The M/M/k with Deterministic Setup Times

JALANI K. WILLIAMS, Carnegie Mellon University, USA
MOR HARCHOL-BALTER, Carnegie Mellon University, USA
WEINA WANG, Carnegie Mellon University, USA

Capacity management, whether it involves servers in a data center, or human staff in a call center, or doctors in a hospital, is largely about balancing a resource-delay tradeoff. On the one hand, one would like to turn off servers when not in use (or send home staff that are idle) to save on resources. On the other hand, one wants to avoid the considerable setup time required to turn an "off" server back "on." This paper aims to understand the delay component of this tradeoff, namely, what is the effect of setup time on average delay in a multi-server system?

Surprisingly little is known about the effect of setup times on delay. While there has been some work on studying the M/M/k with Exponentially-distributed setup times, these works provide only iterative methods for computing mean delay, giving little insight as to how delay is affected by k, by load, and by the setup time. Furthermore, setup time in practice is much better modeled by a Deterministic random variable, and, as this paper shows, the scaling effect of a Deterministic setup time is nothing like that of an Exponentially-distributed setup time.

This paper provides the first analysis of the M/M/k with Deterministic setup times. We prove a lower bound on the effect of setup on delay, where our bound is highly accurate for the common case where the setup time is much higher than the job service time. Our result is a relatively simple algebraic formula which provides insights on how delay scales with the input parameters. Our proof uses a combination of renewal theory, martingale arguments and novel probabilistic arguments, providing strong intuition on the transient behavior of a system that turns servers on and off.

CCS Concepts: • **Mathematics of computing → Queueing theory**; **Markov processes**.

Additional Key Words and Phrases: M/M/k/Setup, deterministic setup times, large-system scaling, exceptional first service, capacity provisioning

ACM Reference Format:
Jalani K. Williams, Mor Harchol-Balter, and Weina Wang. 2022. The M/M/k with Deterministic Setup Times. *Proc. ACM Meas. Anal. Comput. Syst.* 6, 3, Article 56 (December 2022), 45 pages. https://doi.org/10.1145/3570617

1 INTRODUCTION

In many queueing systems, servers have both OFF and ON states. A *setup time* is the amount of time needed to transition a server from being OFF to being ON. Setup times occur in a variety of contexts: the transit time for an on-call doctor, the boot time of a computer, the warmup time for a photocopier, to name a few. In all these cases, we turn off servers for good reason; we're making a tradeoff between delay and other things like power consumption, server health, and service quality.

Authors' addresses: Jalani K. Williams, jalaniw@cs.cmu.edu, Carnegie Mellon University, Computer Science Department, 5000 Forbes Ave, Pittsburgh, PA, USA, 15213; Mor Harchol-Balter, harchol@cs.cmu.edu, Carnegie Mellon University, Computer Science Department, 5000 Forbes Ave, Pittsburgh, PA, USA, 15213; Weina Wang, weinaw@cs.cmu.edu, Carnegie Mellon University, Computer Science Department, 5000 Forbes Ave, Pittsburgh, PA, USA, 15213.

2476-1249/2022/12-ART56
https://doi.org/10.1145/3570617

In order to make these decisions intelligently, we need a good understanding of how setup times affect quality of service.

There are two aspects of setup time which we believe require careful modeling, particularly in multi-server systems. The first is the high ratio between setup times and service times. While setup times have always existed in queueing systems, as time has gone on, the ratio between setup times and service times has grown increasingly larger. When combined with dynamic capacity scaling, wherein one dynamically turns off and on servers in response to current queue lengths, this high setup ratio has the potential to wreak havoc on queueing systems. As a thought experiment, consider the behavior of a dynamically-scaled queueing system as the setup ratio goes to infinity: in this case, turning off a server is like getting rid of it entirely, pushing your system into temporary overload. While the limiting case might seem like a bogeyman, the reality of the situation is not so different: in datacenters, virtual machine boot times can be hundreds or even thousands of times larger than the desired customer latency (minutes versus hundreds of milliseconds) [11, 16, 21].

The second important aspect of setup time is its distribution. We argue that many setup times are best modeled as deterministic. Consider, for example, the setup time in an application system where virtual machines are dynamically being booted and shut down as user demand increases and decreases, as in Google's Autopilot [21]. Although the virtual machine (VM) boot procedure includes steps like resource reservation which rely on communication (and thus could potentially be highly variable), in most cloud computing settings, the variation in VM boot times depends predominantly on static aspects of the task at hand, like the operating system image size [16]. In fact, in [11] the setup times for servers in the data center was found to be a constant, 260s, more than a thousand times the typical service requirement for the jobs in that data center.

To see how variability in the setup time might affect queueing behavior, consider another thought experiment, where your web application has gone viral and you need 10 servers as soon as possible to handle the surge in traffic. If setup times are highly variable, then initializing a setup of 100 servers will very quickly net you the 10 servers you need, and you can simply cancel setting up the rest. On the other hand, if setup times take a fixed (deterministic) amount of time, then initializing a setup for 100 servers won't speed up anything. In some sense, models that include higher variability in setup times can make a system seem more "reactive" to surges in traffic than they actually are, especially in models where the number of servers is large.

While we have argued that large setup times, particularly deterministic ones, can have a significant effect on delay in multi-server systems, at present the effect of setup is only fully understood in the single-server setting. We consider the following notion of *delay*: the delay of a job is defined to be the time the job spends waiting in the queue before entering service. For the M/G/1 queue with setup times of any distribution, [22] gives an exact expression for the Laplace transform of delay as a function of the Laplace transform of the setup time, the Laplace transform of the service time, and the arrival rate λ. In the case of an M/M/1 with deterministic setup times $\frac{1}{\alpha}$, the mean delay from [22] is

$$\mathsf{E}\left[\text{delay in M/M/1 with Deterministic setup}\right] = \frac{1}{\mu}\frac{\rho}{1-\rho} + \frac{1}{2\alpha}\quad\frac{2\alpha+\lambda}{\alpha+\lambda}\quad, \tag{1}$$

where λ is the arrival rate of jobs, μ is the service rate of jobs, and $\rho = \frac{\lambda}{\mu}$ is the load of the system.

Existing work on multiserver systems with setup times ([12, 20]) models setup times as following an Exponential distribution, which allows the authors to model the system via a continuous time Markov chain. Unfortunately, these papers only give iterative methods for computing the mean delay. Thus, even for the case of Exponentially-distributed setup time, there are currently no simple closed-form expressions for mean delay that allow us to understand the effect of setup,

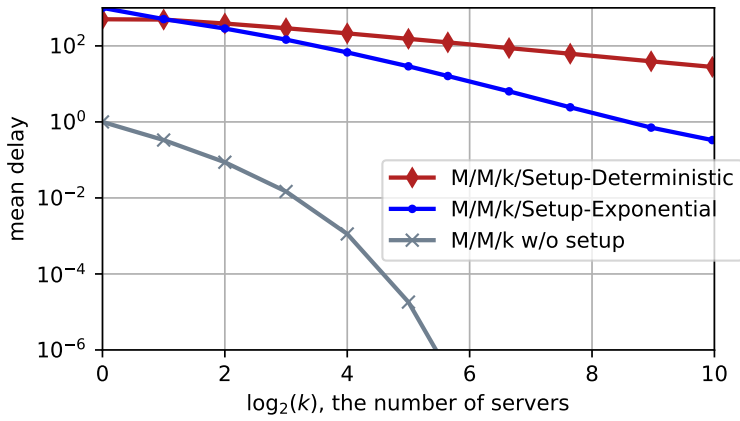

Fig. 1. Simulation results for the M/M/k/Setup-Deterministic, M/M/k/Setup-Exponential, M/M/k (no setup), with $\mu = 1$, setup time $\frac{1}{\alpha} = 1000$, and load kept at a constant $\rho = 0.5$. While there's a huge difference between no-setup and setup, there's also a considerable difference in the shape of the curves for Exponential and Deterministic setups.

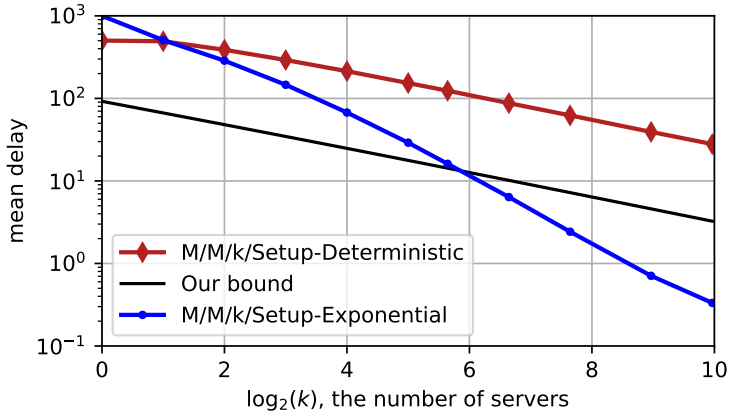

Fig. 2. Comparison of our lower bound for the M/M/k/Setup-Deterministic with simulation results for the M/M/k/Setup-Deterministic and M/M/k/Setup-Exponential, with $\mu = 1$, setup time $\frac{1}{\alpha} = 1000$, and load kept at a constant $\rho = 0.5$. Our bound clearly follows the shape of the M/M/k/Setup-Deterministic. The shape of the M/M/k/Setup-Exponential is quite different.

number of servers, and load, as in (1). Furthermore, in the more practical case where setup times are Deterministic, no analytical results are known for multiserver systems.

To illustrate both the effect of setup time and the effect of Deterministic versus Exponential setup times, we provide some simulations in Figure 1. Here we refer to the M/M/k/Setup with Deterministic (respectively, Exponential) setup times as the M/M/k/Setup-Deterministic (respectively, M/M/k/Setup-Exponential). Roughly, in the M/M/k/Setup-Deterministic, when a server is idle, it is immediately turned off. When a job arrives to find some server(s) off, the job initiates a server

setup. The exact definition of the M/M/k/Setup-Deterministic policy is given in Section 3. In these simulations, the ratio between the mean setup time and mean service time is a constant $\frac{\mu}{\alpha} = 1000$ and load $\rho = 0.5$.

First, we see in Figure 1 that there is a huge difference between the mean delay in systems with setup and the system without setup. This difference grows as the number of servers grows. In particular, given this is a log-log plot, we can see that the mean delay in the systems with setup appears to decay polynomially in k, while we know from Erlang that the delay in the M/M/k without setup decays exponentially in k.

Our second observation in Figure 1 is that the M/M/k/Setup-Deterministic delay curve is much flatter than that of the M/M/k/Setup-Exponential, the latter of which decays much more quickly. Simply put, this tells us that Deterministic setup times are way more painful than Exponentially-distributed ones and must be taken much more seriously in capacity planning. Note that this is in line with our earlier intuition that Exponentially-distributed setup times are more "reactive" (more forgiving) than Deterministic setup times.

Our result. Having seen that Deterministic setup times produce very different results from Exponential setup times, and also very different results from the case of no setup time, we seek to better understand the effect of Deterministic setup times. Our main result is a lower bound on $\mathsf{E}[T_Q]$, the mean delay in the M/M/k with Deterministic setup times, where again delay refers to the time a job spends waiting in the queue before entering service. Our lower bound applies in the case where setup times are much longer than service times. Approximately, it says the following:

Theorem 1 (Informal). *For an M/M/k/Setup with sufficiently large k, sufficiently large Deterministic setup time $\frac{1}{\alpha}$, and some absolute constant C,*

$$\mathsf{E}[T_Q] \geq C \cdot \frac{1}{\alpha} \sqrt{\frac{1}{k\rho}},$$

where one should recall that the arrival rate is $k\lambda$, the service rate is μ, and the load is $\rho = \frac{\lambda}{\mu}$.

This lower bound confirms that the mean delay in an M/M/k/Setup with Deterministic setup time decays, at fastest, *polynomially* in k, consistent with what we observe in Figure 1. Furthermore, we plot this lower bound in Figure 2, where we take a closer look at the difference between the M/M/k/Setup-Exponential and the M/M/k/Setup-Deterministic from Figure 1. We see that our lower bound on the M/M/k/Setup-Deterministic does a very good job of capturing the shape of the M/M/k/Setup-Deterministic, which is noticeably different from the M/M/k/Setup-Exponential. As we hypothesized earlier, the M/M/k/Setup-Exponential has significantly lower delay than the M/M/k/Setup-Deterministic.

Our approach. We now briefly discuss our approach, some technical challenges it raises, and the key ideas which allow us to move beyond those challenges. Recall that our goal is to bound the mean delay, or, equivalently (by Little's Law), to bound the mean number of jobs in queue $\mathsf{E}[Q(\infty)]$. We follow a renewal-reward approach to computing quantities like $\mathsf{E}[Q(\infty)]$: First, we break time into *renewal cycles*, where the system behavior within, say, the first cycle is independently and identically distributed from the system behavior within all other cycles. (For a natural example of such a cycle, consider starting a cycle when the system first becomes empty of jobs.) Once we have chosen how to define a cycle, it follows from renewal theory [2] that

$$\mathsf{E}[Q(\infty)] = \frac{\mathsf{E}\left[\int_{\text{cycle}} Q(t)\mathrm{d}t\right]}{\mathsf{E}[\text{cycle length}]}. \tag{2}$$

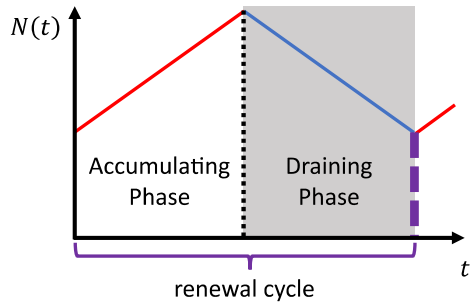

Fig. 3. Idealized depiction of an M/M/k/Setup-Deterministic renewal cycle.

Note that the definition of a cycle is not necessarily unique; in general, there are many definitions of a cycle for which (2) holds, and not all of them are equivalently-easy to reason about.

This gives way to our first technical challenge: defining a cycle which makes the bounding of the numerator and denominator as direct as possible. If, for example, we chose the "empty system" renewal point we discussed before, then we would need to characterize the behavior of the system over a cycle whose mean length is potentially very long (bounded from below by $\approx \frac{e^{k\rho}}{k\lambda}$, via a coupling argument with the M/M/∞).

The key to solving this challenge is to observe the system in simulation. In simulation, a clear two-phase pattern emerges in the behavior of the system, as depicted in Figure 3. In the first phase, the system builds up a large number of queued jobs. In the second phase, once enough servers have turned on, the queue drains in a way reminiscent of a busy period. Upon closer inspection, this two-phase pattern is not a coincidence; the system enters the same state right at the end of every draining phase. As such, we can and do define our cycle as starting with this post-draining state, and ending when the next draining phase ends. This definition turns out to be exactly what we need to make progress.

In some sense, our first challenge is a high-level challenge: we need a good vantage point from which to begin our analysis. In necessary contrast, our second technical challenge is a low-level challenge: in order to actually bound the numerator and denominator in (2), we require a good understanding of the moment-to-moment dynamics of our system. This is *a priori* difficult, since the system's time evolution depends on possibly $k + 1$ pieces of information (the status of each server, plus the number of jobs in queue). Our key low-level insight is that, outside of the precise moments where servers are turning off or on, the behavior of the number of jobs in system $N(t)$ is essentially exactly the behavior of two competing Poisson processes. Combining this insight with martingale theory and the deterministic nature of setup, we have enough to develop our results.

Overview. We give an outline of the remainder of this work.

- In Section 2, we discuss some related work.
- in Section 3, we discuss our model and some preliminary notation.
- In Section 4, we state our main result, a lower bound on the mean delay.
- In Section 5, we go into more depth about our technical approach.
- In Section 6, we prove one of our two main lemmas needed in the main result.
- We leave the rest of the proofs of our results to the Appendices.

2 RELATED WORK

We now discuss some related works which also analyze the effect of setup time in queueing systems. Although we have found no theoretical work analyzing the multi-server systems with deterministic setup times, there is a rich history of work around the analysis of queueing systems with setup.

Single server. The case of setup time in a single server has been understood since the 1960's. Welch, [22], considers a slight generalization of the M/G/1/setup 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. Welch characterizes the steady-state and transient distributions of the queue length and delay. This important result has been extended in a variety of different directions, by adjusting the service discipline or arrival process [3, 4, 13].

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 [8] the authors simplify the solution of the M/M/k with exponential setup times considerably, an prove a decomposition result for mean delay. In [6], the decomposition result is generalized to a hyperexponential job size distribution, and shown to hold approximately for a general job size distribution.

M/M/k/Setup-Exponential, Approximations. Most of the results that deal with an M/M/k/Setup system assume Exponential service times and are approximate. In particular, we highlight the work in [19] and [8]. Gandhi et al. [8] 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 [8] work well, except when both load and setup times are moderately high ($\rho > 0.5$ and $\frac{\mu}{\alpha} > 10$).

Pender and Phung-Duc [19] 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 [19] nor Gandhi et al. [8] 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.

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 [12] and [20]. Gandhi et al. [12] give the first exact analysis of the M/M/k/Setup-Exponential system. To do this, they develop the *Recursive Renewal Reward (RRR)* technique for solving the corresponding Markov chain, algorithmically. Gandhi et al. [12] use RRR to obtain the Laplace transform of delay for any particular value of k, but do not provide a formula as a function of k. Phung-Duc [20] rederives the exact solutions from [12] using generating functions and matrix-analytic methods.

While [12] and [20] are important in that they provide the first exact analysis, their algorithms actually take $O(k^2)$ time to compute. They also do not provide good intuition for the structure of the *solution*, i.e., how the different system parameters (mean setup time, mean service time, arrival rate, number of servers) affect the delay behavior of the system.

Distributed setting. Setup times have also been looked at in distributed systems where a dispatcher routes each incoming job to one of several queues. Mukherjee et al. [18] describe a token-based load

balancing and scaling scheme called TABS that takes into account of setup times on the individual queues. They prove that the performance of TABS (as $k \to \infty$) is asymptotically optimal. While [18] assumes that the queues have finite buffers, Mukherjee and Stolyar [17] generalize their results to infinite buffers. The nature of the questions being asked and answered in [18] and [17] are very different from the central queue-based work we discuss.

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. [14] 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 [15]. Their simulation results corroborate the argument we make in Section 1. In particular, 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.

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 [11]). 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 [11]. 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 [7, 8, 11, 19]. 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 [7]. 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* [11]. The goal of all such algorithms is to minimize the Energy-Response-time-Product (ERP) [7], maximize the Normalized-Performance-Per-Watt (NPPW) [5], or minimize energy given a fixed tail cutoff for response time [11]. 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 [9, 10].

3 MODEL: M/M/k/SETUP-DETERMINISTIC

We now formally describe our model, referred to as M/M/k/Setup-Deterministic, which is a variant on the M/M/k queueing system. An example is illustrated in Figure 4. Just as in the M/M/k model, in our model there are k servers, indexed by $1, 2, \ldots, k$; jobs arrive following a Poisson process of rate $k\lambda$ into a central queue, and job service times follow an exponential distribution with rate μ. The load of the system is denoted as ρ , $\frac{\lambda}{\mu}$, and the quantity R , $k\rho$ is referred to as the offered load of the system following the convention.

