{C_4p_2}\mathbb{E}\left[\left(C_3\sqrt{\mu\beta R}+\mu G\beta\right)^2+\mu 2R\beta\right] \\
&\quad\le \frac{1}{C_4p_2}\mathbb{E}\left[\left(C_3\sqrt{\mu\beta R}+\mu\beta Y+\mu\beta A_5\sqrt{R}\right)^2+\mu 2R\beta\right] \\
&\quad= \frac{1}{C_4p_2}\left[\left(C_3\sqrt{\mu\beta R}+\mu\beta\frac{1}{0.99C_4A_5}\sqrt{R}+\mu\beta A_5\sqrt{R}\right)^2+\mu\beta\frac{1}{(0.99C_4A_5)^2}R+2\mu\beta R\right] \\
&\quad\le \frac{1}{C_4^2p_2^2}\left[\left(C_3\sqrt{\mu\beta R}+\mu\beta\frac{1}{0.99C_4A_5}\sqrt{R}+\mu\beta A_5\sqrt{R}\right)^2\right]+\frac{1}{C_4p_2}\left[\frac{1}{(0.99C_4A_5)^2}+2\right]\mu\beta R \\
&= \left(B_5\sqrt{\mu\beta R}+B_6\mu\beta\sqrt{R}\right)^2+2B_7\mu\beta R.
\end{aligned}
$$

and that

$$
\begin{aligned}
&\mathbb{E}\left[[N(T_A)-R]\,\mathbf{1}_{T_A \text{ during a fall}}\right] \\
&\quad\le \frac{1}{C_4p_2}\sum_{j=1}^{R} C_4p_{\text{rise}}^{(j)}\left[C_3\sqrt{\mu\beta R}+\mu j\beta\right]\prod_{\ell=0}^{j-1}\left(1-C_4p_{\text{rise}}^{(\ell)}\right) \\
&\quad= \frac{1}{C_4p_2}\mathbb{E}\left[C_3\sqrt{\mu\beta R}+\mu\beta G\right] \\
&\quad\le \frac{1}{C_4p_2}\left[C_3\sqrt{\mu\beta R}+\mu\beta\left(A_5\sqrt{R}+\frac{1}{0.99C_4A_5}\sqrt{R}\right)\right] \\
&\quad= \left(B_5\sqrt{\mu\beta R}+B_6\mu\beta\sqrt{R}\right).
\end{aligned}
$$

It follows that

$$\mathbb{E}\left[N(T_A)-R\right]\le\left(B_5\sqrt{\mu\beta R}+B_6\mu\beta\sqrt{R}\right)$$

and that

$$\mathbb{E}\left[[N(T_A)-R]^2\right]\le\left(B_5\sqrt{\mu\beta R}+B_6\mu\beta\sqrt{R}\right)^2+2B_7\mu\beta R.$$

Thus, we have that, using implicitly the formula from Claim A.5,

$$\mathbb{E}\left[I^{\text{busy}}\left([N(T_A)-R],M\right)\right]=I^{\text{busy}}\left(B_5\sqrt{\mu\beta R}+B_6\mu\beta\sqrt{R},M\right)+B_7\frac{\beta R}{M},$$

completing the claim. □

A.7 Hitting Time Bounds

A.7.1 Proof of Claim A.7: Discrete-Time Hitting Time Tail Bound.

Claim A.7 (Discrete-Time Hitting Time Tail Bound)**.** *Suppose one has an upwards-biased discrete random walk $V(t)$ where in each step*

$$\Pr\left(V(t+1) = V(t)+1 | \mathcal{F}_t\right) = p = 1-q,$$

where $p \geq \frac{1}{2} \geq q$. Suppose that $V(0) = 1$ and let the hitting time $\gamma \triangleq \min\{t \in \mathcal{N} : V(t) = 0\}$ be the first timestep where the walk $V(t) = 0$. Then, for $n \geq 1$,

$$\Pr\left(\gamma \geq 2m+1\right) \leq \frac{1}{\sqrt{\pi}} \frac{2q}{\sqrt{m}} \left(1 + \frac{1}{2(m+1)}\right).$$

Moreover, if $p = q = \frac{1}{2}$, then

$$\Pr\left(\gamma \geq 2m+1\right) \geq \frac{1}{\sqrt{\pi}} e^{-\frac{1}{6m}} \frac{1}{\sqrt{m+1}}.$$

Proof

We first note, as in [32], that by a counting argument $\Pr\left(\gamma = 2\ell+1\right) = q\left(qp\right)^{\ell} C_{\ell}$, where $C_{\ell} \triangleq \frac{1}{\ell+1}\frac{(2\ell)!}{\ell!\ell!}$ is the ℓ-th Catalan number; note that γ can not be even, since the number of downward steps must exceed the number of upward steps by exactly 1.