To augment the M/M/k with setup times, we make the following adjustments. When a server completes a job and there are no jobs waiting in the queue, the server turns off. Now if we want to turn an off server back on, it requires a setup time. We assume that the setup times are deterministic with value $\frac{1}{\alpha}$. This is in contrast to the regular M/M/k model, where all the servers are on all the time. With setup times, the dynamics of the system can be described as follows.

- **When a job arrives:** (i) If there are off servers, an off server initializes setup. To avoid ambuguity, we assume that in this case the off server with the smallest index initializes setup. In the example

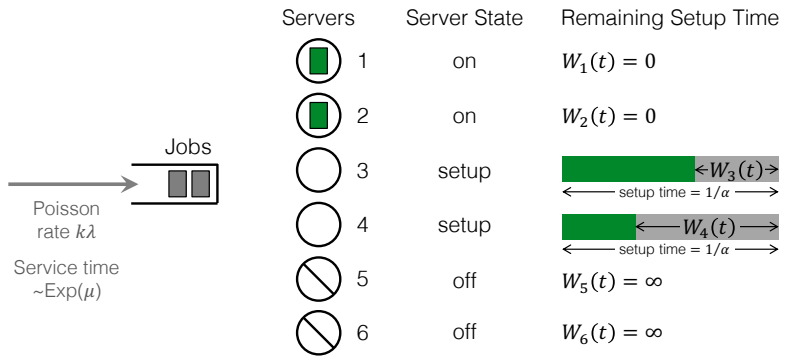

Fig. 4. An example of M/M/k/Setup-Deterministic with $k = 6$. The state pictured has $Z(t) = 2$ busy servers, which means there are 2 jobs in service. There are $Q(t) = 2$ jobs in queue, and thus $N(t) = Z(t) + Q(t) = 4$ jobs in system.

in Figure 4, if a job arrives to the pictured state, server 5 initializes setup. (ii) If all the servers are either on or in setup, their states do not change.

- **When a job is completed on a server:** (i) If the queue is empty, the server turns off. (ii) If the queue is nonempty, the server puts the head-of-queue job into service. Then if the number of jobs left in the queue is smaller than the number of servers in setup, i.e., we are setting up more servers than needed, the setup on the server with the largest remaining setup time is canceled, returning the server back to off. In the example in Figure 4, if server 1 completes the job currently in service, server 1 starts serving the head-of-queue job, the setup on server 4 is canceled, and server 4 returns to off.

Now we explain the notation we use to describe the state of the system at time t. An example is again given in Figure 4. Let $Z(t)$ be the number of on servers (busy servers). Then $Z(t)$ is also the number of jobs in service. Let $Q(t)$ be the number of jobs in queue (not including the jobs in service), and $N(t) = Z(t) + Q(t)$ be the total number of jobs in system. We describe the remaining setup times of servers via a size k vector $\boldsymbol{W}(t)$, where, if server i is in setup, the entry $W_i(t)$ is the remaining amount of time that the i-th server needs to finish setting up before turning ON. If server i is not in setup, we set $W_i(t)$ in the following way for convenience: if server i is on, we set $W_i(t) = 0$; if it is off, we set $W_i(t) = \infty$. Then it is clear that the process

$$\{\mathcal{S}(t) \ , \ (Z(t), Q(t), \boldsymbol{W}(t)) : t \in \mathsf{R}_+\}$$

is a Markov process. Let $s = (z, q, \boldsymbol{w})$ represent a realization of the state. We drop the time index and simply write $\mathcal{S}, Z, Q, N$ and $\boldsymbol{W}$ to represent the corresponding quantities in steady state. We note that by Little's law, the mean number of busy servers in steady state is equal to the offered load, i.e., $\mathsf{E}[Z] = R = k\rho$.

Our goal is to analyze the mean *delay* of jobs, where the delay of a job is defined to be the time the job spends waiting in the queue before entering service. We use T_Q to denote the delay of a job arriving to the steady state of the system. Then by Little's law, if the system is stable, we have

$$\mathsf{E}\ T_Q\ = \frac{\mathsf{E}[Q]}{k\lambda}.$$

Therefore, to bound $\mathsf{E}\ T_Q\ $, it suffices to bound $\mathsf{E}[Q]$. Although our analysis produces bounds of explicit forms, we mainly focus on the asymptotics when the number of servers, k, and the setup time, $\frac{1}{\alpha}$, both become large, while holding other system parameters λ and μ constants.

4 MAIN RESULT

Our main result is a lower bound on the mean delay of M/M/k/Setup-Deterministic. In this section, we first state our lower bound as a scaling result in Theorem 1 to clearly reveal the asymptotic behavior. We then give an explicit form of the lower bound in Theorem 2 to provide more details, which is the form we actually prove in the later sections of the paper. At the end of this section, we include a short proof of Theorem 1 using Theorem 2.

Theorem 1 (Scaling Behavior). *Consider an M/M/k/Setup-Deterministic system with load $\rho = \frac{\lambda}{\mu}$ and setup time $\frac{1}{\alpha}$. We consider the asymptotic regime where k and $\frac{1}{\alpha}$ become large in a way such that $\frac{1/\alpha}{1/\mu} \geq \log^2(k\rho)$, while holding λ and μ constants. Then the expected delay in steady state is lower bounded as*

$$\mathsf{E}[T_Q] = \Omega\left(\frac{1}{\alpha}\sqrt{\frac{1}{k\rho}}\right). \tag{3}$$

As we explained in Section 1, our lower bound shows that in the presence of setup times, the delay is at least on the order of $\frac{1}{\sqrt{k}}$ as $k \to \infty$. This is in sharp contrast to the delay in M/M/k without setup, which decays *exponentially* fast as $k \to \infty$. Moreover, the $\frac{1}{\alpha}$ factor in our bound captures the effect of the setup time $\frac{1}{\alpha}$.

Remark 1. With the lower bound in Theorem 1 alone, we cannot distinguish the asymptotic behavior of M/M/k/Setup-Deterministic from that of M/M/k/Setup-Exponential theoretically. Although the difference is evident in simulations, as noted by Figure 2, we do not have a tight analytical characterization of the delay under M/M/k/Setup-Exponential. We can show that the delay under M/M/k/Setup-Exponential is also on the order of $\frac{1}{\sqrt{k}}$ (with $\text{poly}(\log k)$ modifications), but the dependence on the setup time $\frac{1}{\alpha}$ is unclear.

Our explicit form of the lower bound involves two functions, which we define below. Consider an M/M/1 queue with arrival rate $k\lambda$ and service rate $k\mu$, and consider a busy period started by x jobs in this M/M/1 queue. Then let

$$h(x) = \mathsf{E}[\text{length of busy period started by } x \text{ jobs}] = \frac{x}{k\mu(1-\rho)}, \tag{4}$$

and

$$\begin{aligned} g(x) &= \mathsf{E}\left[\int_{t \in \text{busy period started by } x \text{ jobs}} (\text{number of jobs in system at } t)\,dt\right] \\ &= \frac{x-1}{2} + \frac{1}{1-\rho}\cdot\frac{x}{k\mu(1-\rho)}. \end{aligned} \tag{5}$$

With this notation, our explicit lower bound is presented in Theorem 2 below.

Theorem 2 (Explicit Lower Bound). *Consider an M/M/k/Setup-Deterministic system with load $\rho = \frac{\lambda}{\mu}$ and setup time $\frac{1}{\alpha}$. If the ratio between the setup time and the service time satisfies that $\frac{1/\alpha}{1/\mu} \geq 1000$, the offered load $R = k\rho \geq 128$, and $\frac{1/\alpha}{1/\mu} \geq \log^2(k\rho)$, then the expected delay in steady state is lower bounded as*

$$\mathsf{E}[T_Q] \geq \frac{1}{k\lambda}\cdot\frac{\frac{1}{2}\left(\frac{1}{\alpha}\right)^2\frac{\mu\sqrt{k\rho}}{2} + g\left(\frac{\mu}{\alpha} - 1\right)\frac{\sqrt{k\rho}}{2} - k(1-\rho)\Big)^+}{C_1\left(\frac{3}{\alpha} + \frac{1}{\mu}\right) + \frac{1}{\alpha} + h\left(C_2\frac{\mu\sqrt{k\rho}}{\alpha}\right) + \frac{C_3}{\mu}\log\left(C_2\frac{\mu\sqrt{k\rho}}{\alpha}\right)},$$

where C_1, C_2, and C_3 are constants independent of system parameters k, λ, μ, and α.

Proof of Theorem 1 using Theorem 2.

Proof. We begin with the expression from Theorem 2, converting the numerator and denominator to asymptotic notation. In particular, for the numerator, we drop the g function term because it is nonnegative; for the denominator, we plug in the form of the f function. This gives

$$\mathsf{E}[T_Q] = \Theta\left(\frac{1}{k\lambda}\right) \cdot \frac{\Omega\left(\frac{1}{\alpha}\frac{\mu\sqrt{k\rho}}{\alpha}\right)}{O\left(\frac{1}{\alpha}\right) + O\left(\frac{1}{k\mu(1-\rho)}\frac{\mu\sqrt{k\rho}}{\alpha}\right) + O(\log k) + O\left(\log\frac{1}{\alpha}\right)} \tag{6}$$

$$= \frac{\Omega\left(\frac{1}{\alpha}\frac{\mu}{k\lambda}\frac{\sqrt{k\rho}}{\alpha}\right)}{O\left(\frac{1}{\alpha}\right) + O\left(\frac{1}{\sqrt{k}}\frac{1}{\alpha}\right) + o\left(\frac{1}{\alpha}\right) + o\left(\frac{1}{\alpha}\right)} \tag{7}$$

$$= \Omega\left(\frac{1}{\alpha}\frac{1}{\sqrt{k\rho}}\right), \tag{8}$$

where (7) follows from the assumption that $\frac{1/\alpha}{1/\mu} \geq \log^2(k\rho)$.

With this, the remainder of this paper is devoted to proving Theorem 2.

5 OVERVIEW OF TECHNICAL APPROACH

In this section, we give an overview of our technical approach, which reduces proving Theorem 2 to proving Lemmas 5.1 and 5.2. The proof of Lemma 5.1 is deferred to Appendix A; the proof of Lemma 5.2 follows immediately after this overview, in Section 6.

Recall that the expected delay $\mathsf{E}[T_Q] = \frac{\mathsf{E}[Q]}{k\lambda}$, so it suffices to focus on bounding $\mathsf{E}[Q]$, where Q denotes the number of jobs in the queue in steady state.

Our main idea of the analysis is to decompose time into *renewal* cycles and then express $\mathsf{E}[Q]$ using the Renewal Reward Theorem. This allows us to bound $\mathsf{E}[Q]$ by bounding the expected reward over a cycle and the expected cycle length, respectively.

The key to this approach is to find the right renewal cycles, given the complex dynamics induced by the setup times. We define renewal time points to be times when the number of busy servers, $Z(t)$, decreases from $Z(t^-) = R+1$ to $Z(t) = R$, recalling that $R = k\rho$ is the offered load. Note that the queue is necessarily empty at this transition, i.e., $Q(t^-) = Q(t) = 0$, because otherwise the number of busy servers would not decrease. As a result, there are no servers in setup either at this transition. Without loss of generality, assume that there is a renewal at time 0, and let X be the next renewal point, i.e.,

$$X = \min\{t > 0 : Z(t^-) = R+1, Z(t) = R\}. \tag{9}$$

Then viewing the queue length $Q(t)$ as reward, by the Renewal Reward Theorem, we have

$$\mathsf{E}[Q] = \frac{\mathsf{E}\left[\int_0^X Q(t)dt\right]}{\mathsf{E}[X]}. \tag{10}$$

With this, our lower bound on $\mathsf{E}[T_Q] = \frac{\mathsf{E}[Q]}{k\lambda}$ in Theorem 2 directly follows from the upper bound on $\mathsf{E}[X]$ in Lemma 5.1 and the lower bound on the expected reward $\mathsf{E}\left[\int_0^X Q(t)dt\right]$ in Lemma 5.2.

Before we present the bounds in Lemmas 5.1 and 5.2, we first provide some intuition on the structure of a renewal cycle under our definition, say the time interval $[0, X)$, as illustrated in

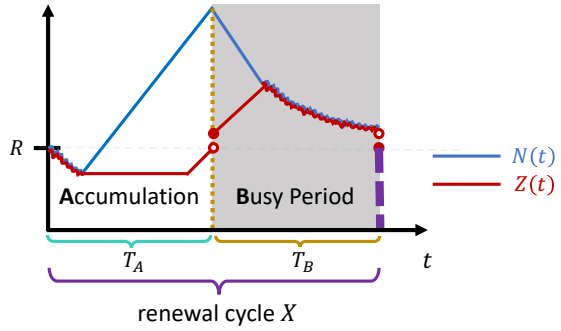

Fig. 5. A visualization of the accumulation and busy period phases which make up a renewal cycle.

Figure 5. By our construction, the renewal cycle $[0, X)$ can be split into two phases, divided by the following time point:

$$T_A = \min\{t > 0 : Z(t) = R + 1\}, \tag{11}$$

i.e., the first time the $(R+1)$-th server turns on. Let $T_B = X - T_A$ be the length of the remaining time in the cycle after T_A. Then T_B can be written as

$$T_B = \min\{s > 0 : Z(T_A + s) = R\}, \tag{12}$$

i.e., the time to the first time after time T_A that the $(R+1)$-th server turns off. With this notation, the cycle $[0, X)$ is split into the following two phases:

- $[0, T_A)$, the *accumulation phase*: During this phase, the system must build up enough jobs in the queue for long enough that the $(R+1)$-th server turns on.
- $[T_A, T_A + T_B)$, the *busy period phase*: During this phase, the system should be (in a net sense) losing jobs over time, until the $(R+1)$-th server turns off.

With this decomposition, we can write $X = T_A + T_B$. Our analyses of $\mathsf{E}[X]$ and $\mathsf{E}\left[\int_0^X Q(t)dt\right]$ are both based on this two-phase structure.

5.1 Upper bound on cycle length $\mathsf{E}[X]$

Our upper bound on $\mathsf{E}[X]$, given in Lemma 5.1 below, makes use of the busy period in an M/M/1 queue with arrival rate $k\lambda$ and service rate $k\mu$. Recall that $h(x) = \frac{x}{k\mu(1-\rho)}$ defined in (4) denotes the expected length of a busy period started by x jobs in this M/M/1 queue.

Lemma 5.1 (Upper Bound on Cycle Length). *Let X be the cycle length defined in (9). Then for three constants C_1, C_2, and C_3 that do not depend on system parameters, we have*

$$\mathsf{E}[X] \le C_1 \left(\frac{3}{\alpha} + \frac{1}{\alpha} + h\left(C_2\frac{\mu\sqrt{R}}{\alpha}\right) + \frac{C_3}{\mu}\log\left(C_2\frac{\mu\sqrt{R}}{\alpha}\right)\right). \tag{13}$$

Recall that $\mathsf{E}[X] = \mathsf{E}[T_A] + \mathsf{E}[T_B]$. The terms in the bound (19) correspond exactly to bounds on $\mathsf{E}[T_A]$ and $\mathsf{E}[T_B]$. We discuss the intuition behind each bound in turn.

Bound on $\mathsf{E}[T_A]$. We first show that

$$\mathsf{E}[T_A] \le C_1\left(\frac{3}{\alpha}\right).$$

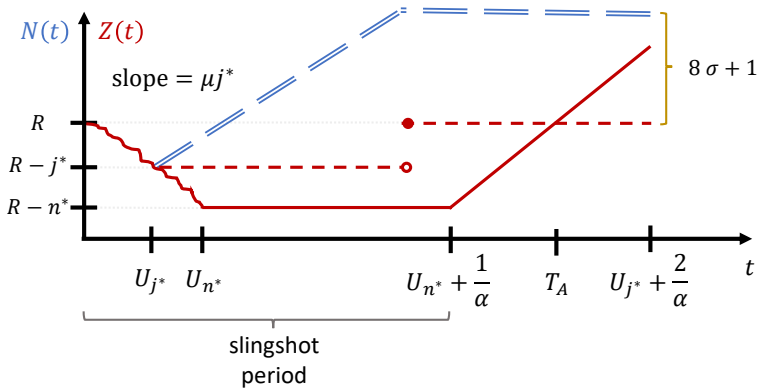

Fig. 6. A visualization of the argument used to bound $\mathsf{E}[T_A]$. The solid red line represents the true value of the number of busy servers $Z(t)$ during some accumulation phase. The dotted red line represents our upper bound on $Z(t)$. The compound blue line represents the expected value of the number of jobs $N(t)$, if we assume the $Z(t)$ takes on the value of the previously mentioned upper bound. Note that our dotted upper bound on $Z(t)$ is valid only until time T_A. Note also that each horizontal line pictured is of length $\frac{1}{\alpha}$.

To bound $\mathsf{E}[T_A]$, we condition on an event $\mathcal{E}$. Before we can define $\mathcal{E}$, we need to define two quantities and two random variables. We begin with the quantities. Let $\sigma \triangleq \sqrt{\frac{k\lambda}{\alpha}}$ represent the standard deviation in the number of job arrivals over $\frac{1}{\alpha}$ time. Let the critical threshold $j^* \triangleq \frac{1}{\frac{1/\alpha}{1/\mu}-1}(8\sigma+1)$. We give more intuition about these quantities later; for now, we define the random variables. Let the hitting time $U_j \triangleq \min\{t > 0 : Z(t) = R - j\}$ be the first time the system has only $R - j$ servers on. Let the slingshot index $n^* \triangleq \min\ j \in Z^+ : U_{j+1} - U_j > \frac{1}{\alpha}$ be the smallest value of j for which the $(j+1)$-th hitting time U_{j+1} occurs more than $\frac{1}{\alpha}$ time away from its predecessor.

We now use these definitions to define the event $\mathcal{E}$. Let $\mathcal{E}$ be the intersection of the following three events:

- Let $\mathcal{E}_1$ be the event that $n^* \geq j^*$. In other words, define $\eta_1 \triangleq U_{n^*} + \frac{1}{\alpha}$ and let
$$\mathcal{E}_1 \triangleq U_{j^*} + \frac{1}{\alpha} \leq \eta_1 .$$
- Let $\mathcal{E}_2$ be the event that, in the $\frac{1}{\alpha}$ time after U_{j^*}, the total number of jobs $N(t)$ reaches $4\sigma + R + 1$. In other words, let $\eta_2 \triangleq \min\{t > \eta_1 : N(t) \geq 4\sigma + R + 1\}$ and define
$$\mathcal{E}_2 \triangleq \eta_2 \leq U_{j^*} + \frac{1}{\alpha} .$$
- Let $\mathcal{E}_3$ be the event that, after reaching $N(\eta_2) = 4\sigma + R + 1$, the total number of jobs $N(t)$ stays above $R + 1$ until the number of busy servers $Z(t)$ is at least $R + 1$. Let $\eta_3 \triangleq \min\{t \geq \eta_2 : Z(t) \geq R + 1\}$. Then we can define $\mathcal{E}_3$ as
$$\mathcal{E}_3 \triangleq \eta_3 \leq \eta_2 + \frac{1}{\alpha} .$$
We explain why this definition suffices. First, note that, if $N(t)$ stays above $R + 1$ for $\frac{1}{\alpha}$ time, then the $(R + 1)$-th server must finish setting up during that time, i.e. the stopping time $\eta_3 \leq \eta_2 + \frac{1}{\alpha}$. Next, note that, if the number of jobs $N(t)$ drops below $R+1$ before $Z(t) \geq R+1$,

then it *must* be the case that $\eta_3 > \eta_2 + \frac{1}{\alpha}$, since it takes at least $\frac{1}{\alpha}$ time before we can set up another server.