We proceed by bounding the Catalan numbers using Stirling's approximation. For $m = 0$, then $\Pr\left(\gamma \geq 1\right) = \Pr\left(\gamma \geq 2\right) = p$, i.e. the probability that the first step is an upward step. For $m \geq 1$, applying Stirling's approximation and simplifying gives

$$e^{-\frac{1}{6\ell}} \frac{1}{\sqrt{\pi \ell}(\ell+1)} q\left(4pq\right)^{\ell} \leq \Pr\left(\gamma = 2\ell+1\right) \leq \frac{1}{\sqrt{\pi\ell}(\ell+1)} q\left(4pq\right)^{\ell}.$$

Lower bound. Since we are interested in the lower bound only when $q = p = \frac{1}{2}$, we obtain that

$$\begin{aligned}
\Pr\left(\gamma \geq 2m+1\right) &\geq \frac{1}{\sqrt{pi}} \frac{1}{2} \sum_{\ell=m}^{\infty} \frac{e^{-\frac{1}{6\ell}}}{\sqrt{\ell}(\ell+1)} \\
&\geq \frac{1}{\sqrt{\pi}} e^{-\frac{1}{6m}} \frac{1}{2} \sum_{\ell=m}^{\infty} \frac{1}{\sqrt{\ell}(\ell+1)} \\
&\geq \frac{1}{\sqrt{\pi}} e^{-\frac{1}{6m}} \frac{1}{2} \int_{m}^{\infty} \frac{1}{(\ell+1)^{3/2}} \mathrm{d}\ell \\
&= \frac{1}{\sqrt{\pi}} e^{-\frac{1}{6m}} \frac{1}{\sqrt{m+1}}.
\end{aligned}$$

Upper bound. Noting that $4pq \le 1$, we have likewise that

$$\begin{aligned}\Pr(\gamma \ge 2m+1) &\le \frac{1}{\sqrt{\pi}} q \sum_{\ell=m}^{\infty} \frac{1}{\sqrt{\ell}(\ell+1)} \\ &\le \frac{1}{\sqrt{\pi}} q \frac{1}{\sqrt{m}(m+1)} + \int_m^\infty \frac{1}{\ell^{3/2}} \mathrm{d}\ell \\ &= \frac{1}{\sqrt{\pi}} q \frac{2}{\sqrt{m}} \left(1 + \frac{1}{2(m+1)}\right).\end{aligned}$$

A.7.2 Proof of Claim A.8:[Continuous-Time Hitting Time Tail Bound.

We further extend this discrete-time bound into a continuous-time bound.

Claim A.8 (Continuous-Time Hitting Time Tail Bound)**.** *Suppose one has an Poisson arrival process $Y_A(t)$ of rate $k\lambda$ and a Poisson departure process $Y_D(t)$ of rate $\mu(R-j)$, for some integer $j \ge 0$. Let the continuous random walk $X_c(t) = Y_A(t) - Y_D(t)$, with $X_c(0) = 1$, and define $\gamma_c \triangleq \min\{t > 0 : X_c(t) = 0\}$. Let $\nu = (2R-j)\mu t$. For any $\nu \ge 3$, we have*

$$\Pr(\gamma_c \ge t) \le \frac{b_1}{\sqrt{2}} \left(\frac{1}{\sqrt{\nu}} + \frac{b_2}{\nu^{3/2}}\right)$$

where $b_1 = \sqrt{\frac{2}{\pi}}$ and $b_2 = 1 + \frac{2.5}{b_1\sqrt{2}}$.

Moreover, if $j = 0$, then

$$\Pr(\gamma_c \ge t) \ge \frac{b_1}{\sqrt{2}} e^{-\frac{1}{3(\nu-1)}} \frac{1}{\sqrt{\nu+2}}.$$

Proof of Upper Bound.

To prove this claim, we first condition on the value of $Y_T = Y_A(t) + Y_D(t)$, the total number of Poisson events during the interval $[0, t]$, then relate that to the same question in a discrete-time random walk, a la Claim A.7. Note that $Y_T \sim \text{Poisson}(\nu)$, and thus

$$\begin{aligned}\Pr(\gamma_c \ge t) &= \Pr(\gamma \ge Y_T) \\ &= \sum_{j=0}^{\infty} e^{-\nu} \frac{\nu^j}{j!} \Pr(\gamma \ge j) \\ &= e^{-\nu} + 2p\nu e^{-\nu} + \sum_{j=3}^{\infty} e^{-\nu} \frac{\nu^j}{j!} \Pr(\gamma \ge j + \mathbf{1}_{j \text{ is even}}) \\ &= e^{-\nu} + 2p\nu e^{-\nu} + \sum_{j=0}^{\infty} e^{-\nu} \frac{\nu^j}{j!} \Pr\left(\gamma \ge 2\left(\frac{j + \mathbf{1}_{j \text{ is even}} - 1}{2}\right) + 1\right).\end{aligned}$$

Applying the discrete upper bound to the sum, we obtain

$$\begin{aligned}
&\sum_{j=3}^{\infty} e^{-\nu}\frac{\nu^j}{j!}\Pr\left(\gamma \geq 2\left(\frac{j+\mathbf{1}_{j \text{ is even}}-1}{2}\right)+1\right) \\
&\leq b_1\sqrt{2}q\sum_{j=3}^{\infty} e^{-\nu}\frac{\nu^j}{j!}\frac{1}{\sqrt{j+\mathbf{1}_{j \text{ is even}}-1}}\left(1+\frac{1}{(j+\mathbf{1}_{j \text{ is even}}+1)}\right) \\
&= b_1\sqrt{2}q\frac{1}{\nu}\sum_{j=3}^{\infty} e^{-\nu}\frac{\nu^{(j+1)}}{(j+1)!}\frac{1}{\sqrt{j+\mathbf{1}_{j \text{ is even}}-1}}\left(j+1+\frac{j+1}{j+\mathbf{1}_{j \text{ is even}}+1}\right) \\
&\leq b_1\sqrt{2}q\frac{1}{\nu}\sum_{j=3}^{\infty} e^{-\nu}\frac{\nu^{(j+1)}}{(j+1)!}\frac{j+2}{\sqrt{j+\mathbf{1}_{j \text{ is even}}-1}} \\
&\leq b_1\sqrt{2}q\frac{1}{\nu}\sum_{j=3}^{\infty} e^{-\nu}\frac{\nu^{(j+1)}}{(j+1)!}\frac{j+\mathbf{1}_{j \text{ is even}}-1+3}{\sqrt{j+\mathbf{1}_{j \text{ is even}}-1}} \\
&= b_1\sqrt{2}q\frac{1}{\nu}\sum_{j=3}^{\infty} e^{-\nu}\frac{\nu^{(j+1)}}{(j+1)!}\left(\sqrt{j+\mathbf{1}_{j \text{ is even}}-1}+\frac{3}{\sqrt{j+\mathbf{1}_{j \text{ is even}}-1}}\right).
\end{aligned}$$

From here, we note that the function $f(x) = \sqrt{x} + \frac{3}{\sqrt{x}}$ is both increasing and concave for all $x \geq 3$. After increasing the argument and applying Jensen's inequality, we find that

$$\begin{aligned}
&\leq b_1\sqrt{2}q\frac{1}{\nu}\sum_{j=3}^{\infty} e^{-\nu}\frac{\nu^{(j+1)}}{(j+1)!}\left(\sqrt{j+1}+\frac{3}{\sqrt{j+1}}\right) \\
&\leq b_1\sqrt{2}q\frac{1}{\nu}\left(\sqrt{\nu}+\frac{3}{\sqrt{\nu}}\right),
\end{aligned}$$

where in the final line we have used that the function $f(x)$ is increasing in x for any $x \geq 3$, and that $\mathbb{E}\left[Y_T\mathbf{1}_{Y_T\geq 4}\right] \geq \nu - 3 \geq 3$. Thus, we have that

$$\begin{aligned}
\Pr\left(\gamma_c \geq t\right) &\leq (3\nu)\, e^{-\nu} + 2q\sqrt{\frac{2}{\pi}}\left(\frac{1}{\sqrt{\nu}}+\frac{1}{\nu^{3/2}}\right) \\
&\leq \frac{2.5}{\nu^{3/2}} + 2q\sqrt{\frac{2}{\pi}}\left(\frac{1}{\sqrt{\nu}}+\frac{1}{\nu^{3/2}}\right)
\end{aligned}$$

Proof of Lower Bound.

We approach the initial stages of the proof in the precisely the same way, obtaining

$$\begin{aligned}
\Pr(\gamma_c \geq t) &= \Pr(\gamma \geq Y_T) \\
&= e^{-\nu} + 2p\nu e^{-\nu} + \sum_{j=3}^{\infty} e^{-\nu}\frac{\nu^j}{j!}\Pr\left(\gamma \geq 2\left(\frac{j + \mathbf{1}_{j \text{ is even}} - 1}{2}\right) + 1\right) \\
&\geq \sum_{j=3}^{\infty} e^{-\nu}\frac{\nu^j}{j!} b_1 e^{-\frac{1}{3(j+\mathbf{1}_{j \text{ is even}}-1)}} \frac{q\sqrt{2}}{\sqrt{(j + \mathbf{1}_{j \text{ is even}} + 1)}} \\
&\geq \sum_{j=3}^{\infty} e^{-\nu}\frac{\nu^j}{j!} b_1 e^{-\frac{1}{3(j-1)}} \frac{q\sqrt{2}}{\sqrt{(j+2)}}.
\end{aligned}$$

Applying Jensen's inequality, we obtain

$$\geq b_1 q\sqrt{2} e^{-\frac{1}{3(\nu-1)}} \frac{1}{\sqrt{\nu+2}}.$$

□

A.7.3 Proof of Claim A.9: Bound on Expected Length of Stopped Random Walk.