We illustrate some of these critical time instants in Figure 6.

With the event $\mathcal{E}$ defined as such, it follows that

$$\mathsf{E}[T_A|\mathcal{E}] \leq \mathsf{E}[\eta_3|\mathcal{E}] \leq \mathsf{E}[\eta_2|\mathcal{E}] + \frac{1}{\alpha} \leq \mathsf{E}[\eta_1|\mathcal{E}] + \frac{1}{\alpha}.$$

This implies

$$\begin{aligned}\mathsf{E}[T_A] &= \Pr(\mathcal{E})\,\mathsf{E}[T_A|\mathcal{E}] + \Pr(\mathcal{E}^c)\,\mathsf{E}[T_A|\mathcal{E}^c] \\ &\leq \Pr(\mathcal{E})\ \ \mathsf{E}[\eta_1|\mathcal{E}] + \frac{1}{\alpha}\ \ + \Pr(\mathcal{E}^c)\,\mathsf{E}[T_A|\mathcal{E}^c] \\ &\leq \mathsf{E}[\eta_1] + \Pr(\mathcal{E})\,\frac{1}{\alpha} + \Pr(\mathcal{E}^c)\,\mathsf{E}[T_A|\mathcal{E}^c]\,.\end{aligned}$$

We spend the rest of this section describing how we upper bound the probability $\Pr(\mathcal{E}^c)$, upper bound the expectation $\mathsf{E}[\eta_1]$, and upper bound the conditional expectation $\mathsf{E}[T_A|\mathcal{E}^c]$. Our upper bound on $\mathsf{E}[T_A|\mathcal{E}^c]$ will be larger than $\frac{1}{\alpha}$, so this will suffice to prove the claim.

Upper bound on $\Pr(\mathcal{E}^c)$. We argue here that the probability $\Pr(\mathcal{E}^c)$ is small, i.e. that the events of $\mathcal{E}$ are likely. We do so by giving an intuitive explanation of the early dynamics of our system. We begin by noting three facts:

- First, jobs arrive at rate $k\lambda$ and depart at rate $\mu Z(t)$.
- Second, whenever the number of busy servers $Z(t)$ decreases, the number of jobs in queue $Q(t)$ must be 0, since we only turn off servers when there are no jobs available for them to work on.
- Third, if at time t' the number of busy servers $Z(t') = R - j$ and the queue is empty, then, for at least $\frac{1}{\alpha}$ time afterwards, $Z(t)$ will be at most $R - j$, since no additional servers could possibly set up during that time.

It follows from these facts that, in a sense, the system behaves like a slingshot: if the number of busy servers decreases to $R - j$, then, for the next $\frac{1}{\alpha}$ time, the system will accumulate jobs at rate $k\lambda - \mu Z(t) \geq \mu j$. If the number of turned off servers j was large enough, then, after this period, the system will likely have a large number of jobs queued. If there are a large number of jobs queued and the departure rate $\mu Z(t)$ does not exceed the arrival rate $k\lambda$ (i.e. we are in the accumulation phase), then it's unlikely that the queue will empty before we turn the $(R+1)$-th server on. We visualize this chain of events in Figure 6. Nicely, because the events are based on the stopping times U_j, η_1, η_2 and η_3, each event $\mathcal{E}_i$ concerns the behavior of the system for $\frac{1}{\alpha}$ after a state-determined stopping time. This state information is precisely what makes it possible to bound each of $\Pr\ \mathcal{E}_i^c$; applying a union bound, we gain a bound for $\Pr(\mathcal{E}^c)$.

Bound on $\mathsf{E}[\eta_1]$. To bound $\mathsf{E}[\eta_1]$, we first describe η_1 as the length of a sequence of stopped random walks, then bound the expected length of each step in the sequence. Recall the definitions of the hitting time $U_j = \min\{t > 0 : Z(t) = R - j\}$, the slingshot index $n^* = \min\ \ j \in \mathsf{W} : U_{j+1} - U_j > \frac{1}{\alpha}$, and the stopping time $\eta_1 = U_{n^*} + \frac{1}{\alpha}$. One can describe η_1 as the length of the following process, which we call the slingshot period. Starting from time $U_0 = 0$, after each hitting time U_j, we give the system $\frac{1}{\alpha}$ time to reach the next hitting time U_{j+1}. If the system succeeds, then the slingshot period continues; if not, the period ends. We first observe that, during the j-th step in this process, the behavior of the number of jobs $N(t)$ is exactly the behavior of a biased random walk driven by

a Poisson process, i.e. the time between each arrival/departure event is i.i.d. $\text{Exp}(k\lambda + \mu(R-j))$ and is an arrival with probability $p = \frac{R}{2R-j}$ and a departure with probability $1-p = \frac{R-j}{2R-j}$.

After making this observation, we use it to bound the expected contribution of each step in the slingshot period. Let $V_j(t)$, with $V_j(0) = 0$, be the value of the Poisson-driven random walk in the previous paragraph and let $\tau_j = \min\ \ t > 0 : V_j(t) = -1\ \ $ be the first passage time of the continuous-time random walk $V_j(t)$. We first develop an upper bound on the tail of τ_j (Lemma B.2), then use that tail bound to bound both the probability that the j-th step occurs and the expected contribution of the j-th step given that it does occur. Summing the expected contribution from each step together, we find that, for some constant C_4 independent of system parameters,

$$\mathsf{E}[\eta_1] \leq \frac{1}{\alpha} + C_4 \sum_{j=1}^{\mathsf{s}} \frac{1}{\alpha\mu} \leq \frac{2}{\alpha}.$$

Upper bound on $\mathsf{E}[T_A|\mathcal{E}^c]$. To bound the conditional expectation $\mathsf{E}[T_A|\mathcal{E}^c]$, we break $T_A \mid \mathcal{E}^c$ down into two terms, a stopping time $\psi \mid \mathcal{E}^c$ and a remainder $(T_A - \psi) \mid \mathcal{E}^c$. In particular, event $\mathcal{E}$ describes a set of allowed initial system trajectories: all trajectories where $Z(t)$ reaches $R - j^*$ quickly, the number of jobs in queue $Q(t)$ crosses some threshold, and the number of jobs stays above $R+1$. Let the stopping time ψ be the first moment that the system deviates from that set of allowed trajectories. With ψ defined, we have

$$\mathsf{E}[T_A|\mathcal{E}^c] = \mathsf{E}[\psi|\mathcal{E}^c] + \mathsf{E}[T_A - \psi|\mathcal{E}^c].$$

We now discuss how to bound $\mathsf{E}[\psi|\mathcal{E}^c]$, then give some brief insight into how we bound the remainder $\mathsf{E}[T_A - \psi|\mathcal{E}^c]$. First, we argue that, if the event $\mathcal{E}^c$ occurs, then $\psi < \frac{3}{\alpha}$. This follows from the fact that the length of every allowed trajectory is at most $\frac{3}{\alpha}$. In other words, if the first deviation from the set of allowed trajectories has not occurred by $\frac{3}{\alpha}$, then it will never occur, i.e. the event $\mathcal{E}^c$ is first. This implies that $\psi \mid \mathcal{E}^c < \frac{3}{\alpha}$.

To bound the remainder $\mathsf{E}[T_A - \psi|\mathcal{E}^c]$, we use a worst-case bound of the remainder over every possible state $\mathcal{S}(\psi)$. Let S_ψ denote the set of all possible exit states $\mathcal{S}(\psi)$. Using the Markovian structure of the system, we obtain

$$\begin{aligned}
\mathsf{E}[T_A - \psi|\mathcal{E}^c] &= \tilde{\sum_{s \in S_\psi}} \Pr(\mathcal{S}(\psi) = s|\mathcal{E}^c)\, \mathsf{E}[T_A - \psi|\mathcal{E}^c \cap \mathcal{S}(\psi) = s] \\
&= \tilde{\sum_{s \in S_\psi}} \Pr(\mathcal{S}(\psi) = s|\mathcal{E}^c)\, \mathsf{E}[T_A - \psi|\mathcal{S}(\psi) = s] \\
&\leq \max_{s \in S_\psi} \mathsf{E}[T_A - \psi|\mathcal{S}(\psi) = s].
\end{aligned}$$

We relax things further. Let S_A be the set of all states consistent with being in the accumulation phase, i.e. all states s such that $Z(t) \leq R$. To upper bound the conditional expectation of $\mathsf{E}[T_A - \psi|\mathcal{S}(\psi) = s]$ for every state in S_ψ, we derive an upper bound on $\mathsf{E}[T_A - t|\mathcal{S}(t) = s]$ that applies to every state in S_A.

The full derivation of this bound is of course deferred to Appendix A; however, we explain here the rough strategy, visualized in Figure 7. The key idea is to argue about the dynamics of the system using only the total number of jobs $N(t)$. First, we split all states in S_A into 3 different buckets based on their value of $N(t)$, with each number of jobs being classified as high, medium, or low. In particular, we call a number of jobs $N(t)$ low if it is strictly less than R, call it high if that number of jobs is above some threshold $R + M$ (whose value we discuss later), and medium otherwise. Then, for each bucket of states B, we give a bound on the worst-case expected remaining time

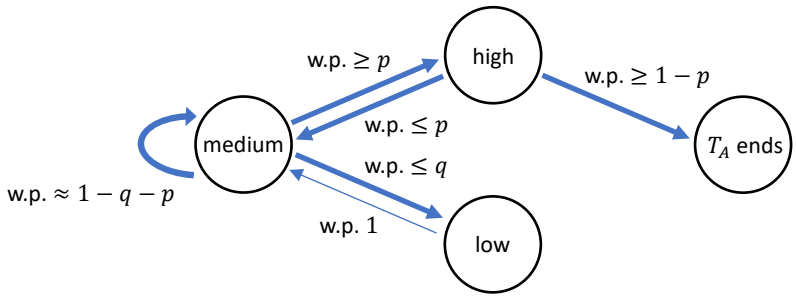

Fig. 7. A depiction of our method for bounding $\mathsf{E}[T_A - \ell | \mathcal{S}(\ell) = s]$. We partition a state's number of jobs $N(t)$ into three bins: low, medium, and high. We say that a state has a medium number of jobs if $N(t)$ lies in some range $[R, R+M)$, a low number of jobs if $N(t) < R$, and a high number of jobs if $N(t) \geq R+M$. Each circle represents a bin of states (low, medium, or high), and each arrow represents a possible transition between states, which occurs with the probability written alongside it. The single thin arrow represents a step which takes at most $\approx \frac{1}{\mu}$ time, in expectation. The remaining thick arrows represent steps which take at most $\frac{1}{\alpha}$ time.

$\max_{s \in B} \mathsf{E}[T_A - t | \mathcal{S}(t) = s]$ using the worst-case bounds of the *other* buckets. Roughly, it suffices to show the following claims:

- In expectation, a low state becomes a medium state in $\approx \frac{1}{\mu}$ time.
- With some probability p depending on M, a system beginning in a medium state enters a high state within $\frac{1}{\alpha}$ time.
- With probability at least $1-p$, a system beginning in a high state completes the accumulation phase within $\frac{1}{\alpha}$ time.

By setting M appropriately, we guarantee that the probability p is not too small and also not too large (i.e., that $1-p$ is also not too small) and, by completing some casework, we arrive at a bound on $\mathsf{E}[T_A - t | \mathcal{S}(t) = s]$ which holds for any state s that might occur in the accumulation phase. This implies a bound $\mathsf{E}[T_A | \mathcal{E}^c]$ and thus $\mathsf{E}[T_A]$, completing the proof.

Bound on $\mathsf{E}[T_B]$. The right two terms of Lemma 5.1 correspond to an upper bound on $\mathsf{E}[T_B]$:

$$\mathsf{E}[T_B] \leq \frac{1}{\alpha} + h\left(C_2 \frac{\mu\sqrt{R}}{\alpha}\right) + \frac{C_3}{\mu} \log\left(C_2 \frac{\mu\sqrt{R}}{\alpha}\right)$$

We prove this bound in two steps. First, we give a concave function $f(y)$ such that $f(y) \geq \mathsf{E}[T_B | Q(T_A) = y]$, then use Jensen's inequality to show

$$\begin{aligned} \mathsf{E}[T_B] &= \mathsf{E}[\mathsf{E}[T_B | Q(T_A)]] \\ &\leq \mathsf{E}[f(Q(T_A))] \\ &\leq f(\mathsf{E}[Q(T_A)]). \end{aligned}$$

Second, we show that $\mathsf{E}[Q(T_A)] \leq C_2 \frac{\mu\sqrt{R}}{\alpha}$. We describe each step in more detail below.

Bound on $\mathsf{E}[T_B | Q(T_A)]$. We set out to show

$$\mathsf{E}[T_B | Q(T_A) = y] \leq \frac{1}{\alpha} + \frac{y}{\mu k(1-\rho)} + \frac{C_3}{\mu}(\log(y) + 1) \;\; , \;\; f(y).$$

We prove this bound via a stronger state-specific result. Before stating that result, we give some necessary definitions and assumptions. Let $S_B = \{s = (z, q, \boldsymbol{w}) : z \geq R + 1\}$ be the set of all possible states in the draining phase; these are the states we are bounding over. We consider the system at some time $\ell < T_B$ in some state $\mathcal{S}(\ell) = s = (z, q, \boldsymbol{w})$. Let $v \in \mathsf{Z}_+ \cup 0$ be the unique index such that the number of jobs in the system $n = q + z$ lies somewhere in $[2^v + R, 2^{v+1} + R)$. Note that such an index must exist, since the total number of jobs in the system n must be at least the number of jobs being served $z \geq R + 1$. Let $z_h \;,\; \min(2^v + R, k)$ be some "hoped for" number of busy servers, and recall that w_{z_h} is the remaining amount of setup time on the z_h-th server.

The stronger result is this: For any such state $\mathcal{S}(\ell) = s \in S_B$,

$$\mathsf{E}[T_B - \ell | \mathcal{S}(\ell) = s] \leq w_{z_h} + \frac{n - z_h}{\mu(z_h - R)} + \frac{1}{\mu}(v + 1).$$

We prove this state-dependent result via induction on $v \;,\; \lfloor \log n - R \rfloor$. The base case is very simple: in the case where $Z(\ell) = R + 1$ and $Q(\ell) = 0$, the remaining time in the system $T_B - \ell$ must be stochastically dominated by the length of an M/M/1 busy period with arrival rate $k\lambda$ and departure rate $\mu(R + 1) = k\lambda + \mu$. In the inductive case, we observe the system for $W_{z_h}(\ell)$ time and condition on whether $N(t)$ ever dips below $2^v + R$. The moment that it does, we can invoke our inductive hypothesis. If the number of jobs $N(t)$ never dips, then $N(t)$ has stayed above z_h for enough time that we can be sure at least z_h servers are on. Intuitively, having a large number of servers should help us drain the queue very quickly. Unfortunately, if $N(t)$ never dips below $2^v + R$, then the number of jobs in our system $N(t)$ is likely large. We use the observation that $N(t)$ is a supermartingale to bound the number of jobs in queue $Q(t)$ in this bad case. To complete the result, we couple the original system at time $\ell + w_{z_h}$ to an OFF system, a system that begins in the same state but can only turn servers off.

Bound on $\mathsf{E}[Q(T_A)]$. We now discuss the bound on the expected number of jobs in queue at the end of the accumulation phase, $Q(T_A)$. The proof has two steps. In the first step, we show that

$$\mathsf{E}[Q(T_A)] = O\left(\mathsf{E}[R - Z^*] + \sigma\right),$$

where Z^* is the minimal number of servers that are on during the accumulation phase and the variation $\sigma = \sqrt{\frac{k\lambda}{\alpha}}$. In the second step, we upper bound $\mathsf{E}[R - Z^*]$.

To upper bound $\mathsf{E}[Q(T_A)]$ using $\mathsf{E}[R - Z^*]$, we condition on whether the number of jobs $N(t)$ exceeded $R + 4\sigma + 1$ near the end of the accumulation phase. If not, then clearly we can upper bound $Q(T_A)$ by 4σ. If $N(t)$ *did* exceed that amount, we argue that the expected value of $N(T_A)$ is the conditional expectation of some stopped random walk, and that we can bound this conditional expectation by making a martingale argument. Informally,

$$\mathsf{E}[Q(T_A) | N(t) \text{ gets large}] \leq \frac{1}{p} \;\; 4\sigma + \frac{1}{\alpha}[\text{maximal upward drift experienced by } N(t)] \;\;,$$

where the p is a probability near 1 and the maximal upward drift is just $\mu(R - Z^*)$.