Claim A.9 (Bound on Expected Length of Stopped Random Walk)**.** *Suppose we have a critically loaded M/M/1 queue with arrival rate and departure rate both equal to $k\lambda$. Suppose also that at time 0, a job arrives. Let τ be the length of the busy period which follows. Then,*

$$\mathbb{E}[\min(\beta, \tau)] \leq \frac{1}{\mu}\frac{2b_1\sqrt{\mu\beta}}{\sqrt{R}} + \frac{6}{\mu R}.$$

Proof of Claim A.9.

From Claim A.8, the continuous-time hitting time bound, we have that

$$\Pr(\tau \geq t) \leq \frac{b_1}{\sqrt{2}}\left(\frac{1}{\sqrt{\nu}} + \frac{b_2}{\nu^{3/2}}\right), \tag{A.7}$$

where $\nu = 2\mu R t$ and we require that $\nu \geq 3$. By integrating this bound (using a bound of 1 wherever this bound is ≥ 1), and using the fact that the setup time β and the offered load R are both at least 100, we obtain

$$\mathbb{E}[\min(\beta, \tau)] = \int_0^{\beta} \Pr(\tau > t)\, \mathrm{d}t \leq \frac{1}{\mu}\frac{2b_1\sqrt{\mu\beta}}{\sqrt{R}} + \frac{6}{\mu R}.$$

as desired. □

A.8 Claim A.10: M/M/∞ Passage Time Bound.

Claim A.10 (M/M/∞ Passage Time Bound)**.** *Given an M/M/∞ queue, let $T_{x\to y}$ denote the random amount of time taken to go from state x to state y. Suppose this system has an arrival rate of $k\lambda$ and a per-server departure rate of μ. Let $R \triangleq k\frac{\lambda}{\mu}$. Then, for any h such that $1 \le h \le \sqrt{R}$,*

$$\mathbb{E}\left[T_{(R+h-1)\to(R+h)}\right] \le \frac{\sqrt{2\pi}}{\mu\sqrt{R}}\left(1+\frac{h}{R}\right)^{h-\frac{1}{2}} e^{\frac{1}{12R}} \le D_2\frac{\sqrt{\pi}}{\mu\sqrt{R}}.$$

Proof.

The proof here is quite simple. First, we note that the passage time in the M/M/∞ from state $(R+h-1)$ to state $(R+h)$ is exactly the passage time from those states in the M/M/$(R+h)$/$(R+h)$. This new system has a nice product form, so that

$$\begin{aligned}
\mathbb{E}\left[T_{(R+h-1)\to(R+h)}\right] &\le \mathbb{E}\left[T_{(R+h)\to(R+h)}\right] \\
&= \frac{1}{\mu(R+h)}\frac{1}{\pi_{R+h}} \\
&= \frac{1}{\mu(R+h)}\frac{\sum_{i=0}^{R+h}\frac{R^i}{i!}}{\frac{R^{R+h}}{(R+h)!}} \\
&\le \frac{1}{\mu(R+h)}e^R\frac{(R+h)!}{R^{R+h}} \\
&\le \frac{1}{\mu(R+h)}e^R\frac{e^{\frac{1}{12(R+h)}}\sqrt{2\pi(R+h)}(R+h)^{R+h}e^{-(R+h)}}{R^{R+h}} \\
&\le e^{\frac{1}{12R}}\frac{1}{\mu}\frac{1}{\mu\sqrt{R+h}}\sqrt{2\pi}\left(1+\frac{h}{R}\right)^{R+h} e^{-h} \\
&\le \frac{\sqrt{2\pi}}{\mu\sqrt{R}}\left(1+\frac{h}{R}\right)^{h-\frac{1}{2}} e^{\frac{1}{12R}} \\
&\le \frac{1}{\mu}\frac{\sqrt{2\pi}}{\sqrt{R}}e^{\frac{h^2}{R}}e^{\frac{1}{12R}} \\
&\le \frac{7}{\mu\sqrt{R}},
\end{aligned}$$

where we have made extensive use of Stirling's approximation and the bound $(1+x) \le e^x$.

A.9 Proof of (5.3): Bound on the first long index $\mathbb{E}[L]$.