To bound $\mathsf{E}[R - Z^*]$, we hearken back to our bound on T_A. In particular, we need to lower bound the probability that, if we have just reached $Z(t) = R - j$ servers, then the accumulation phase ends without the number of busy servers $Z(t)$ dropping lower than $R - j$. To do this, we argue that, if the number of turned off servers j is sufficiently large ($\approx \sqrt{R}$), then the probability we follow our previously discussed path (up to $N(t) = 4\sigma + R + 1$ and onward) *without* turning off a server is at least $\approx \frac{1}{\sqrt{R}}$. By analogy to a Geometric random variable, it follows that

$$\mathsf{E}[R - Z^*] = O\left(\sqrt{R}\right) \;\;,$$

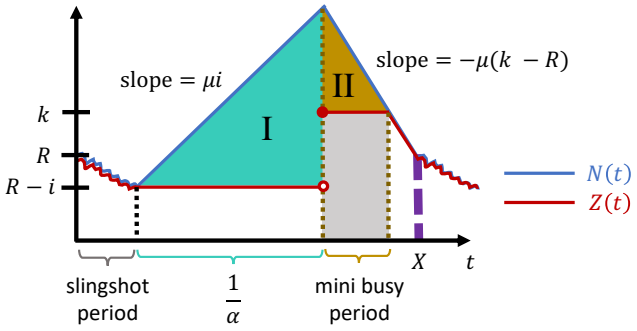

Fig. 8. A visualization of our approach to bounding $\mathsf{E}\left[\int_0^X Q(t)\mathrm{d}t\right]$.

which, since $\sigma = \sqrt{\frac{\mu R}{\alpha}} \ll \frac{\mu\sqrt{R}}{\alpha}$, implies

$$\mathsf{E}\left[Q(T_A)\right] \leq C_2 \frac{\mu\sqrt{R}}{\alpha}.$$

Bound on $\mathsf{E}[X]$. This allows us to complete our bound on the length of the draining phase $\mathsf{E}[T_B]$, and in turn our bound on the cycle length $\mathsf{E}[X]$. Recalling that $h(x) = \frac{x}{\mu k(1-\rho)}$ and using our previous results $\mathsf{E}[T_B|Q(T_A)]$ and $\mathsf{E}[Q(T_A)]$, we obtain

$$\mathsf{E}[T_B] \leq \frac{1}{\alpha} + h\left(C_2\frac{\mu\sqrt{R}}{\alpha}\right) + \frac{C_3}{\mu}\log\left(C_2\frac{\mu\sqrt{R}}{\alpha}\right),$$

as desired. Combining our results on $\mathsf{E}[T_A]$ and $\mathsf{E}[T_B]$, we obtain Lemma 5.1,

$$\mathsf{E}[X] \leq C_1\left(\frac{3}{\alpha}\right) + \frac{1}{\alpha} + h\left(C_2\frac{\mu\sqrt{R}}{\alpha}\right) + \frac{C_3}{\mu}\log\left(C_2\frac{\mu\sqrt{R}}{\alpha}\right).$$

5.2 Lower bound on expected reward $\mathsf{E}\left[\int_0^X Q(t)dt\right]$

Our lower bound on $\mathsf{E}\left[\int_0^X Q(t)dt\right]$ also makes use of the M/M/1 queue with arrival rate $k\lambda$ and service rate $k\mu$. Recall that $g(x) = \frac{x-1}{2} + \frac{1}{1-\rho}\frac{x}{k\mu(1-\rho)}$ denotes the expected time integral of the number of jobs in system (in queue plus in service) during a busy period started by x jobs in this M/M/1 queue. Then our lower bound is given as Lemma 5.2.

Lemma 5.2 (Lower Bound on Expected Reward).

$$\mathsf{E}\left[\int_0^X Q(t)\,dt\right] \geq \frac{1}{2}\left(\frac{1}{\alpha}\right)^2\frac{\mu\sqrt{R}}{2} + g\left(\frac{\mu}{\alpha} - 1\right)\left(\frac{\sqrt{R}}{2} - k(1-\rho)\right)^+.$$

We illustrate our reward bounds in Figure 8. The first term in Lemma 5.2 corresponds to the area of Triangle I before the slingshot period plus $\frac{1}{\alpha}$, and the second term corresponds to the area of Triangle II in the mini busy period. The initial dynamics of our system are like a slingshot, as discussed in the previous subsection. During this period, servers are being turned off. As we turn off servers, the upward drift in our number of jobs $N(t)$ gets higher. Suppose we turn off $n^* = i$ servers before we turn on servers in this initial period. Then when we start turning on servers, the

number of jobs $N(t)$ will, for at least $\frac{1}{\alpha}$ time, drift upward at a rate of μi. The cumulative reward up to the slingshot period plus $\frac{1}{\alpha}$ then is lower bounded by the area of Triangle I, which is $\approx \frac{1}{2}\left(\frac{1}{\alpha}\right)^2 \mu i$.

Moreover, once we reach the time point that is the slingshot period plus $\frac{1}{\alpha}$, we have accumulated roughly $\mu i \cdot \frac{1}{\alpha} + R - i = \left(\frac{\mu}{\alpha} - 1\right) i + R$ jobs, with at most $\left(\frac{\mu}{\alpha} - 1\right) i + R - k$ jobs in the queue. The fastest way we can get rid of these jobs is if all k servers are working. Therefore, we consider an M/M/1 queue with arrival rate $k\lambda$ and service rate $k\mu$. Then the expected reward afterwards can be lower bounded by the time integral of the number of jobs in the M/M/1 queue during a busy period started by $\left(\frac{\mu}{\alpha} - 1\right) i + R - k$ jobs. This lower bound corresponds to the area of Triangle II.

In our proof, we will lower bound the expected number of servers that get turned off during the slingshot period by relating the number of jobs in queue to a continuous-time random walk.

We note here the scaling of the terms mentioned. As k gets large, the dominating term will be the first term, since $\frac{\mu}{\alpha}\sqrt{R} \le k(1-\rho)$ for sufficiently large k, making the last term go to 0.

6 LOWER BOUND ON EXPECTED REWARD

In this section, we prove the lower bound given in Lemma 5.2 on the expected reward $\mathsf{E}\left[\int_0^{X} Q(t)\mathrm{d}t\right]$. Recall that the renewal cycle starts with $Z(0^-) = R+1$ and $Z(0) = R$. Our proof relies on the following hitting times that we defined earlier. For convenience, let $U_0 = 0$, and we think of U_0 as the first time for $Z(t)$ to reach R. Recall that we have defined

$$U_j = \min\left\{t > U_{j-1} \colon Z(t) = R - j\right\}, j = 1, 2, \ldots, R.$$

I.e., U_j is the first time that the number of busy servers reaches $R-j$. Note that by this construction, we necessarily have that $Z(U_j^-) = R - j + 1, Z(U_j) = R - j$ and $Q(U_j) = 0$. We have also defined

$$n^* = \min\left\{j \in \mathsf{Z}_+ \colon U_{j+1} - U_j > \frac{1}{\alpha}\right\}. \tag{14}$$

Let $n^* = R$ when $U_{j+1} - U_j \le \frac{1}{\alpha}$ for all $j = 0, 1, \ldots, R-1$ for convenience. Note that $U_{n^*} + \frac{1}{\alpha}$ is then a stopping time with respect to the process of the system state $\{\boldsymbol{S}(t), (Z(t), Q(t), \boldsymbol{W}(t)) \colon t \in \mathsf{R}_+\}$. Based on these definitions, our proof of Lemma 5.2 consists of proving the following three claims, where Claim 6.1 is used in proving Claims 6.2 and 6.3. Lemma 5.2 then follows immediately by taking a sum of the lower bounds in Claims 6.2 and 6.3.

Claim 6.1. *Consider the n^* in (14) and recall that $R = k\rho$. Then*

$$\mathsf{E}[n^*] \ge \frac{1}{2}\sqrt{R}.$$

Claim 6.2. *The expected cumulative reward up to time $U_{n^*} + \frac{1}{\alpha}$ is lower bounded as*

$$\mathsf{E}\left[\int_0^{U_{n^*}+\frac{1}{\alpha}} Q(t)\,dt\right] \ge \frac{1}{2}\left(\frac{1}{\alpha}\right)^2 \frac{\mu\sqrt{R}}{2}.$$

Claim 6.3. *The expected cumulative reward after time $U_{n^*} + \frac{1}{\alpha}$ is lower bounded as*

$$\mathsf{E}\left[\int_{U_{n^*}+\frac{1}{\alpha}}^{X} Q(t)\,dt\right] \ge g\left(\left(\frac{\mu}{\alpha} - 1\right)\frac{\sqrt{R}}{2} - k(1-\rho)\right)^+,$$

where recall that $g(x) = \frac{x-1}{2} + \frac{1}{1-\rho}\frac{x}{k\mu(1-\rho)}$ is the expected time integral of the number of jobs during a busy period of the M/M/1 queue with arrival rate $k\lambda$ and service rate $k\mu$.

We prove these three claims in the remainder of this section.

6.1 Proof of Claim 6.1

Proof. We observe that $n^* \geq i$ is equivalent to $U_{j+1} - U_j \leq \frac{1}{\alpha}$ for all $j = 0, 1, \ldots, i-1$. Consider any $j = 0, 1, \ldots, i-1$. Since $Z(U_j) = n - j$, $Q(U_j) = 0$, and no servers can finish setting up within a time duration of $\frac{1}{\alpha}$, we can couple the queueing dynamics $Q(t)$ for $t \in \left[U_j, U_j + \frac{1}{\alpha}\right]$ with a random walk defined by two independent Poisson processes: (1) $\{Y_a(t) : t \in \mathbb{R}_+\}$ with rate $k\lambda = \mu R$, and (2) $\{Y_d(t) : t \in \mathbb{R}_+\}$ with rate $k\mu(R-j)$. Let a random walk $\tilde{Q}_j(t)$ start from $\tilde{Q}_j(0) = 0$ and be defined as $\tilde{Q}_j(t) = Y_a(t) - Y_d(t)$. Let $\tau = \min\{t > 0 : \tilde{Q}_j(t) < 0\}$. Then we can couple the arrival and departure processes of our system during the time interval $\left[U_j, U_j + \min\left\{\frac{1}{\alpha}, \tau\right\}\right]$ with $\{Y_a(t) : t \in \mathbb{R}_+\}$ and $\{Y_d(t) : t \in \mathbb{R}_+\}$ during $\left[0, \min\left\{\frac{1}{\alpha}, \tau\right\}\right]$, respectively. As a result,

$$\Pr\left\{U_{j+1} - U_j \leq \frac{1}{\alpha}\right\} = \Pr\left\{\tau \leq \frac{1}{\alpha}\right\} \geq \frac{R-j}{R} e^{-\gamma},$$

where $\gamma = -\frac{1}{2}\ln(1 - e^{-4}) \approx 0.009$, and the inequality follows from Lemma B.2 in Appendix B. Noting that $U_{j+1} - U_j$ for $j = 0, 1, \ldots, R-1$ are independent, we get

$$\Pr\left(n^* \geq i\right) \geq e^{-\gamma i} \prod_{j=0}^{i-1} \frac{R-j}{R}.$$

Continuing from the tail sum formula for expectation, we have

$$\begin{aligned}
\mathsf{E}[n^*] &= \sum_{i=1}^{\infty} \Pr\left(n^* \geq i\right) \\
&\geq \sum_{i=1}^{\infty} e^{-\gamma i} \prod_{j=0}^{i-1} \frac{R-j}{R} \\
&= \sum_{i=1}^{\infty} e^{-\gamma i} \frac{R!}{(R-i)!R^i} \\
&\geq \sum_{i=1}^{R-1} e^{-\gamma i} \left(1 + \frac{i}{R-i}\right)^{R-i+\frac{1}{2}} e^{-i-\frac{1}{12}} + e^{-\gamma R} \frac{R!}{R^R} \qquad (15) \\
&\geq e^{-\frac{1}{12}} \sum_{i=1}^{R-1} e^{-\gamma i} e^{-\frac{i^2}{R}}, \qquad (16)
\end{aligned}$$

where (15) is obtained by applying Stirling's lower and upper bounds to $R!$ and $(R-i)!$, respectively, for $i = 1, 2, \ldots, R-1$; and (16) follows from the inequality $\left(1 + \frac{x}{y}\right)^y \geq e^{\frac{yx}{x+y}}$ and some simple bounding. Notice that the form of the sum in (16) is similar to that of a Gaussian integral. We can indeed lower-bound it by making use of the integral, thus obtaining the final bound as desired. We refer the readers to Appendix C for the bounding and calculation details.

6.2 Proof of Claim 6.2

Proof. We first write the expected cumulative reward up to time $U_{n^*} + \frac{1}{\alpha}$ as

$$\begin{aligned}
&\mathsf{E}\left[\int_0^{U_{n^*}+\frac{1}{\alpha}} Q(t)\mathrm{d}t\right] \\
&= \sum_{i=0}^{\infty} \mathsf{E}\left[\int_0^{U_i+\frac{1}{\alpha}} Q(t)\mathrm{d}t \;\middle|\; n^* = i\right] \cdot \Pr\left(n^* = i\right)
\end{aligned}$$

$$
\begin{aligned}
&\geq \sum_{i=0}^{R-1} \mathbb{E}\left[\int_{U_i}^{U_i+\frac{1}{\alpha}} Q(t)\mathrm{d}t \;\middle|\; n^* = i\right] \cdot \Pr\left(n^* = i\right) \\
&= \sum_{i=0}^{R-1} \mathbb{E}\left[\int_{U_i}^{U_i+\frac{1}{\alpha}} Q(t)\mathrm{d}t \;\middle|\; U_1 - U_0 \leq \frac{1}{\alpha}, \ldots, U_i - U_{i-1} \leq \frac{1}{\alpha}, U_{i+1} - U_i > \frac{1}{\alpha}\right] \cdot \Pr\left(n^* = i\right) \\
&= \sum_{i=0}^{R-1} \mathbb{E}\left[\int_{U_i}^{U_i+\frac{1}{\alpha}} Q(t)\mathrm{d}t \;\middle|\; U_{i+1} - U_i > \frac{1}{\alpha}\right] \cdot \Pr\left(n^* = i\right) \qquad (17)
\end{aligned}
$$

To bound $\mathbb{E}\left[\int_{U_i}^{U_i+\frac{1}{\alpha}} Q(t)\mathrm{d}t \;\middle|\; n^* = i\right]$ for each $i = 0, 1, \ldots, R-1$, we consider the same coupling between $\{Q(t) : t \in [U_i, U_i + \frac{1}{\alpha}]\}$ and $\{\tilde{Q}_i(t) : t \in [0, \frac{1}{\alpha}]\}$ as that in the proof of Claim 6.1. It is not hard to see that

$$
\{Q(t) : t \in [U_i, U_i + \frac{1}{\alpha}] \mid U_{i+1} - U_i > \frac{1}{\alpha}\} \overset{d}{=} \{\tilde{Q}_i(t) : t \in [0, \frac{1}{\alpha}] \mid \tilde{Q}_i(t) \geq 0 \text{ for all } t \in [0, \frac{1}{\alpha}]\}.
$$

Therefore, by Lemma B.3 in Appendix B,

$$
\begin{aligned}
\mathbb{E}\left[\int_{U_i}^{U_i+\frac{1}{\alpha}} Q(t)\mathrm{d}t \;\middle|\; U_{i+1} - U_i > \frac{1}{\alpha}\right] &= \mathbb{E}\left[\int_{0}^{\frac{1}{\alpha}} \tilde{Q}_i(t)\mathrm{d}t \;\middle|\; \tilde{Q}_i(t) \geq 0 \text{ for all } t \in [0, \frac{1}{\alpha}]\right] \\
&\geq \frac{1}{2}\left(\frac{1}{\alpha}\right)^2 \mu i.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\mathbb{E}\left[\int_{0}^{U_{n^*}+\frac{1}{\alpha}} Q(t)\mathrm{d}t\right] &\geq \sum_{i=0}^{R-1} \frac{1}{2}\left(\frac{1}{\alpha}\right)^2 \mu i \cdot \Pr\left(n^* = i\right) \\
&= \frac{1}{2}\left(\frac{1}{\alpha}\right)^2 \mu \mathbb{E}\left[n^*\right] \\
&\geq \frac{1}{2}\left(\frac{1}{\alpha}\right)^2 \frac{\mu\sqrt{R}}{2},
\end{aligned}
$$

where the last inequality follows from Claim 6.1. This completes the proof.

6.3 Proof of Claim 6.3

Proof. We first write the expected cumulative reward after time $U_{n^*} + \frac{1}{\alpha}$ as

$$
\begin{aligned}
&\mathbb{E}\left[\int_{0}^{U_{n^*}+\frac{1}{\alpha}} X\, Q(t)\mathrm{d}t\right] \\
&= \sum_{i=0}^{R-1} \mathbb{E}\left[\int_{0}^{U_i+\frac{1}{\alpha}} X\, Q(t)\mathrm{d}t \;\middle|\; n^* = i\right] \cdot \Pr\left(n^* = i\right) \\
&= \sum_{i=0}^{R-1} \mathbb{E}\left[\mathbb{E}\left[\int_{U_i}^{U_i+\frac{1}{\alpha}} X\, Q(t)\mathrm{d}t \;\middle|\; \mathcal{S}\left[U_i + \frac{1}{\alpha}\right], n^* = i\right] \;\middle|\; n^* = i\right] \cdot \Pr\left(n^* = i\right) \\
&= \sum_{i=0}^{R-1} \mathbb{E}\left[\mathbb{E}\left[\int_{U_i}^{U_i+\frac{1}{\alpha}} X\, Q(t)\mathrm{d}t \;\middle|\; \mathcal{S}\left[U_i + \frac{1}{\alpha}\right], U_1 - U_0 \leq \frac{1}{\alpha}, \ldots, U_i - U_{i-1} \leq \frac{1}{\alpha}, U_{i+1} - U_i > \frac{1}{\alpha}\right] \;\middle|\; n^* = i\right] \\
&\quad \cdot \Pr\left(n^* = i\right).
\end{aligned}
$$

Note that $U_{i+1} - U_i > \frac{1}{\alpha}$ is equivalent to $Z(t) \geq n - i$ for all $t \in \left[U_i, U_i + \frac{1}{\alpha}\right)$. Therefore, since $U_i + \frac{1}{\alpha}$ is a stopping time, given the state $\mathcal{S}\left(U_i + \frac{1}{\alpha}\right)$, the queue length $Q(t)$ for $t \in \left[U_i + \frac{1}{\alpha}, X\right)$ is independent from $U_1 - U_0 \leq \frac{1}{\alpha}, \ldots, U_i - U_{i-1} \leq \frac{1}{\alpha}, U_{i+1} - U_i > \frac{1}{\alpha}$. Thus,

$$\mathbb{E}\left[\int_{U_{n^*}+\frac{1}{\alpha}}^{X} Q(t)\mathrm{d}t\right] = \sum_{i=0}^{k} \mathbb{E}\left[\mathbb{E}\left[\int_{U_i+\frac{1}{\alpha}}^{X} Q(t)\mathrm{d}t \;\middle|\; \mathcal{S}\left(U_i + \frac{1}{\alpha}\right), n^* = i\right] \cdot \Pr\left(n^* = i\right)\right]. \tag{18}$$

We now construct a coupling to lower-bound $\mathbb{E}\left[\int_{U_i+\frac{1}{\alpha}}^{X} Q(t)\mathrm{d}t \;\middle|\; \mathcal{S}\left(U_i + \frac{1}{\alpha}\right)\right]$ for each $i = 0, 1, \ldots, R$. Note that the number of busy servers in the system is always smaller than or equal to k. So we can construct an M/M/1 queue $\tilde{Q}(t)$ with arrival rate $k\lambda$, service rate $k\mu$, and initial state $\tilde{Q}(0) = \left[N\left(U_i + \frac{1}{\alpha}\right) - k\right]^+ \leq Q\left(U_i + \frac{1}{\alpha}\right)$, where the arrival process is coupled with the arrival process in our system and the service process dominates the service process in our system. Let $U^* := \min\{t \geq 0 : \tilde{Q}(t) = 0\}$ be the end of the initial busy period. Then given $\mathcal{S}\left(U_i + \frac{1}{\alpha}\right)$, the queue length $Q(t)$ in our system for each $t \in \left[U_i + \frac{1}{\alpha}, U_i + \frac{1}{\alpha} + U^*\right)$ is lower bounded by $\tilde{Q}\left(t - U_i - \frac{1}{\alpha}\right)$, and it can be verified that $U_i + \frac{1}{\alpha} + U^* \leq X$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\int_{U_i+\frac{1}{\alpha}}^{X} Q(t)\mathrm{d}t \;\middle|\; \mathcal{S}\left(U_i + \frac{1}{\alpha}\right)\right] &\geq \mathbb{E}\left[\int_{0}^{U^*} \tilde{Q}(t)\mathrm{d}t \;\middle|\; \tilde{Q}(0) = \left[N\left(U_i + \frac{1}{\alpha}\right) - k\right]^+\right] \\ &= g\left(\left[N\left(U_i + \frac{1}{\alpha}\right) - k\right]^+\right).\end{aligned}$$

Inserting the bound above back to (18) gives

$$\mathbb{E}\left[\int_{U_{n^*}+\frac{1}{\alpha}}^{X} Q(t)\mathrm{d}t\right] = \sum_{i=0}^{k} \mathbb{E}\left[g\left(\left[N\left(U_i + \frac{1}{\alpha}\right) - k\right]^+\right) \middle| n^* = i\right] \cdot \Pr\left(n^* = i\right).$$

Consider the same coupling as that in the proof of Claim 6.2 for $t \in \left[U_i, U_i + \frac{1}{\alpha}\right)$. By Lemma B.3 in Appendix B, we have that $\mathbb{E}\left[N\left(U_i + \frac{1}{\alpha}\right) \middle| n^* = i\right] = \frac{\mu}{\alpha} i + R - i$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\int_{U_{n^*}+\frac{1}{\alpha}}^{X} Q(t)\mathrm{d}t\right] &\geq \sum_{i=0}^{k} g\left(\left[\frac{\mu}{\alpha} - 1\right) i - k(1-\rho)\right]^+\right) \cdot \Pr\left(n^* = i\right) \\ &= \mathbb{E}\left[g\left(\left[\left(\frac{\mu}{\alpha} - 1\right) n^* - k(1-\rho)\right]^+\right)\right] \\ &\geq g\left(\left[\left(\frac{\mu}{\alpha} - 1\right) \mathbb{E}\left[n^*\right] - k(1-\rho)\right]^+\right) \\ &= g\left(\left[\left(\frac{\mu}{\alpha} - 1\right) \frac{\sqrt{R}}{2} - k(1-\rho)\right]^+\right),\end{aligned}$$

where both inequalities follow from the convexity of the function $g([\cdot]^+)$, and the last equality follows from Claim 6.1. This completes the proof.