Proof Outline. We prove (5.3) by first showing that

$$\Pr\left(L > j | L \ge j\right) \ge \left(1-\frac{j}{R}\right)\left(1-\frac{b_1}{\sqrt{\mu\beta R}}\right), \tag{A.8}$$

where $b_1 = \frac{2}{\sqrt{\pi}}$. Next, we show that this implies that, for any $\delta \in (0,1)$ and any $j < \delta R$,

$$\Pr(L > j) \geq \left(1 - \frac{b_1}{\sqrt{\mu\beta R}}\right)^{j+1} e^{-\frac{j(j+1)}{2R}\frac{1}{1-\delta}}. \tag{A.9}$$

From here, we use the sum of tails formula $\mathbb{E}[L] = \sum_{j=0}^{\infty} \Pr(L > j)$ to show

$$\mathbb{E}[L] \geq \left(1 - \frac{b_1}{\sqrt{\mu\beta R}}\right)\left(\left[\sqrt{\frac{\pi}{2}}(1-\delta) - \frac{1.15(1-\delta)}{\sqrt{\mu\beta}}\right]\sqrt{R} - \frac{1}{2} - \frac{2(1-\delta)}{\delta}e^{-R\frac{\delta^2}{1-\delta}}\right). \tag{A.10}$$

Choosing $\delta = \frac{2}{\sqrt{R}}$ then noting that $\mu\beta \geq 100$ and $R \geq 100$ gives the result.

Proof of (A.8)**.**

Recall that an epoch j is *long* if $\tau_{j+1} - \tau_j > \beta$, that L is the index of the first long epoch, and that, if $L \geq j$, then we learn that $L \geq j$ precisely at time τ_j, i.e. when epoch j begins. Moreover, since the system is Markovian, the behavior of the system from τ_j onwards is completely independent of what happened previously. Thus,

$$\Pr(L > j | L \geq j) = \Pr\left(L > j \middle| \mathcal{F}_{\tau_j}, L \geq j\right) = \Pr\left(\tau_{j+1} - \tau_j \leq \beta \middle| \mathcal{F}_{\tau_j}, L \geq j\right) = \Pr(\tau_{j+1} - \tau_j \leq \beta).$$

From here, we note that the random time $\tau_{j+1} - \tau_j$ is a stopping time; a hitting time, to be exact. Moreover, since the number of servers $Z(t)$ can not increase before time $\tau_j + \beta$ and can not decrease until τ_{j+1}, we have that the coupled process $\tilde{N}(t)$ defined as

$$\tilde{N}(t - \tau_j) \triangleq 1 + A(\tau_j, t) - \mathcal{D}[R - j]((\tau_j, t))$$

is in correspondence with $N(t)$; in particular,

$$N(t) = \tilde{N}(t - \tau_j) + R - j - 1$$

for any time $t \in [\tau_j, \min(\tau_j + \beta, \tau_{j+1})]$. If we define the coupled hitting time $\gamma_c \triangleq \min\left\{t > 0 : \tilde{N}(t) \leq 0\right\}$, then we also have that the hitting time $\gamma_c = \tau_{j+1} - \tau_j$, whenever the event $\{\tau_{j+1} - \tau_j \leq \beta\}$ occurs. From here, we can apply Claim A.8 to find that

$$\Pr(\gamma_c \leq \beta) \geq \left(1 - \frac{j}{R}\right)\left(1 - \frac{b_1}{\sqrt{\mu\beta R}}\right), \text{ as desired.}$$

Proof of (A.9)**.**

Having shown the above bound on the conditional extension of the tail, we note that, for $j \leq \delta R$

$$\begin{aligned}
\Pr\left(L \geq j+1\right) &= \Pr\left(L \geq j+1 | L \geq j\right) \Pr\left(L \geq j | L \geq j-1\right) \cdots \Pr\left(L \geq 1\right) \\
&\geq \prod_{i=0}^{j} \left(1 - \frac{i}{R}\right) \left(1 - \frac{b_1}{\sqrt{\mu\beta R}}\right) \\
&\geq \left(1 - \frac{b_1}{\sqrt{\mu\beta R}}\right)^{j+1} \prod_{i=0}^{j} e^{-\frac{i}{R-i}} \\
&\geq \left(1 - \frac{b_1}{\sqrt{\mu\beta R}}\right)^{j+1} e^{-\sum_{i=0}^{j} \frac{i}{R-i}} \\
&= \left(1 - \frac{b_1}{\sqrt{\mu\beta R}}\right)^{j+1} e^{-\sum_{i=0}^{j} \frac{i}{R} \frac{R}{R-j}} \\
&= \left(1 - \frac{b_1}{\sqrt{\mu\beta R}}\right)^{j+1} e^{-\frac{j(j+1)}{2R} \frac{1}{1-\delta}},
\end{aligned}$$

as desired. □

Proof of (A.10)**.**

We now complete the proof. Let $a \triangleq \frac{1}{2R} \frac{1}{1-\delta}$ and $\psi \triangleq -\ln\left(1 - \frac{b_1}{\sqrt{\mu\beta R}}\right)$ as a shorthand. Then we can rewrite (A.9) as

$$\Pr\left(L \geq j+1\right) \geq e^{-aj^2 - (\psi + a)j - \psi}.$$

Now, using the sum-of-tails formula for expectations, we find that

$$\begin{aligned}
\mathbb{E}\left[L\right] &= \sum_{j=0}^{R-1} \Pr\left(L \geq j+1\right) \\
&\geq \sum_{j=0}^{\delta R - 1} \Pr\left(L \geq j+1\right) \\
&\geq \sum_{j=0}^{\delta R - 1} e^{-aj^2 - (\psi + a)j - \psi} \\
&\geq \int_0^{\delta R} e^{-aj^2 - (\psi + a)j - \psi} \mathrm{d}j \\
&= \int_0^{\delta R} e^{-a\left(j^2 + \left(\frac{\psi}{a} + 1\right)j\right) - \psi} \mathrm{d}j \\
&= \int_0^{\delta R} e^{-a\left(j + \frac{1}{2}\left(\frac{\psi}{a} + 1\right)\right)^2 + \frac{a}{4}\left(\frac{\psi}{a} + 1\right)^2 - \psi} \mathrm{d}j \\
&= e^{\frac{a}{4}\left(\frac{\psi}{a} + 1\right)^2 - \psi} \int_0^{\delta R} e^{-a\left(j + \frac{1}{2}\left(\frac{\psi}{a} + 1\right)\right)^2} \mathrm{d}j.
\end{aligned}$$

Evaluating the integral further, we find that

$$\int_0^{\delta R} e^{-a\left(j+\frac{1}{2}\left(\frac{\psi}{a}+1\right)\right)^2}\mathrm{d}j = \int_{\frac{1}{2}\left(\frac{\psi}{a}+1\right)}^{\delta R+\frac{1}{2}\left(\frac{\psi}{a}+1\right)} e^{-aj^2}\mathrm{d}j$$
$$= \int_0^{\infty} e^{-aj^2}\mathrm{d}j - \int_0^{\frac{1}{2}\left(\frac{\psi}{a}+1\right)} e^{-aj^2}\mathrm{d}j - \int_{\delta R+\frac{1}{2}\left(\frac{\psi}{a}+1\right)}^{\infty} e^{-aj^2}\mathrm{d}j.$$

We now bound each of these integrals in turn. First, we know classically that

$$\int_0^{\infty} e^{-aj^2}\mathrm{d}j = \frac{1}{2}\sqrt{\frac{\pi}{a}} = \sqrt{\frac{\pi}{2}}\cdot\sqrt{1-\delta}\sqrt{R} \geq \sqrt{\frac{\pi}{2}}\cdot(1-\delta)\sqrt{R}$$

Next, we note that, since the integrand is ≤ 1,

$$\begin{aligned}\int_0^{\frac{1}{2}\left(\frac{\psi}{a}+1\right)} e^{-aj^2}\mathrm{d}j &\leq \frac{1}{2}\left(\frac{\psi}{a}+1\right)\\ &= \frac{1}{2}\left(2R(1-\delta)\ln\left(\frac{1}{1-\frac{b_1}{\sqrt{\mu\beta R}}}\right)\right)+\frac{1}{2}\\ &\leq R(1-\delta)\frac{b_1}{\sqrt{\mu\beta R}}\frac{1}{1-\frac{b_1}{\sqrt{\mu\beta R}}}+\frac{1}{2}\\ &\leq \left(\frac{1}{\sqrt{\beta}}\right)(1-\delta)\cdot\frac{100\cdot b_1}{100-b_1}\sqrt{R}+\frac{1}{2}\\ &\leq \frac{1.15(1-\delta)}{\sqrt{\mu\beta}}\sqrt{R}+\frac{1}{2}.\end{aligned}$$

Finally, we have that,

$$\int_{\delta R+\frac{1}{2}\left(\frac{\psi}{a}+1\right)}^{\infty} e^{-aj^2}\mathrm{d}j \leq \int_{\delta R}^{\infty} e^{-aj^2}\mathrm{d}j \leq \int_{\delta R}^{\infty} e^{-a\delta Rj}\mathrm{d}j = \frac{1}{a\delta R}e^{-a\delta^2R^2} = \frac{2(1-\delta)}{\delta}e^{-R\frac{\delta^2}{1-\delta}}.$$