7 CONCLUSION AND FUTURE WORK

Summary. This paper is the first to analyze multiserver systems (M/M/k) with setup times, where the setup time is Deterministic (M/M/k/Setup-Deterministic). We derive a lower bound on the mean delay (waiting time) in the M/M/k/Setup-Deterministic, showing that the mean delay scales as $1/\sqrt{k}$, with the number of servers k.

Impact. Our work has three main takeaways. First, we show that, when modeling real systems with large setup times, it is imperative that one models the effect of setup times. Our lower bound proves that, when considering mean delay (waiting time), systems with setup exhibit a fundamentally different scaling behavior than systems without setup (visualized in Figure 1). In particular, the ratio of the delay of systems with setup and those without setup grows exponentially in the number of servers, k. Second, we demonstrate in simulation that the scaling behavior of setup systems depends on the setup distribution. When setup times are large compared to service times, modeling setup times as Exponential random variables (as was done in nearly all previous theoretical work) tends to *severely underestimate* the detrimental effect of setup times on delay. Third, our theoretical analysis suggests that the prevailing wisdom of "be very careful when turning off servers" is well-founded. In our analysis, most of the waiting happens during the "accumulation phase," which is a period where so many servers have been turned off that we enter a period of transient instability, which doesn't end until the Deterministic setup time completes. This motivates the need to consider better policies for dealing with Deterministic setup times.

Future work. One natural direction of future work is creating an upper bound on delay to match our existing lower bound. While our lower bound is important in that it gives us a lower bound on the needed capacity, we need the upper bound to tell us whether our lower bound is close to tight. Another direction of future work is the analysis of more sophisticated setup policies, that avoid shutting off servers too aggressively just to turn them on again later. Several such policies are discussed at the end of Section 2. Many of these policies are interested in both minimizing mean delay and also minimizing power consumption. Unfortunately, these policies have only been analyzed under Exponentially-distributed setup times. It would be interesting to study how their performance is affected by Deterministic setup times.

8 ACKNOWLEDGMENTS

This research was supported by NSF CMMI-1938909, NSF CSR-1763701, NSF CNS-200773, and NSF ECCS-2145713, in addition to the Gates Millennium and GEM Fellowships.

REFERENCES

[1] J. R. Artalejo, A. Economou, and M. J. Lopez-Herrero. Analysis of a Multiserver Queue with Setup Times. *Queueing Syst.*, 51(1):53–76, 2005.
[2] S. Asmussen. *Applied Probability and Queues*, volume 2. Springer, 2003.
[3] W. Bischof. Analysis of M/G/1-Queues with Setup Times and Vacations under Six Different Service Disciplines. *Queueing Syst.*, 39(4):265–301, 2001.
[4] G. Choudhury. On a batch arrival Poisson queue with a random setup time and vacation period. *Comp. & Oper. Res.*, 25(12):1013–1026, 1998.
[5] A. Gandhi and M. 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.
[6] A. Gandhi and M. Harchol-Balter. M/G/k with staggered setup. *Oper. Res. Lett.*, 41(4):317–320, 2013.
[7] A. Gandhi, V. Gupta, M. Harchol-Balter, and M. 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.
[8] A. Gandhi, M. Harchol-Balter, and I. Adan. Server farms with setup costs. *Performance Evaluation*, 67(11):1123–1138, 2010.
[9] A. Gandhi, M. Harchol-Balter, and M. 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.
[10] A. Gandhi, M. Harchol-Balter, and M. Kozuch. Are sleep states effective in data centers? In *Int. Conf. Green Computing (IGCC)*, pages 1–10, San Jose, CA, 2012.
[11] A. Gandhi, M. Harchol-Balter, R. Raghunathan, and M. A. Kozuch. AutoScale: Dynamic, Robust Capacity Management for Multi-Tier Data Centers. *ACM Trans. Comput. Syst.*, 30(4):1–26, 2012.

[12] A. Gandhi, S. Doroudi, M. Harchol-Balter, and A. Scheller-Wolf. Exact analysis of the M/M/k/setup class of Markov chains via Recursive Renewal Reward. In *Queueing Syst.*, pages 153–166, 2013.
[13] Q.-M. 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.
[14] E. Hyytiä, D. Down, P. Lassila, and S. Aalto. Dynamic Control of Running Servers. In *Int. Conf. Measurement, Modelling and Evaluation of Comput. Systems*, pages 127–141, Erlangen, Germany, 2018. Springer.
[15] A. Kara. Energy Consumption in Data Centers with Deterministic Setup Times. Master's thesis, Middle East Technical University, 2017.
[16] M. Mao and M. Humphrey. A Performance Study on the VM Startup Time in the Cloud. In *IEEE Int. Conf. Cloud Computing (CLOUD)*, pages 423–430, Honolulu, HI, 2012.
[17] D. Mukherjee and A. Stolyar. Join Idle Queue with Service Elasticity: Large-Scale Asymptotics of a Nonmonotone System. *Stoch. Syst.*, 9(4):338–358, 2019.
[18] D. Mukherjee, S. Dhara, S. C. Borst, and J. S. 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.
[19] J. Pender and T. 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.
[20] T. Phung-Duc. Exact solutions for M/M/c/setup queues. *Telecommun. Syst.*, 64(2):309–324, 2017.
[21] K. Rzadca, P. Findeisen, J. Swiderski, P. Zych, P. Broniek, J. Kusmierek, P. Nowak, B. Strack, P. Witusowski, S. Hand, et al. Autopilot: workload autoscaling at Google. In *Proc. European Conf. Computer Systems (EuroSys)*, pages 1–16, Heraklion, Crete, Greece, 2020.
[22] P. 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.

A UPPER BOUND ON CYCLE LENGTH

In this section, we prove our upper bound on the expected cycle length $\mathsf{E}[X]$ in Lemma 5.1, which we restate as Lemma A.1 below for ease of reference. Recall that, at time 0, the system is in the state $\mathcal{S}(0)$ where $Z(0) = N(0) = R$. Recall also that the cycle length $X = \min\{t > 0 : Z(t^-) = R+1, Z(t) = R\}$ is the amount of time until we next turn off the $(R+1)$-th server.

Lemma A.1 (Upper Bound on Cycle Length). *Let X be the cycle length defined in (9). Then for three constants C_1, C_2, and C_3 that do not depend on system parameters, we have*

$$\mathsf{E}[X] \le C_1 \left(\frac{3}{\alpha} + \frac{1}{\alpha} + h\left(C_2 \frac{\mu\sqrt{R}}{\alpha} \right) + \frac{C_3}{\mu} \log\left(C_2 \frac{\mu\sqrt{R}}{\alpha} \right) \right). \tag{19}$$

As discussed, we split our analysis into two claims bounding $\mathsf{E}[T_A]$ and $\mathsf{E}[T_B]$. We spend the remainder of this section proving these claims.

Claim A.1. *For any $\frac{1/\alpha}{1/\mu} > 1000$ and $R > 128$,*

$$\mathsf{E}[T_A] \le \left(1 - 7e^{-4}\right)\left(\frac{3}{\alpha}\right) + 7e^{-4}\left(\frac{10}{\alpha} + \frac{10}{\mu}\right).$$

Claim A.2. *Let $h(x) = \frac{x}{k\mu(1-\rho)}$ be the expected length of a busy period started by x jobs in an M/M/1 with arrival rate $k\lambda$ and departure rate $k\mu$. Let j^* be the smallest index such that $\frac{\mu}{\alpha} j^* \ge 8\sigma + j^* + 1$, where $\sigma = \sqrt{\frac{k\lambda}{\alpha}}$.*

$$\mathsf{E}[T_B] \le \frac{1}{\alpha} + \frac{2}{(1-e^{-4})^2} \left[\frac{4\sigma + 1 + 3\frac{\mu\sqrt{R}}{\alpha}}{k\mu(1-\rho)} + \frac{2}{\mu} \log\left(\frac{1}{1-e^{-4}} \left(4\sigma + 1 + 3\frac{\mu\sqrt{R}}{\alpha} \right) \right) \right].$$

A.1 Proof of Claim A.1

We begin by showing the upper bound on $\mathsf{E}[T_A]$. We repeat the approach given in Section 5.

To bound $\mathsf{E}[T_A]$, we condition on an event $\mathcal{E}$. Before we can define $\mathcal{E}$, we need to define two quantities and two random variables. We begin with the quantities. Let $\sigma \triangleq \sqrt{\frac{k\lambda}{\alpha}}$ represent the standard deviation in the number of job arrivals over $\frac{1}{\alpha}$ time. Let the critical threshold $j^* \triangleq \frac{1}{\frac{1/\alpha}{1/\mu}-1}(8\sigma+1)$. We give more intuition about these quantities later; for now, we define the random variables. Let the hitting time $U_j \triangleq \min\{t > 0 : Z(t) = R - j\}$ be the first time the system has only $R-j$ servers on. Let the slingshot index $n^* \triangleq \min\ j \in \mathbb{Z}^+ : U_{j+1} - U_j > \frac{1}{\alpha}$ be the smallest value of j for which the $(j+1)$-th hitting time U_{j+1} occurs more than $\frac{1}{\alpha}$ time away from its predecessor.

We now use these definitions to define the event $\mathcal{E}$. Let $\mathcal{E}$ be the intersection of the following three events:

- Let $\mathcal{E}_1$ be the event that $n^* \geq j^*$. In other words, define $\eta_1 \triangleq U_{n^*} + \frac{1}{\alpha}$ and let
$$\mathcal{E}_1 \triangleq U_{j^*} + \frac{1}{\alpha} \leq \eta_1 .$$
- Let $\mathcal{E}_2$ be the event that, in the $\frac{1}{\alpha}$ time after U_{j^*}, the total number of jobs $N(t)$ reaches $4\sigma + R + 1$. In other words, let $\eta_2 \triangleq \min\{t > \eta_1 : N(t) \geq 4\sigma + R + 1\}$ and define
$$\mathcal{E}_2 \triangleq \eta_2 \leq U_{j^*} + \frac{1}{\alpha} .$$
- Let $\mathcal{E}_3$ be the event that, after reaching $N(\eta_2) = 4\sigma + R + 1$, the total number of jobs $N(t)$ stays above $R+1$ until the number of busy servers $Z(t)$ is at least $R+1$. Let $\eta_3 \triangleq \min\{t \geq \eta_2 : Z(t) \geq R+1\}$. Then we can define $\mathcal{E}_3$ as
$$\mathcal{E}_3 \triangleq \eta_3 \leq \eta_2 + \frac{1}{\alpha} .$$
We explain why this definition suffices. First, note that, if $N(t)$ stays above $R+1$ for $\frac{1}{\alpha}$ time, then the $(R+1)$-th server must finish setting up during that time, i.e. the stopping time $\eta_3 \leq \eta_2 + \frac{1}{\alpha}$. Next, note that, if the number of jobs $N(t)$ drops below $R+1$ before $Z(t) \geq R+1$, then it *must* be the case that $\eta_3 > \eta_2 + \frac{1}{\alpha}$, since it takes at least $\frac{1}{\alpha}$ time before we can set up another server.

We illustrate some of these critical time instants in Figure 6.

With the event $\mathcal{E}$ defined as such, it follows that
$$\mathsf{E}[T_A|\mathcal{E}] \leq \mathsf{E}[\eta_3|\mathcal{E}] \leq \mathsf{E}[\eta_2|\mathcal{E}] + \frac{1}{\alpha} \leq \mathsf{E}[\eta_1|\mathcal{E}] + \frac{1}{\alpha}.$$
This implies
$$\begin{aligned}\mathsf{E}[T_A] &= \Pr(\mathcal{E})\,\mathsf{E}[T_A|\mathcal{E}] + \Pr(\mathcal{E}^c)\,\mathsf{E}[T_A|\mathcal{E}^c] \\ &\leq \Pr(\mathcal{E})\ \mathsf{E}[\eta_1|\mathcal{E}] + \frac{1}{\alpha}\ + \Pr(\mathcal{E}^c)\,\mathsf{E}[T_A|\mathcal{E}^c] \\ &\leq \mathsf{E}[\eta_1] + \Pr(\mathcal{E})\,\frac{1}{\alpha} + \Pr(\mathcal{E}^c)\,\mathsf{E}[T_A|\mathcal{E}^c] .\end{aligned}$$
We spend the rest of this section proving upper bounds on the probability $\Pr(\mathcal{E}^c)$, the expectation $\mathsf{E}[\eta_1]$, and the conditional expectation $\mathsf{E}[T_A|\mathcal{E}^c]$. Our upper bound on $\mathsf{E}[T_A|\mathcal{E}^c]$ will be larger than $\frac{1}{\alpha}$, so this will suffice to prove the claim. In particular, we prove the following claims.

Claim A.3. *For the event $\mathcal{E}$ described above,*
$$\Pr(\mathcal{E}^c) = \Pr\ \mathcal{E}_1^c \cup \mathcal{E}_2^c \cup \mathcal{E}_3^c\ \leq 7e^{-4}.$$

Claim A.4. *Recall the definitions of the hitting time* $U_j = \min\{t > 0 : Z(t) = R - j\}$, *the slingshot index* $n^* = \min\ j \in \mathbb{W} : U_{j+1} - U_j > \frac{1}{\alpha}$, *and the stopping time* $\eta_1 = U_{n^*} + \frac{1}{\alpha}$. *Then the expected duration of the slingshot period* η_1 *can be bounded as*

$$\mathsf{E}[\eta_1] \le \frac{1}{\alpha} + C_4 \frac{1}{\sqrt{\alpha\mu}} \le \frac{2}{\alpha}.$$

Claim A.5. *For the event* $\mathcal{E}$ *outlined above,*

$$\mathsf{E}[T_A | \mathcal{E}^c] \le 5\ \frac{2}{\alpha} + \frac{1.01}{\mu}\ .$$

Clearly, these claims suffice to prove Claim A.1. We prove each in turn.

A.1.1 Proof of Claim A.3. To prove Claim A.3, we obtain upper bounds of $\approx e^{-4}$ for each probability $\Pr\ \mathcal{E}_i^c$, then apply a union bound.

Bound of $\Pr\ \mathcal{E}_1^c$. We begin by bounding the probability of event $\mathcal{E}_1$. Before we begin manipulating the probability $\Pr(\mathcal{E}_1) = \Pr(n^* \ge j^*)$, we describe the behavior of the system in a useful way. Recall the definitions of the hitting time $U_j = \min\{t > 0 : Z(t) = R - j\}$ and the slingshot index $n^* = \min\ j \in \mathbb{Z}^+ : U_{j+1} - U_j > \frac{1}{\alpha}$.

We begin by arguing that the slingshot period is actually a sequence of stopped random walks in continuous time. First, we observe that each step j in this process ends at a well-defined stopping time $\min\ U_{j+1}, U_j + \frac{1}{\alpha}$. Next, we show that the departure rate is static during each step. As mentioned in the previous paragraph, since the queue is empty at every hitting time U_j and each step has a maximum length of $\frac{1}{\alpha}$, the departure rate of the system can not increase during a step. Furthermore, because no additional servers are set up during a step, every time we turn off a server must be the *first* time we turn off that server. Since a step ends when a server is turned off, it follows that, during each step, the number of busy servers $Z(t)$ (and thus the departure rate) stays the same. Since the arrival rate is also fixed at $k\lambda$, during each step, the behavior of the number of jobs $N(t)$ during step j is exactly the behavior of a biased discrete random walk driven by a Poisson process, i.e. the time between each job arrival/departure is distributed $\mathrm{Exp}(2k\lambda - \mu j)$ and is an arrival with probability $p = \frac{R}{2R-j}$ and a departure with probability $q = 1 - p$.

We now use this observation to prove our bound. Let τ_j be the 1-to-0 first passage time in the Poisson-driven random walk, like that discussed in the previous paragraph. Note that, due to the Markovian structure of our system, the distribution of τ_j (and, in fact, $U_{j+1} - U_j$) is independent of all other τ_i's (and $(U_{i+1} - U_i)$'s, respectively) for $i < j$. Applying this knowledge, along with Lemma B.2, gives

$$\begin{aligned}
\Pr(n^* \ge j^*) &= \Pr\ \cap_{i=0}^{j^*-1}\ U_{i+1} - U_i \le \frac{1}{\alpha} \\
&= \Pr\ \cap_{i=0}^{j^*-1}\ U_{i+1} - U_i \le \frac{1}{\alpha} \\
&= \prod_{i=0}^{j^*-1} \Pr\ U_{i+1} - U_i \le \frac{1}{\alpha} \\
&= \prod_{i=0}^{j^*-1} \Pr\ \tau_i \le \frac{1}{\alpha} \\
&\ge \prod_{i=0}^{j^*-1} e^{-\frac{7}{\sigma}}\ \frac{R-i}{R}
\end{aligned}$$

$$= e^{-\frac{7}{\sigma}j^*} \frac{1}{R^{j^*}} \frac{R!}{(R-j^*)!}.$$

Applying Stirling's approximation,

$$\begin{aligned}
\frac{1}{R^{j^*}} \frac{R!}{(R-j^*)!} &\geq \frac{1}{R^{j^*}} e^{-\frac{1}{12}} \left(\frac{R}{e}\right)^R \left(\frac{e}{R-j^*}\right)^{R-j^*} \sqrt{\frac{R}{R-j^*}} \\
&\geq e^{-\frac{1}{12}} \left(1+\frac{j^*}{R-j^*}\right)^{R-j^*+\frac{1}{2}} e^{-j^*} \\
&\geq e^{-\frac{1}{12}} \left(1+\frac{j^*}{R-j^*}\right)^{R-j^*} e^{-j^*} \\
&\geq e^{-\frac{1}{12}} e^{j^* - \frac{(j^*)^2}{R}} e^{-j^*} \qquad (20) \\
&= e^{-\frac{1}{12} - \frac{(j^*)^2}{R}},
\end{aligned}$$

where we have made use of the inequality $\left(1+\frac{x}{y}\right)^y \geq e^{\frac{xy}{x+y}}$. Note that $j^* = \frac{8\sigma+1}{\frac{\mu}{\alpha}-1} \leq \frac{1000}{998}(8.01)\sqrt{\frac{\alpha}{\mu}}\sqrt{R}$, thus,

$$\begin{aligned}
\Pr(n^* \geq j^*) &\geq e^{-\frac{3.5}{\sigma}j^*} e^{-\frac{1}{12} - \frac{(j^*)^2}{R}} \\
&\geq e^{-93\frac{\mu}{\alpha}} \geq e^{-0.093},
\end{aligned}$$

which implies that

$$\Pr\left(\mathcal{E}_1^c\right) \leq 5e^{-4}.$$

Bound of $\Pr\left(\mathcal{E}_2^c\right)$. Recall that $\mathcal{E}_2$ is the event that, in the $\frac{1}{\alpha}$ time after U_{j^*}, the total number of jobs $N(t)$ reaches $4\sigma + R + 1$. We bound $\Pr\left(\mathcal{E}_2^c\right)$ by constructing a coupled system with a smaller number of jobs $\tilde{N}(t)$, then proving the result for that system using martingale techniques.

We begin by constructing the coupled system $\tilde{N}(t)$. Recall that, within $\frac{1}{\alpha}$ of any hitting time U_j, the departure rate $\mu Z(t)$ can only decrease. Let $Y_a(t)$ be a Poisson process of rate $k\lambda$ and $Y_d(t)$ be a Poisson process of rate $\mu(R-j^*)$. If we consider the system $\tilde{N}(t) = R - j + Y_a(t - U_{j^*}) - Y_d(t - U_{j^*})$, then, by choosing the natural coupling, the departure rate of the original system is always greater than the departure rate of the coupled system. Using the natural coupling, we maintain that the number of jobs $N(t) \geq \tilde{N}(t)$ for $t \in \left[U_{j^*}, U_{j^*} + \frac{1}{\alpha}\right]$. As such, it suffices to show the result for the coupled system.

To show the result for the coupled system, we use Doob's submartingale inequality. First, note that the coupled number of jobs $\tilde{N}(t) = R - j + Y_a(t - U_{j^*}) - Y_d(t - U_{j^*})$ is a submartingale, since Poisson processes have independent increments, and the departure rate is smaller than the arrival rate. Applying Doob's submartingale inequality,

$$\begin{aligned}
\Pr\left(\sup_{t\in[U_{j^*}, U_{j^*}+\frac{1}{\alpha}]} N(t) \leq 4\sigma + R + 1\right) &\leq \Pr\left(\sup_{t\in[U_{j^*}, U_{j^*}+\frac{1}{\alpha}]} \tilde{N}(t) \leq 4\sigma + R + 1\right) \\
&= \Pr\left(\sup_{t\in[0,\frac{1}{\alpha}]} Y_a(t) - Y_d(t) \leq 4\sigma + j^* + 1\right) \\
&\leq \Pr\left(Y_a\left(\frac{1}{\alpha}\right) - Y_d\left(\frac{1}{\alpha}\right) \leq 4\sigma + j + 1\right)
\end{aligned}$$

$$= \Pr\left[\exp\left(-\theta\left(Y_a\left(\frac{1}{\alpha}\right) - Y_d\left(\frac{1}{\alpha}\right)\right)\right) \geq \exp\left(-\theta(4\sigma + j + 1)\right)\right]$$

For brevity, we let $Y_a^* = Y_a\left(\frac{1}{\alpha}\right)$ and $Y_d^* = Y_d\left(\frac{1}{\alpha}\right)$. Recall that, for any $Y \sim \text{Poisson}(\nu)$ and any real θ, the moment generating function $\mathbb{E}\left[e^{\theta Y}\right] = \exp\left(-\nu + \nu e^{\theta}\right)$, that $\mathbb{E}\left[Y_a^*\right] = \frac{k\lambda}{\alpha} = \sigma^2$, and that $\mathbb{E}\left[Y_d^* - Y_a^*\right] = -\frac{\mu}{\alpha} j^* = -(8\sigma + j^* + 1)$. Letting $\epsilon = \frac{2}{\sigma}$ and $\theta = \ln(1+\epsilon) \leq \epsilon$, and noting that Y_d^* and Y_a^* are independent,

$$\begin{aligned}
&\Pr\left[\exp\left(-\theta\left(Y_a\left(\frac{1}{\alpha}\right) - Y_d\left(\frac{1}{\alpha}\right)\right)\right) \geq \exp\left(-\theta(4\sigma + j + 1)\right)\right] \\
&\leq \frac{\mathbb{E}\left[e^{-\theta(Y_a^*)}\right]\mathbb{E}\left[e^{\theta(Y_d^*)}\right]}{e^{-\theta(4\sigma + j^* + 1)}} \\
&\leq \exp\left[-\sigma^2(1 - e^{-\theta}) - (\sigma^2 - \frac{\mu j^*}{\alpha})(1 - e^{\theta}) + \theta(4\sigma + j^* + 1)\right] \\
&\leq \exp\left[-\sigma^2(\frac{\epsilon}{1+\epsilon}) - (\sigma^2 + \frac{\mu j^*}{\alpha})(-\epsilon) + \epsilon(4\sigma + j^* + 1)\right] \\
&= \exp\left[\sigma^2 \frac{\epsilon^2}{1+\epsilon} - \epsilon(4\sigma)\right] \leq \exp\left[\sigma^2\epsilon^2 - \epsilon(4\sigma)\right] \\
&\leq \exp(-4),
\end{aligned}$$

as desired.

Bound on $\Pr\left[\mathcal{E}_3^c\right]$. Recall the definition of $\eta_2 = \min\{t > \eta_1 : N(t) = 4\sigma + R + 1\}$ as the first moment after the slingshot period that $N(t)$ reaches $4\sigma + R + 1$ and $\eta_3 = \min\{t > \eta_2 : Z(t) \geq R + 1\}$ as the first moment after η_2 that occurs after the accumulation phase. To bound the probability of the event $\mathcal{E}_3 = \left[\eta_3 \leq \eta_2 + \frac{1}{\alpha}\right]$, we again make an argument based on a coupling to a martingale. Just as in our bound for $\Pr\left[\mathcal{E}_2^c\right]$, upon reaching $N(\eta_2) = 4\sigma + R + 1$, until the $R+1$-th server turns on, we can lower bound the change in $N(t)$ with the difference between two Poisson processes $X_a(t)$ and $X_d(t)$ of rate $k\lambda = \mu(R)$ again coupling each process with the arrival and departure process of our original system. Without loss of generality, assume that, at time $t < T_A$, we are in a state $\mathcal{S}(t) = s$ with $Z \leq R$ and $N = 4\sigma + R + 1$. Let $\beta = W_{R+1}(t)$ be the remaining setup time for the $R+1$-th server. Then, we show the following

$$\Pr\left[\inf_{\ell \in (t, t+\beta]} N(\ell) < R + 1 \,\middle|\, \mathcal{S}(t) = s\right] \leq e^{-4}. \tag{21}$$

Since we begin with $N(\eta_2) = 4\sigma + R + 1$, we must bound the probability that $X_d(\ell) - X_a(\ell)$ never exceeds 4σ for $\ell \in [\eta_2, \eta_2 + \beta]$; note that $X_d(\ell) - X_a(\ell)$ is itself a martingale, since

$$\begin{aligned}
\mathbb{E}\left[X_d(s+\delta) - X_a(s+\delta) | \mathcal{F}_s\right] &= X_d(s) + \mathbb{E}\left[X_d(s+\delta) - X_d(s) | \mathcal{F}_s\right] - X_a(s) - \mathbb{E}\left[X_a(s+\delta) - X_a(s) | \mathcal{F}_s\right] \\
&= X_d(s) - X_a(s)
\end{aligned}$$

from the independent increments property. Note that $\mathbb{E}\left[X_a(t + \frac{1}{\alpha})\right] = \frac{k\lambda}{\alpha} = \sigma^2$. Proceeding with our extended Chernoff bound, noting that $\beta \leq \frac{1}{\alpha}$,

$$\begin{aligned}
\Pr\left[\inf_{\ell \in (\eta_2, \eta_2 + \beta]} N(\ell) < R + 1 \,\middle|\, \mathcal{S}(\eta_2) = s\right] &\leq \Pr\left[\inf_{\ell \in (\eta_2, \eta_2 + \beta]} 4\sigma + R + 1 + X_a(\ell) - X_d(\ell) < R + 1 \,\middle|\, \mathcal{S}(\eta_2) = s\right] \\
&= \Pr\left[\sup_{\ell \in (\eta_2, \eta_2 + \beta]} X_d(\ell) - X_a(\ell) \geq 4\sigma + 1\right]
\end{aligned}$$

$$\leq \frac{\mathsf{E}\left[e^{\theta X_d(\beta+\eta_2)}\right]\mathsf{E}\left[e^{-\theta X_d(\beta+\eta_2)}\right]}{e^{\theta(4\sigma+1)}}$$
$$= \exp\left(\sigma^2(-2+e^{\theta}+e^{-\theta}-\theta(4\sigma+1)\right).$$

Let $\theta = \frac{2}{\sigma}$ and note the inequalities $e^{\theta}+e^{-\theta} = 2cosh(\theta) \leq e^{\frac{\theta^2}{2}}$ and $e^x \geq \frac{1}{1-x}$. Proceeding,

$$\begin{aligned}\exp\left(\sigma^2(-2+e^{\theta}+e^{-\theta}-\theta(4\sigma+1)\right) &\leq \exp\left(\sigma^2(-2+2e^{\frac{\theta^2}{2}}-\theta(4\sigma+1)\right)\\ &\leq \exp\left(\sigma^2(-2+2e^{\frac{\theta^2}{2}}-\theta(4\sigma+1)\right)\\ &\leq \exp\left(\sigma^2\left(-2+2\left(\frac{1}{1-\frac{\theta^2}{2}}\right)-\theta(4\sigma+1)\right)\right),\end{aligned}$$

which, after simplification, becomes $\leq e^{-4}$. Note the last line holds for $\sigma > 9 > \sqrt{18}+4$.

Combining the bounds. Now, applying a union bound over the events $\mathcal{E}_1^c, \mathcal{E}_2^c$, and $\mathcal{E}_3^c$,

$$\Pr\left[\mathcal{E}^C\right] \leq \Pr\left[E_1^c\right] + \Pr\left[E_2^c\right] + \Pr\left[E_3^c\right] \leq 7e^{-4}.$$

A.1.2 Proof of Claim A.4. We now bound the expected length of the slingshot period $\mathsf{E}[\eta_1]$. Recall the definitions of the hitting time $U_j = \min\{t > 0 : Z(t) = R - j\}$, the slingshot index $n^* = \min\left(j \in \mathbb{Z}^+ : U_{j+1} - U_j > \frac{1}{\alpha}\right)$, and the stopping time $\eta_1 = U_{n^*} + \frac{1}{\alpha}$. We prove this result by using the random walk tools developed in the proof of Claim A.3 to bound the expected contribution of each step in the slingshot period.

We first describe the expected contribution of each step, then bound this expected contribution, then sum these bounds across all steps. To begin, note that the j-th step, if it occurs, takes time $\min\left(U_{j+1}-U_j, \frac{1}{\alpha}\right)$, and that the length of this step is independent of the previous system history. Thus, the expected contribution of the j-th step is

$$\Pr\left(j\text{-th step occurs}\right)\mathsf{E}\left[\min\left(U_{j+1}-U_j, \frac{1}{\alpha}\right)\right] = \Pr\left(n* \geq j\right)\mathsf{E}\left[\min\left(U_{j+1}-U_j, \frac{1}{\alpha}\right)\right].$$

For the bound on $\Pr(n^* \geq j)$, note that

$$\begin{aligned}\Pr\left(n^* \geq j\right) &= \prod_{i=0}^{j-1} \Pr\left[\tau_j \leq \frac{1}{\alpha}\right]\\ &\leq \prod_{i=0}^{j-1} \Pr\left[\tau_j \leq \infty\right]\\ &= \prod_{i=0}^{j-1} \frac{R-i}{R}\\ &\leq e^{-\frac{1}{R}\sum_{i=0}^{j-1} i}\\ &= e^{-\frac{j(j-1)}{2R}}\end{aligned}$$

For the bound on $\mathsf{E}\left[\tau_j \wedge \frac{1}{\alpha}\right]$, first note that $\mathsf{E}\left[\tau_j \wedge \frac{1}{\alpha}\right] = \int_0^{\frac{1}{\alpha}} \Pr\left[\tau_j > x\right] dx$. Applying our bound on the tail from Lemma B.2,

$$\mathsf{E}\left[\tau_j \wedge \frac{1}{\alpha}\right] = \int_0^{\frac{1}{\alpha}} \Pr\left[\tau_j > x\right] dx$$

$$
\begin{aligned}
&= \int_0^{\frac{1}{\alpha}} \min\left\{1 - \frac{R-j}{R} + \frac{C}{\sqrt{k\lambda x}}, 1\right\} \mathrm{d}x \\
&= \frac{C^2}{k\lambda} + \frac{j}{R}\left(\frac{1}{\alpha} - \frac{C^2}{k\lambda}\right) + \int_{\frac{C^2}{k\lambda}}^{\frac{1}{\alpha}} \frac{C}{\sqrt{k\lambda x}} \mathrm{d}x \\
&\le \frac{C^2}{k\lambda} + \frac{j}{R}\left(\frac{1}{\alpha} - \frac{C^2}{k\lambda}\right) + \left[\frac{2C\sqrt{x}}{\sqrt{k\lambda}}\right]_{\frac{C^2}{k\lambda}}^{\frac{1}{\alpha}} \\
&= \frac{C^2}{k\lambda} + \frac{j}{R}\left(\frac{1}{\alpha} - \frac{C^2}{k\lambda}\right) + \left(\frac{2C}{\sqrt{R}}\sqrt{\frac{1}{\mu\alpha}} - \frac{2C^2}{k\lambda}\right)^+ \\
&\le \frac{j}{R}\frac{1}{\alpha} + \frac{2C}{\sqrt{R}}\sqrt{\frac{1}{\mu\alpha}}.
\end{aligned}
$$

Now we split the last two terms and then sum them individually.

$$
\begin{aligned}
\sum_{j=0}^{\infty} \Pr(n^* \ge j) \mathsf{E}\left[\tau_j \wedge \frac{1}{\alpha}\right] &\le \sum_{j=0}^{\infty} \left(\frac{j}{R}\frac{1}{\alpha} + \frac{2C}{\sqrt{R}}\sqrt{\frac{1}{\mu\alpha}}\right)^+ \sum_{i=0}^{j-1} \frac{R-i}{R} \\
&= \frac{1}{\alpha} \sum_{j=0}^{\infty} \frac{j}{R} \sum_{i=0}^{j-1} \frac{R-i}{R} + \frac{2C}{\sqrt{R}}\sqrt{\frac{1}{\mu\alpha}} \sum_{j=0}^{\infty} \sum_{i=0}^{j-1} \frac{R-i}{R} \\
&= \frac{1}{\alpha} + \frac{2C}{\sqrt{R}}\sqrt{\frac{1}{\mu\alpha}} \sum_{j=0}^{\infty} \sum_{i=0}^{j-1} \frac{R-i}{R} \\
&\le \frac{1}{\alpha} + \frac{2C}{\sqrt{R}}\sqrt{\frac{1}{\mu\alpha}} \sum_{j=0}^{\infty} e^{-\frac{1}{2R}j(j-1)} \\
&= \frac{1}{\alpha} + \frac{2C}{\sqrt{R}}\sqrt{\frac{1}{\mu\alpha}} \sum_{j=0}^{\infty} e^{-\frac{1}{2R}\left((j-\frac{1}{2})^2 - \frac{1}{4}\right)}.
\end{aligned}
$$

Turning the sum into an integral,

$$
\sum_{j=0}^{\infty} e^{-\frac{1}{2R}\left((j-\frac{1}{2})^2 - \frac{1}{4}\right)} \le 2 + \int_1^R e^{-\frac{1}{2R}\left((x-\frac{1}{2})^2 - \frac{1}{4}\right)} \mathrm{d}x \le 2 + \frac{1}{2}\sqrt{2\pi R}.
$$

Substituting this into our previous formula, we obtain

$$
\begin{aligned}
\sum_{j=0}^{\infty} \Pr(n^* \ge j) \mathsf{E}\left[\tau_j \wedge \frac{1}{\alpha}\right] &\le \frac{1}{\alpha} + \frac{2C}{\sqrt{R}}\sqrt{\frac{1}{\mu\alpha}} \sum_{j=0}^{\infty} e^{-\frac{1}{2R}\left((j-\frac{1}{2})^2 - \frac{1}{4}\right)} \\
&\le \frac{1}{\alpha} + \frac{2C}{\sqrt{R}}\sqrt{\frac{1}{\mu\alpha}} \left(2 + \frac{1}{2}\sqrt{2\pi R}\right) \\
&\le \frac{1}{\alpha} + \frac{3C}{\sqrt{\alpha\mu}},
\end{aligned}
$$

where in the last step we have used that the offered load $R \ge 100$. For $C = 7$ and $\frac{1/\alpha}{1/\mu} > 1000$, we obtain

$$
\mathsf{E}[\eta_1] \le \frac{2}{\alpha}.
$$

A.1.3 Proof of Claim A.5. We prove Claim A.5 by showing that, from any state s, with probability $\frac{1}{5}$ the residual time $T_A - t \leq \frac{2}{\alpha} + \frac{1.01}{\mu}$. The argument comes down to two simple subclaims.