To complete the proof, we note that $e^{-\psi} = \left(1-\frac{b_1}{\sqrt{\mu\beta R}}\right)$, thus

$$\begin{aligned}\mathbb{E}\left[L\right] &\geq e^{\frac{a}{4}\left(\frac{\psi}{a}+1\right)^2-\psi}\int_0^{\delta R} e^{-a\left(j+\frac{1}{2}\left(\frac{\psi}{a}+1\right)\right)^2}\mathrm{d}j\\ &\geq e^{-\psi}\int_0^{\delta R} e^{-a\left(j+\frac{1}{2}\left(\frac{\psi}{a}+1\right)\right)^2}\mathrm{d}j\\ &\geq \left(1-\frac{b_1}{\sqrt{\mu\beta R}}\right)\left[\sqrt{\frac{\pi}{2}}(1-\delta)-\frac{1.15(1-\delta)}{\sqrt{\mu\beta}}\right]\sqrt{R}-\frac{1}{2}-\frac{2(1-\delta)}{\delta}e^{-R\frac{\delta^2}{1-\delta}}\\ &= \left(1-\frac{b_1}{\sqrt{\mu\beta R}}\right)\left[\sqrt{\frac{\pi}{2}}(1-\delta)-\frac{1.15(1-\delta)}{\sqrt{\mu\beta}}\right]\sqrt{R}-\frac{1}{2}-\frac{2(1-\delta)}{\delta}e^{-R\frac{\delta^2}{1-\delta}}\\ &= \left(1-\frac{b_1}{\sqrt{\mu\beta R}}\right)\left(\left[\sqrt{\frac{\pi}{2}}(1-\delta)-\frac{1.15(1-\delta)}{\sqrt{\mu\beta}}\right]\sqrt{R}-\frac{1}{2}-\frac{2(1-\delta)}{\delta}e^{-R\frac{\delta^2}{1-\delta}}\right).\end{aligned}$$

From here, we could choose δ to maximize our lower bound further based on system parameters, but a simple choice is $\delta = \frac{2}{\sqrt{R}}$. This gives

$$\begin{aligned}
\mathbb{E}\left[L\right] &\geq \left(1 - \frac{b_1}{\sqrt{\mu\beta R}}\right)\left[\left(1 - \frac{2}{\sqrt{R}}\right)\left(\sqrt{\frac{\pi}{2}} - \frac{1.15}{\sqrt{\mu\beta}} - 2e^{-4}\right) - \frac{1}{2\sqrt{R}}\right]\sqrt{R} \\
&\geq \frac{2}{3}\sqrt{\frac{\pi}{2}}\sqrt{R},
\end{aligned}$$

as desired.

Bibliography

[1] Jesus R Artalejo, Antonis Economou, and Maria Jesus Lopez-Herrero. Analysis of a Multiserver Queue with Setup Times. *Queueing Syst.*, 51(1):53–76, 2005. 2.1.2

[2] Wolfgang Bischof. Analysis of M/G/1-Queues with Setup Times and Vacations under Six Different Service Disciplines. *Queueing Syst.*, 39(4):265–301, 2001. 2.1.1

[3] Gautam Choudhury. On a batch arrival Poisson queue with a random setup time and vacation period. *Comp. & Oper. Res.*, 25(12):1013–1026, 1998. 2.1.1

[4] John D. Cook. Details for error bound on normal approximation to the Poisson distribution. `https://www.johndcook.com/blog/berry_esseen_poisson/`. Accessed: 01/16/2024. A.1.3, A.4.1

[5] Andrew Daw and Jamol Pender. Queues driven by hawkes processes. *Stochastic Systems*, 8(3):192–229, 2018. 8.3

[6] Andrew Daw, Robert C Hampshire, and Jamol Pender. How to staff when customers arrive in batches. *arXiv preprint arXiv:1907.12650*, 2019. 9.3.2

[7] Zohar Feldman, Avishai Mandelbaum, William A Massey, and Ward Whitt. Staffing of time-varying queues to achieve time-stable performance. *Management Science*, 54(2):324–338, 2008. 8.3, 9.3.2

[8] Anshul Gandhi and Mor Harchol-Balter. How Data Center Size Impacts the Effectiveness of Dynamic Power Management. In *Proc. Ann. Allerton Conf. Communication, Control and Computing*, pages 1164–1169, Urbana-Champaign, IL, September 2011. 2.4

[9] Anshul Gandhi and Mor Harchol-Balter. M/G/k with staggered setup. *Oper. Res. Lett.*, 41 (4):317–320, 2013. 2.1.2

[10] Anshul Gandhi, Varun Gupta, Mor Harchol-Balter, and Michael Kozuch. Optimality analysis of energy-performance trade-off for server farm management. In *Proc. Int. Symp. Computer Performance, Modeling, Measurements and Evaluation (IFIP Performance)*, Namur, Belgium, November 2010. 2.4