(1) For any state $\mathcal{S}(t) = s_1$ with $Z \leq R$ and $N \geq R$,

$$\Pr\left[T_A - t < \frac{2}{\alpha} \,\middle|\, \mathcal{S}(t) = s_1\right] \geq \frac{1}{5}.$$

(2) Consider the alternative case, where state $\mathcal{S}(t) = s_2$ with $Z < N < R - 1$. Given t, let $T_R = \inf\{\ell > 0 : N(t+\ell) = R\}$ be the first time you return to $N(t + T_R) = R$. Then

$$\mathsf{E}\left[T_R | \mathcal{S}(t) = s_2\right] \leq \frac{1.01}{\mu}. \tag{22}$$

Before proving these subclaims, we complete the proof of the main claim. Without loss of generality, consider a state $\mathcal{S}(t)$ with $N \geq R$. Conditioning on whether N stays above $R+1$,

$$\begin{aligned}\mathsf{E}\left[T_A - t | \mathcal{S}(t) = s\right] &\leq \frac{2}{\alpha} + \frac{1.01}{\mu} + \frac{1}{5}(0) + \frac{4}{5}\left(\frac{1.01}{\mu} + \mathsf{E}\left[T_A - T_R\right]\right) \\ &\leq 5\left(\frac{2}{\alpha} + \frac{2}{\mu}\right),\end{aligned}$$

where we've used the slack in our estimate of $\frac{1}{5}$ to clean up the 1.01.

We now prove each subclaim in turn.

First subclaim. We prove this again through a coupling argument. Consider two competing independent Poisson processes of rate $k\lambda$, $X_a(\ell)$ and $X_d(\ell)$, with $X_a(t) = X_d(t)$. As we argued before, the change in N can be lower bounded by the change in $X_a(\ell) - X_d(\ell)$. By symmetry,

$$\Pr\left[\sup_{t \in [0, \frac{1}{\alpha}]} X_a(t) - X_d(t) > C\right] = \Pr\left[\inf_{t \in [0, \frac{1}{\alpha}]} X_a(t) - X_d(t) < -C\right],$$

for any C. Fix C and call this probability p_1. Then the probability that $N(\ell)$ reaches $N(\ell)+C+1$ in $\frac{1}{\alpha}$ time is p_1, and the probability that, after reaching $N(\ell)+C+1$, it does not come back down to $N(\ell)$ in $\frac{1}{\alpha}$ time is $(1-p_1)$. All that remains is to show that there exists some C such that $p_1(1-p_1) \geq \frac{1}{5}$. To see this, note that, by Poisson splitting, this supremum is essentially the supremum of an unbiased discrete random walk $V(j)$ with a random number of steps $T \sim \text{Poisson}(2\frac{k\lambda}{\alpha})$. Let $p_C^{(T)} = \Pr\left[\sup_{1 \leq i \leq T} V(i) > C\right]$ be the probability that $V(i)$ reaches C in T steps. By conditioning on the first outcome, we can see that, for $C \geq 0$

$$\begin{aligned}p_C^{(T)} &= \frac{1}{2}p_{C-1}^{(T-1)} + \frac{1}{2}p_{C+1}^{(T-1)} \\ &\leq \frac{1}{2}p_{C-1}^{(T)} + \frac{1}{2}p_{C+1}^{(T)};\end{aligned}$$

It follows that the differences $p_{C-1}^{(T)} - p_C^{(T)}$ are decreasing with C. To complete the proof we need only show that $p_{C-1}^{(T)} - p_C^{(T)}$ does not decrease too quickly. One can easily verify that the supremum is > 1 with probability $> \frac{1}{2}$. It follows from the previous property that, since at least two points are above $\frac{1}{2}$ (including 0), there must exist a point with $p(1-p) \geq \frac{1}{3}\frac{2}{3} \geq \frac{1}{5}$.

Second subclaim. We prove this subclaim via an analogy to the $M/M/1$. Consider a state with $N(t) = R - i$. Then the number of busy servers $Z(t) \leq R - i$. Now consider a flipped system, where arrivals occur at rate $Z(t)$ and departures occur at rate $k\lambda$. We desire the expected time of the busy period in this flipped system. With some probability, we finish within $\frac{1}{\alpha}$ time. If we do, our expected time must be $\leq \frac{1}{\mu}$, via a coupling argument. Moreover, the probability that such a busy period lasts longer than $\frac{1}{\alpha}$ is $\leq \frac{\alpha}{\mu}$, by Markov's inequality. It follows that the expected time to return is

$$\begin{aligned}\mathsf{E}[\tau_{\to R}] &\leq \frac{1}{\mu} + \frac{\alpha}{\mu}\mathsf{E}[\tau_{\to R}] \leq \frac{1}{\mu - \alpha} \\ &\leq \frac{1.01}{\mu},\end{aligned}$$

for $\frac{\mu}{\alpha} > 100$. This completes the proof.

A.2 Proof of Claim A.2

Recall that $T_B = \min\{t > 0 : Z(T_A + t) = R\}$ is the next time the $R+1$th server is turned off. Recall that j^* is the smallest index such that $\frac{\mu}{\alpha} j^* \geq 8\sigma + j^* + 1$. We now prove Claim A.2, which states that

$$\mathsf{E}[T_B] \leq \frac{1}{\alpha} + \frac{2}{(1-e^{-4})^2}\left(\frac{4\sigma + 1 + \frac{\mu}{\alpha}3\sqrt{R}}{k\mu(1-\rho)} + \frac{2}{\mu}\log\left(\frac{1}{1-e^{-4}}\left(4\sigma + 1 + \frac{\mu}{\alpha}3\sqrt{R}\right)\right)\right),$$

where $h(y)$ is the length of a busy period started by y jobs. To prove the claim, we first give an upper bound on the conditional expectation $\mathsf{E}[T_B | \mathcal{S}(T_A) = s]$ for any state s with $Q(T_A) = x$; note that $Z(T_A) = R + 1$ by definition. After proving that bound, we give an upper bound on $\mathsf{E}[Q(T_A)]$, by relating it to $\mathsf{E}\left[\inf_{t\in(0,T_A)} Z(t)\right]$. More specifically, we show the following claims.

Claim A.6. *Let $\mathcal{S}(T_A) = s$ be a state with $Q(T_A) = x$. Then,*

$$\mathsf{E}[T_B | \mathcal{S}(T_A) = s] \leq \frac{1}{\alpha} + 2\frac{x}{k\mu(1-\rho)} + 2\frac{\log(x)}{\mu}.$$

Claim A.7. *Recall that $T_A = \min\{t > 0 : Z(t) = R + 1\}$. Let $Z^* = \inf_{t\in(0,T_A)} Z(t)$ be the minimal number of busy servers reached during a cycle. Recall that $\sigma = \sqrt{\frac{k\lambda}{\alpha}}$. Then*

$$\mathsf{E}[Q(T_A)] \leq \frac{1}{1-e^{-4}}\left(4\sigma + \frac{\mu}{\alpha}\mathsf{E}[R - Z^*]\right).$$

Claim A.8. *Recall that $Z^* = \inf_{t\in(0,T_A)} Z(t)$ is the minimal number of busy servers reached during a cycle.*

$$\mathsf{E}[R - Z^*] \leq (2 + \frac{1}{1-e^{-4}})\sqrt{R} \leq \frac{3}{1-e^{-4}}\sqrt{R},$$

Combined these claims clearly give way to Claim A.6. We devote the remainder of this section to their proofs.

A.2.1 *Proof of Claim A.6.*

Recall that Claim A.6 says that, given a state $\mathcal{S}(T_A) = s$ with $Q(T_A) = x$,

$$\mathsf{E}[T_B | \mathcal{S}(T_A) = s] \leq \frac{1}{\alpha} + 2\frac{\lceil \log(x) \rceil + 1}{\mu} + \frac{x}{\mu k(1-\rho)}$$

Note that the same bound must also hold for the conditional expectation $\mathsf{E}[T_B | Q(T_A) = x]$.

Proof. We prove this via casework and induction. There are two major cases, based on whether $x \leq k(1-\rho)$. Before discussing the major cases, we treat the base case, when $x = 1$.

Base case. In the base case, the system is in a state where $Q(t) = 1$. The number of jobs $N(t)$ here is upper bounded by a the number of jobs in a system where servers can only turn off; denote these with a hat and a subscript OFF. The expected time remaining time before $\hat{N}_{\text{OFF}}(t) = \hat{Z}_{\text{OFF}}(t) = R$ is simply the length of two M/M/1 busy periods, where the arrival rate is $k\lambda$ and the service rate is $\mu(R+1) = k\lambda + \mu$. Thus, the expected remaining time in this over system is precisely $\frac{2}{\mu}$.

Case 1: $x \leq k(1-\rho)$. Without loss of generality, assume $x \in [2^{l-1}, 2^{\ell}]$, and let $\beta = W_{R+2^{\ell-1}}$ be the remaining setup time for the $(R+2^{\ell-1})$-th server. In this case, we show that

$$\mathsf{E}\left[T_B | \mathcal{S}(T_A) = s\right] \leq \beta + 2\frac{\log(x)+1}{\mu}$$

Assume inductively that, for any state $\mathcal{S}(T_A + t) = s_2$ with $Q(T_A + t) = [2^{\ell-2}, 2^{\ell-1})$, number of busy servers $Z(T_A + t) \geq R + 1$, and for $W_{R+2^{\ell-1}} \leq \beta - \gamma$, we know that

$$\mathsf{E}\left[T_B - t | \mathcal{S}(T_A + t) = s_2\right] \leq \beta - \gamma + 2\frac{\ell}{\mu}.$$

Without loss of generality, call the current time 0. Let $\tau = \min \quad t > 0 : N(t) \leq 2^{\ell-1} + R \quad$. We case on whether $\tau < \beta$.

Consider the case where $\tau < \beta$. After τ time, the system has $N(\tau) = 2^{\ell-1} + R + 1$ and we can apply our inductive hypothesis, with $W_{R+1+2^{\ell-2}} \leq \beta - \tau$. From that hypothesis, the remaining expected time from this point is $\leq \beta - \tau + 2\frac{\ell}{\mu}$.

Now, consider the case $\tau \geq \beta$. Using martingale arguments, we bound the combined conditional expectation

$$Pr(\tau \geq \beta)\mathsf{E}\left[\text{remaining time} | \tau \geq \beta\right] \leq \frac{2}{\mu} + \Pr\left\{\tau_{\text{spec}} = \beta\right\} \beta + \frac{2\ell}{\mu} \quad .$$

Define $\tau_{\text{spec}} = \min\{\tau, \beta\}$; then τ_{spec} is a.s. bounded (by β). Note that, until the number of busy servers Z becomes less than R (which must happen after time β), the number of jobs $N(t)$ is a supermartingale, i.e., it has negative drift in a strong sense. From Doob's Optional Stopping Theorem, it follows that

$$\mathsf{E}\left[N(0)\right] \geq \mathsf{E}\left\{N(\tau_{\text{spec}}) \right\} \geq \Pr\left\{\tau_{\text{spec}} = \beta\right\} \mathsf{E}\left\{N(\tau_{\text{spec}}) \, \tau_{\text{spec}} \leq \beta\right\} \quad .$$

We bound the conditional remaining time by considering a coupled system, which decomposes into a sequence of M/M/1 busy periods. Consider a coupled system where servers can only turn OFF from this point onward. Then the number of jobs in this coupled system $\hat{N}_{\text{OFF}} \geq N$, meaning that the time until the cycle ends ($\hat{Z}_{\text{OFF}} = \hat{N}_{\text{OFF}} = R$) in the coupled system is strictly larger than in the original system. Let $BP(a, b, c)$ denote the expected length of an M/M/1 busy period with arrival rate a, service rate b, and initial number of jobs c. Then $BP(a, b, c) = \frac{c}{b-a}$. It is direct that

$$\begin{aligned}
&\mathsf{E}\left[\text{remaining time original system}\right] \\
&\leq \mathsf{E}\left[\text{remaining time in OFF system}\right] \\
&= BP(k\lambda, \mu(R + 2^{\ell-1}), y) + \sum_{j=1}^{2^{\ell-1}} BP(k\lambda, \mu(R+i), 1) \\
&= \frac{y}{\mu 2^{\ell-1}} + \sum_{j=1}^{2^{\ell-1}} \frac{1}{\mu i} \\
&= \frac{y}{\mu 2^{\ell-1}} + \frac{H_{2^{\ell-1}}}{\mu}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\Pr\{\tau_{\text{spec}} = \beta\} \cdot \mathsf{E}\left[\text{remaining time} \mid \tau_{\text{spec}} = \beta\right] \\
&\leq \Pr\{\tau_{\text{spec}} = \beta\} \cdot \left(\beta + \frac{\mathsf{E}\left[N(\tau_{\text{spec}}) \mid \tau_{\text{spec}} = \beta\right]}{\mu 2^{\ell-1}} + \frac{H_{2^{\ell-1}}}{\mu}\right)^{\#}. \\
&\leq \frac{\mathsf{E}\left[N(0)\right]}{\mu 2^{\ell-1}} + \Pr\{\tau_{\text{spec}} = \beta\} \cdot \left(\beta + \frac{H_{2^{\ell-1}}}{\mu}\right) \\
&\leq \frac{2}{\mu} + \Pr\{\tau_{\text{spec}} = \beta\} \cdot \left(\beta + \frac{2\ell}{\mu}\right)
\end{aligned}$$

Combining the two cases, we find that, for any state s such that $Z \geq R+1$, $2^{\ell-1} + R \leq N < 2^{\ell} + R]$, and $W_{R+2^{\ell-1}} = \beta$,

$$\begin{aligned}
\mathsf{E}\left[\text{remaining time} \mid \mathcal{S}(0) = s\right] &\leq \Pr\{\tau_{\text{spec}} < \beta\} \cdot \left(\tau_{\text{spec}} + \beta - \tau_{\text{spec}} + \frac{\ell}{\mu}\right) \\
&\quad + \frac{2}{\mu} + \Pr\{\tau_{\text{spec}} = \beta\} \cdot \left(\beta + \frac{2\ell}{\mu}\right) \\
&= \beta + 2\frac{\ell+1}{\mu}, \\
&\leq \frac{1}{\alpha} + 2\frac{\lceil \log(x) \rceil + 1}{\mu}
\end{aligned}$$

as desired.

Second case: $x > k(1-\rho)$. To handle this case, we use an exactly analogous argument as the inductive case above, except that we define $\tau_{\text{spec}} = \min\{t > 0 : N(t) \leq k\}$ and set $\beta = W_k$. Casing on whether $\tau_{\text{spec}} < \beta$, we find that, if $\tau_{\text{spec}} < \beta$, then the expected conditional remaining time $\leq \tau_{\text{spec}} + \beta - \tau_{\text{spec}} + 2\frac{\lceil \log(k(1-\rho)) \rceil + 1}{\mu}$. In the other case, if $\tau_{\text{spec}} \geq \beta$, then we again couple to the OFF system. This gives that

$$\mathsf{E}\left[\text{remaining time}; \tau_{\text{spec}} = \beta\right] \leq + \frac{x - k(1-\rho)}{\mu k(1-\rho)} + \Pr\{\tau_{\text{spec}} = \beta\} \cdot \left(\beta + 2\frac{\lceil \log(k(1-\rho)) \rceil + 1}{\mu}\right).$$

Combined, these show that

$$\mathsf{E}\left[\text{remaining time}\right] \leq \beta + 2\frac{\lceil \log(k(1-\rho)) \rceil + 1}{\mu} + \frac{x - k(1-\rho)}{\mu k(1-\rho)},$$

proving the larger case.

Combining the results for these cases proves the claim, since, for any state s such that $Q = x$ and $Z = R+1$,

$$\begin{aligned}
\mathsf{E}\left[T_B \mid \mathcal{S}(T_A) = s\right] &\leq \frac{1}{\alpha} + 2\frac{\lceil \log\left(\min\{x, k(1-\rho)\}\right) \rceil + 1}{\mu} + \frac{[x - k(1-\rho)]^+}{\mu k(1-\rho)} \\
&\leq \frac{1}{\alpha} + 2\frac{\lceil \log(x) \rceil + 1}{\mu} + \frac{x}{\mu k(1-\rho)}.
\end{aligned}$$

A.2.2 Proof of Claim A.7. Since our bound on the $\mathsf{E}[T_B|Q(T_A) = x]$ is concave in x, to complete our result it suffices to bound $\mathsf{E}[Q(T_A)]$. Recall the minimal number of servers $Z^* = \min_{t\in[0,T_A)} Z(t)$. In Claim A.7, we reduce the bounding of $\mathsf{E}[Q(T_A)]$ to bounding $\mathsf{E}[R - Z^*]$.

$$\mathsf{E}[Q(T_A)] \le \frac{1}{1-e^{-4}}\left[4\sigma + 1 + \frac{\mu}{\alpha}\mathsf{E}[R - Z^*]\right].$$

We prove this claim by considering the behavior of the system immediately prior to the accumulation time T_A.

Proof. Notice that, since $Z(T_A) = R + 1$ by definition, it suffices to estimate $\mathsf{E}[N(T_A)]$. We begin by first classifying the trajectories leading up to time T_A into two different classes. Let $T_R = \sup\{0 < t < T_A : N(t) \le R\}$ be the last time before T_A that $N(t) \le R$. Recall the threshold $M = 4\left\lceil\frac{k\lambda}{\alpha}\right\rceil$. We now condition on whether, in the interval $[T_R, T_A]$, the number of jobs $N(t)$ was ever larger than $M + R$.

$$\begin{aligned}\mathsf{E}[N(T_A)] = \Pr\left[\sup_{t\in(T_R,T_A]} N(t) \le M+R\right]\mathsf{E}\left[N(T_A)\,\middle|\,\sup_{t\in(T_R,T_A]} N(t) \le M+R\right]\\ + \Pr\left[\sup_{t\in(T_R,T_A]} N(t) > M+R\right]\mathsf{E}\left[N(T_A)\,\middle|\,\sup_{t\in(T_R,T_A]} N(t) > M+R\right].\end{aligned}$$

The first conditional expectation we can upper bound directly by $M + R$. We focus on the second conditional expectation, and let $T_M = \min\{T_R < t < T_A : N(t) > M + R\}$ be the moment after time T_R when $N(t)$ rises above the threshold $M + R$. Given the observed state at time T_M, the value of $N(T_A)$ is a sample of a specific conditional random variable, whose expectation we bound via martingale arguments.