[11] Anshul Gandhi, Mor Harchol-Balter, and Ivo Adan. Server farms with setup costs. *Performance Evaluation*, 67(11):1123–1138, 2010. 2.1.2, 2.2.1, 2.2.2, 2.4

[12] Anshul Gandhi, Mor Harchol-Balter, and Mike Kozuch. The case for sleep states in servers. In *SOSP Workshop on Power-Aware Computing and Systems (HotPower)*, pages 1–5, Cascais, Portugal, October 2011. 2.4

[13] Anshul Gandhi, Mor Harchol-Balter, and Mike Kozuch. Are sleep states effective in data

centers? In *Int. Conf. Green Computing (IGCC)*, pages 1–10, San Jose, CA, 2012. 2.4

[14] 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):1–26, 2012. 2.4

[15] Anshul Gandhi, Sherwin Doroudi, Mor Harchol-Balter, and Alan Scheller-Wolf. Exact analysis of the M/M/k/setup class of Markov chains via Recursive Renewal Reward. In *Queueing Syst.*, pages 153–166, 2013. 1.2.2, 2.2.2

[16] Shlomo Halfin and Ward Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations research*, 29(3):567–588, 1981. 8.3

[17] Mor Harchol-Balter. *Performance modeling and design of computer systems: queueing theory in action*. Cambridge University Press, 2013. 4.1.1

[18] Qi-Ming He and E Jewkes. Flow time in the MAP/G/1 queue with customer batching and setup times. *Stochastic Models*, 11(4):691–711, 1995. 2.1.1

[19] Yige Hong and Ziv Scully. Performance of the gittins policy in the g/g/1 and g/g/k, with and without setup times. *ACM SIGMETRICS Performance Evaluation Review*, 51(2):33–35, 2023. 2.3

[20] Esa Hyytiä, Douglas Down, Pasi Lassila, and Samuli Aalto. Dynamic Control of Running Servers. In *Int. Conf. Measurement, Modelling and Evaluation of Comput. Systems*, pages 127–141, Erlangen, Germany, 2018. Springer. 1.2.2, 2.4

[21] Otis B Jennings, Avishai Mandelbaum, William A Massey, and Ward Whitt. Server staffing to meet time-varying demand. *Management Science*, 42(10):1383–1394, 1996. 8.3, 9.3.2

[22] Aytaç Kara. Energy Consumption in Data Centers with Deterministic Setup Times. Master's thesis, Middle East Technical University, 2017. 1.2.2, 2.4

[23] Ioannis Karatzas and Steven Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 2012. A.1.4

[24] Debankur Mukherjee and Alexander Stolyar. Join Idle Queue with Service Elasticity: Large-Scale Asymptotics of a Nonmonotone System. *Stoch. Syst.*, 9(4):338–358, 2019. 2.2.3

[25] Debankur Mukherjee, Souvik Dhara, Sem C Borst, and Johan SH van Leeuwaarden. Optimal Service Elasticity in Large-Scale Distributed Systems. *Proc. ACM SIGMETRICS Int. Conf. Measurement and Modeling of Computer Systems*, 1:1–28, 2017. 2.2.3

[26] Jerome Niyirora and Jamol Pender. Optimal staffing in nonstationary service centers with constraints. *Naval Research Logistics (NRL)*, 63(8):615–630, 2016. 8.3, 9.3.2

[27] Jamol Pender and Tuan Phung-Duc. A law of large numbers for m/m/c/delayoff-setup queues with nonstationary arrivals. In *Int. Conf. on Analytical and Stochastic Modeling Techniques and Applications*, pages 253–268, Cardiff, UK, 2016. Springer. 1.2.2, 2.2.1, 2.4

[28] Tuan Phung-Duc. Exact solutions for M/M/c/setup queues. *Telecommun. Syst.*, 64(2): 309–324, 2017. 1.2.2, 2.2.2

[29] Irina Shevtsova. On the absolute constants in the Berry-Esseen type inequalities for identically distributed summands, 2011. A.1.3

[30] Peter D Welch. On a Generalized M/G/1 Queuing Process in Which the First Customer of Each Busy Period Receives Exceptional Service. *Oper. Res.*, 12(5):736–752, 1964. 1.1.3, 2.1.1

[31] Jalani Williams, Weina Wang, and Mor Harchol-Balter. Average waiting time in the M/M/k/Setup with Deterministic Setup Times. *Oper. Res.*, 2024, in preparation. 5.1

[32] Jalani K. Williams, Mor Harchol-Balter, and Weina Wang. The M/M/k with Deterministic Setup Times. *Proc. ACM SIGMETRICS Int. Conf. Measurement and Modeling of Computer Systems*, 6(3), dec 2022. doi: 10.1145/3570617. 5.1, 5.2, A.7.1