In particular, consider any state $\mathcal{S}(T_M)$ with number of jobs $N(T_M) = M+R+1$, number of jobs in service $Z(T_M) = R - i$, and remaining setup for $R + 1$-th server $W_{R+1}(T_M) = \beta$. By memorylessness, the distribution of $N(T_A)$ under our conditions is precisely

$$N(T_A) =^d \left[N(\beta)\,\middle|\,\mathcal{S}(0) = s, \min_{t\in(0,\beta)} N(t) \ge R+1\right].$$

We bound the conditional expectation of this random variable by relating it to a stopped martingale. Consider the system which begins in state s at time 0. Let the stopping time $\tau_{\text{work}} = \beta \wedge \min\{t > 0 : N(t) \le R\}$ be either when the $R+1$-th server stops setting up or when the $R+1$-th server next turns on, whichever comes sooner. Let $d(t) = k\lambda - \mu Z(t)$ be the signed difference between the arrival rate and service rate; note that, for $t \in [0, \tau_{\text{work}})$, one can compute $d(t)$ exactly from the initial state s. Recalling the view of the arrival and departure processes as independent Poisson processes with (possibly) variable rates, it is immediate that $N(t) - \int_0^t d(s)\mathrm{d}s$ is a martingale on the interval $[0, \tau_{\text{work}})$. Applying Doob's Optional Stopping Theorem, we have that

$$\mathsf{E}[N(0)] = \mathsf{E}\left[N(\tau_{\text{work}}) - \int_0^{\tau_{\text{work}}} d(s)\mathrm{d}s\right].$$

Since $N(\tau_{\text{work}}) \mid \tau_{\text{work}} < \beta = R$ by definition,

$$\mathsf{E}[N(\tau_{\text{work}})] = \Pr(\tau_{\text{work}} < \beta)\,R + \Pr(\tau_{\text{work}} = \beta)\,\mathsf{E}[N(\tau_{\text{work}})|\tau_{\text{work}} = \beta].$$

Rearranging (and subtracting R for convenience), we find that

$$\mathsf{E}[N(\tau_{\text{work}}) - R|\tau_{\text{work}} = \beta] = \frac{1}{\Pr(\tau_{\text{work}} = \beta)}\,\mathsf{E}[N_0] - R + \mathsf{E}\int_0^{\tau_{\text{work}}} d(s)\mathrm{d}s$$

$$
\begin{aligned}
&= \frac{1}{\Pr\left(\tau_{\text{work}} = \beta\right)} \left[M + 1 + \mathsf{E}\left[\int_0^{\tau_{\text{work}}} d(s)\mathrm{d}s \right] \right] \\
&\leq \frac{1}{\Pr\left(\tau_{\text{work}} = \beta\right)} \left[M + 1 + \mu i \mathsf{E}\left[\tau_{\text{work}}\right] \right], \\
&\leq \frac{1}{\Pr\left(\tau_{\text{work}} = \beta\right)} \left[M + 1 + \mu i \frac{1}{\alpha} \right],
\end{aligned}
$$

where the second-to-last line follows from the fact that the number of busy servers $Z(t)$ is increasing for $t \in [0, \tau_{\text{work}}]$, and the last line from the fact that $\tau_{\text{work}} \leq \beta \leq \frac{1}{\alpha}$. To complete our analysis of the conditional expectation for the state s, we require only a lower bound for $\Pr\left(\tau_{\text{work}} = \beta\right)$. But, by coupling the system to an M/M/1 with both arrival rate and departure rate $k\lambda$, we have already computed a lower bound, in Claim A.1. There, we showed that, keeping in mind that $\mathcal{S}(0) = s$ here,

$$\Pr\left(\tau_{\text{work}} = \beta\right) \geq p_2 = 1 - e^{-4}.$$

Thus,

$$\mathsf{E}\left[N(\tau_{\text{work}}) - R | \tau_{\text{work}} = \beta\right] \leq \frac{1}{p_2}\left[M + 1 + \mu i \frac{1}{\alpha} \right],$$

And, moving back to our original system,

$$\mathsf{E}\left[N(T_A) \sup_{t \in (T_R, T_A]} N(t) > M + R, \mathcal{S}(T_M) = s \right] \leq \frac{1}{p_2}\left[M + 1 + \frac{\mu}{\alpha} i \right] + R,$$

where one should recall that $i = R - Z(T_M)$ is the only term on the right side of this inequality which depends on the initial state s. This completes our bound of the conditional expectation; we proceed by extending this bound to the unconditioned case. First, note that, where T_M is well-defined, we can apply the lower bound $Z(T_M) \geq Z^*$. Moreover, for a state $\mathcal{S}(T_M) = s$ such that $i = R - Z(T_M)$, we have that $i \leq R - \mathsf{E}\left[Z^* \mid \mathcal{S}(T_M) = s, \sup_{t \in (T_R, T_A]} N(t) > M + R \right]$. Applying the tower property over the time T_M state $\mathcal{S}(T_M)$, we have that

$$
\begin{aligned}
&\mathsf{E}\left[N(T_A) \sup_{t \in (T_R, T_A]} N(t) > M + R \right] \\
&\leq \frac{1}{p_2}\left[M + 1 + \mu \mathsf{E}\left[R - Z^* \sup_{t \in (T_R, T_A]} N(t) > M + R \right] \frac{1}{\alpha} \right] + R.
\end{aligned}
$$

Since $Z(0) = R$ and the probability $p_2 \leq 1$, we also know that

$$
\begin{aligned}
&\mathsf{E}\left[N(T_A) \sup_{t \in (T_R, T_A]} N(t) \leq M + R \right] \\
&\leq M + R + 1 \\
&\leq \frac{1}{p_2}\left[M + 1 + \frac{\mu}{\alpha} \mathsf{E}\left[R - Z^* \sup_{t \in (T_R, T_A]} N(t) \leq M + R \right] \right] + R.
\end{aligned}
$$

It follows that

$$\mathsf{E}\left[N(T_A)\right] \leq \frac{1}{p_2}\left[M + 1 + \frac{\mu}{\alpha} \mathsf{E}\left[R - Z^*\right] \right] + R,$$

And, since $Q(T_A) = N(T_A) - Z(T_A) = N(T_A) - (R + 1)$,

$$\mathsf{E}\left[Q(T_A)\right] \leq \frac{1}{p_2}\left[M + 1 + \frac{\mu}{\alpha} \mathsf{E}\left[R - Z^*\right] \right] - 1$$

$$\leq \frac{1}{1-e^{-4}} \left(4\sigma + 1 + \frac{\mu}{\alpha} \mathsf{E}\left[R - Z^*\right] \right),$$

which proves the claim.

A.2.3 Proof of Claim A.8. To complete our bound on $\mathsf{E}\left[Q(T_A)\right]$ (and thus $\mathsf{E}\left[T_B\right]$ and $\mathsf{E}\left[X\right]$), we bound the expected maximum number of servers we shut off during a cycle. In particular, we show Claim A.8:

$$\mathsf{E}\left[R - Z^*\right] \leq \frac{3}{1-e^{-4}}\sqrt{R}$$

Proof. We upper bound $\mathsf{E}\left[R - Z^*\right]$ by first upper bounding the tail probability $\Pr\left(R - Z^* > i\right)$, then upper bounding the sum of these tail probabilities.

For positive integer j, recall that

$$U_j = \min\left\{t > 0 : Z(t) = R - i\right\}$$

is the first time the number of busy servers $Z(t) = R - i$. Recall that $T_A = \min\left\{t > 0 : Z(t) = R + 1\right\}$. We state a few observations. First, since busy servers turn off one at a time, the inequality $U_{j+1} \geq U_j$ must hold. Second, note that $U_j \leq T_A$ if and only if $Z^* \leq R - j$. Third, note that, at time U_j, we know the precise state of the system, since a departure must occur while the queue is empty in order to turn off a server. In particular, the number of busy servers: $Z(U_j) = R - j$, the queue length $Q(U_j) = 0$, and, accordingly, there are no servers in setup. Thus,

$$\begin{aligned}
\Pr\left(R - Z^* \geq i\right) &= \Pr\left(Z^* \leq R - i\right) = \Pr\left(U_i \leq T_A\right) \\
&= \Pr\left(\bigcap_{j=1}^{i} U_j \leq T_A \right) = \prod_{j=1}^{i} \Pr\left(U_j \leq T_A \,\middle|\, \bigcap_{\ell=1}^{j-1} U_\ell \leq T_A \right) \\
&= \prod_{j=1}^{i} \Pr\left(U_j \leq T_A \,\middle|\, U_{j-1} \leq T_A \right) \\
&= \prod_{j=1}^{i} 1 - \Pr\left(U_j > T_A \,\middle|\, U_{j-1} \leq T_A \right).
\end{aligned}$$

We proceed from here by lower bounding $\Pr\left(U_j > T_A \,\middle|\, U_{j-1} \leq T_A \right)$. Let $c_2 = 2$ and $C_3 = \frac{1}{1-e^{-4}}$. In particular, we claim that, for $j \geq c_2\sqrt{R}$,

$$\Pr\left(U_j > T_A \,\middle|\, U_{j-1} \leq T_A \right) \geq \frac{1}{C_3\sqrt{R}}.$$

Before proving this subordinate claim, we show how to use it to complete the upper bound on $\mathsf{E}\left[R - Z^*\right]$. The idea is to bound it via a Geometric series:

$$\begin{aligned}
\mathsf{E}\left[R - Z^*\right] &= \sum_{i=1}^{R} \Pr\left(R - Z^* \geq i\right) \\
&\leq \left\lceil c_2\sqrt{R} \right\rceil - 1 + \sum_{i=\lceil c_2\sqrt{R} \rceil}^{R} \Pr\left(R - Z^* \geq i\right) \\
&\leq \left\lceil c_2\sqrt{R} \right\rceil - 1 + \sum_{i=0}^{R} \left(1 - \frac{1}{C_3\sqrt{R}}\right)^j \\
&\leq \left\lceil c_2\sqrt{R} \right\rceil - 1 + C_3\sqrt{R}
\end{aligned}$$

$$\leq (c_2 + C_3)\sqrt{R}.$$

All that remains is to show that for $j \geq c_2\sqrt{R}$,

$$\Pr\ U_{j+1} > T_A \, U_j \leq T_A \ \geq \frac{1}{C_3\sqrt{R}}.$$

In other words, we show that, if we begin with $Z = R - j - 1$ and $Q = 0$, then with decent probability the $R+1$-th server turns on before we hit $Z = R - j$. We prove this directly, by analyzing a particular class of trajectories where our desired event occurs.

We first define three events. Let

$$A = \left(\inf_{t \in [U_j, U_j + \frac{1}{\alpha}]} N(t) \geq R - j \right)$$

be the event that the time $U_{j+1} > U_j + \frac{1}{\alpha}$. Recall that $M = 4\sigma$. Let

$$B = \left(\sup_{t \in [U_j, U_j + \frac{1}{\alpha}]} N(t) > M + R \right)$$

be the event that N rises above the threshold M in the first $\frac{1}{\alpha}$ after time U_j. Define

$$\tau_M = \min \ t > U_j : N(t) > M + R$$

as the moment that N crosses that threshold, and recall that W_{R+1} is the remaining setup time left on the $R+1$-th server. Finally, let

$$C = \min_{t \in [\tau_M, \tau_M + W_{R+1}(\tau_M))} N(t) \geq R + 1$$

be the event that the $R+1$-th server turns on before the number of jobs N has a chance to dip back down below R. Examining our probability of interest,

$$\begin{aligned} \Pr\ U_{j+1} > T_A \, U_j \leq T_A \ &\geq \Pr(A \cap B \cap C) \\ &= \Pr(C | A \cap B) \Pr(A \cap B) \\ &\geq \Pr(C | A \cap B) \left(\Pr(A) + \Pr(B) - 1\right). \end{aligned}$$

We have previously shown that for any state s such that $Z \leq R$ and $N = M + R + 1$,

$$\Pr(C | A \cap B | \mathcal{S}(\tau_M) = s) \geq p_2 = 1 - e^{-4};$$

it follows that the same bound holds for $\Pr(C | A \cap B)$.

Continuing on, since no servers can turn within $\frac{1}{\alpha}$ of time U_j, we can bound $\Pr(A) = \Pr\ \inf_{t \in [U_j, U_j + \frac{1}{\alpha}]} N(t) \geq R - j$ via analogy to the return probability of a biased random walk. Consider a biased random walk V with upwards probability $p = \frac{k\lambda}{k\lambda + \mu(R-j)}$ and downwards probability $q = 1 - p$. We can upper bound the probability we stay up with the classical $1 - \frac{q}{p} = \frac{j}{R}$ result. We now focus on bounding the final term $\Pr(B) = \Pr\ \sup_{t \in [U_j, U_j + \frac{1}{\alpha}]} N(t) > M + R$. We lower bound the probability that the N crosses the threshold by time $U_j + \frac{1}{\alpha}$ by considering the probability that $N(U_j + \frac{1}{\alpha}) > M + R$, then performing a Chernoff-style bound on the negated event. Let $Y_a \sim \text{Poisson}\ \frac{k\lambda}{\alpha}$ be the distribution of arrivals between time U_j and $U_j + \frac{1}{\alpha}$. One can see via

a coupling argument that the departure rate can only decrease in this interval. Thus, we can upper bound the number of departures with $Y_d \sim \text{Poisson}\left(\frac{\mu(R-j)}{\alpha}\right)$. Summarizing,

$$\begin{aligned}\Pr(B) &\geq \Pr\left(N\left(U_j + \frac{1}{\alpha}\right) > M + R\right)\\ &\geq \Pr(Y_a - Y_d > M + j)\\ &= 1 - \Pr(Y_a - Y_d \leq M + j).\end{aligned}$$

We continue with our Chernoff bound.

$$\begin{aligned}\Pr(Y_a - Y_d \leq M + j) &= \Pr(-\theta(Y_a - Y_d) \geq -\theta(M + j))\\ &= \Pr\left(e^{-\theta(Y_a - Y_d)} \geq e^{-\theta(M+j)}\right)\\ &\leq \mathsf{E}\left[e^{\theta Y_d}\right]\mathsf{E}\left[e^{-\theta Y_a}\right]e^{\theta(M+j)}\\ &= \exp\left(\mathsf{E}[Y_d](e^{\theta} - 1) + \mathsf{E}[Y_a](e^{-\theta} - 1) + \theta(M + j)\right).\end{aligned}$$

Choosing $\theta = \ln(1 + \epsilon)$ and $\epsilon = c\left(\frac{k\lambda}{\alpha}\right)^{-1} = c(\sigma)^{-1}$,

$$\begin{aligned}\Pr(Y_a - Y_d \leq M + j) &\leq \exp\left(\mathsf{E}[Y_d](e^{\theta} - 1) + \mathsf{E}[Y_a](e^{-\theta} - 1) + \theta(M + j)\right)\\ &= \exp\left(\mathsf{E}[Y_d](\epsilon) + \mathsf{E}[Y_a]\left(-\frac{\epsilon}{1+\epsilon}\right) + \ln(1 + \epsilon)(M + j)\right)\\ &\leq \exp\left(\mathsf{E}[Y_d]\epsilon^2 + [\mathsf{E}[Y_a] - \mathsf{E}[Y_d]]\left(-\frac{\epsilon}{1+\epsilon}\right) + \epsilon(M + j)\right)\\ &\leq \exp\left(\frac{k\lambda}{\alpha}\epsilon^2 + \left(\frac{j}{\alpha}\right)\left(-\frac{\epsilon}{1+\epsilon}\right) + \epsilon(M + j)\right)\\ &\leq \exp\left(c^2 + \left(\frac{j}{\sqrt{R}}\sqrt{\frac{\mu}{\alpha}}\right)\left(-\frac{c}{1+\epsilon}\right) + c + \sqrt{\frac{\alpha}{\mu}}\frac{j}{\sqrt{R}}\right)\\ &\leq \exp\left(c^2 + c + c_2\sqrt{\frac{\mu}{\alpha}}\left(-\frac{c}{1+\epsilon} + \frac{\alpha}{\mu}\right)\right)\end{aligned}$$

We make some simplifying assumptions to get an intelligible result. Set $c = 1$. Let $\epsilon^{-1} = \sqrt{\frac{k\lambda}{\alpha}} \geq 3$, let the ratio $\sqrt{\frac{\mu}{\alpha}} \geq \ln(R)$, and let the offered load $R \geq e^4$. Then

$$\begin{aligned}\Pr(Y_a - Y_d \leq M + j) &\leq \exp\left(2 + c_2\sqrt{\frac{\mu}{\alpha}}\left(-\frac{1}{2}\right)\right)\\ &\leq \exp\left(2 - \frac{c_2}{2}\ln(R)\right) = \frac{e^2}{R^{\frac{c_2}{2}}}\\ &= \frac{1}{\sqrt{R}}\frac{e^2}{R^{\frac{c_2-1}{2}}} = \frac{1}{\sqrt{R}}\frac{e^2}{R^{1/2}}\\ &\leq \frac{1}{\sqrt{R}},\end{aligned}$$

where we have used the fact that $c_2 = 2$. Returning to our previous claim, we find that

$$\Pr(B) \leq 1 - \frac{1}{\sqrt{R}},$$

And thus,

$$\begin{aligned}\Pr\ U_{j+1} > T_A \, U_j \le T_A \ &\ge \Pr(C|A \cap B)\,(\Pr(A) + \Pr(B) - 1)\\ &\ge p_2\ \frac{2}{\sqrt{R}} + 1 - \frac{1}{\sqrt{R}} - 1\\ &\ge p_2 \frac{1}{\sqrt{R}},\end{aligned}$$

where $p_2 = 1 - e^{-4}$. Returning to our original claim, it follows that

$$\mathsf{E}[R - Z^*] \le (2 + \frac{1}{1-e^{-4}})\sqrt{R} \le \frac{3}{1-e^{-4}}\sqrt{R},$$

as desired.

Plugging this in, we get that (upper bounding with p_2's)

$$\mathsf{E}[Q(T_A)] \le \left(\frac{1}{1-e^{-4}}\right)^{2\lceil 4\sigma + 1 + \frac{\mu}{\alpha}3\sqrt{R}\rceil},$$

which implies

$$\mathsf{E}[T_B] \le \frac{1}{\alpha} + \frac{2}{(1-e^{-4})^2}\,\frac{\lceil 4\sigma + 1 + \frac{\mu}{\alpha}3\sqrt{R}\rceil}{k\mu(1-\rho)} + \frac{2}{\mu}\log\left(\frac{1}{1-e^{-4}}\right)^{2\lceil 4\sigma + 1 + \frac{\mu}{\alpha}3\sqrt{R}\rceil},$$

as desired.

B LEMMAS ON RANDOM WALKS

This section presents lemmas on the continuous-time random walks used in the proof of Lemma 5.2. We first consider a discrete-time random walk in Lemma B.1. We then use this lemma to study the continuous-time random walks in Lemma B.2. Lemma B.3 derives bounds on the continuous-time random walks given that their values stay nonnegative during a period of time.

Lemma B.1. *Consider a discrete-time random walk $X(\cdot)$ with upwards probability $p = 1 - q > \frac{1}{2}$, starting at $X(0) = 0$. Let the first passage time $\tau_{-1} = \min\{i \in \mathsf{Z}_+ : X(i) = -1\}$. Then,*

$$\frac{q}{p}\left(1 - \frac{3p}{\pi(n+1)}\right) \le \Pr(\tau_{-1} \le 2n+1) \le \frac{q}{p}. \tag{23}$$

Proof. The upper bound directly follows from the fact that $\Pr(\tau_{-1} \le 2n+1) \le \Pr(\tau_{-1} < \infty)$ and the classical result $\Pr(\tau_{-1} < \infty) \le \frac{q}{p}$. To prove the lower bound, we first compute $\Pr(\tau_{-1} = 2\ell + 1)$ for each $\ell > n$. Since $X(0) = 0$, to hit the state -1 at time slot $2\ell+1$, the random walk must be at state 0 at time slot 2ℓ and does not hit -1 from time 0 to time 2ℓ. So to find $\Pr(\tau_{-1} = 2\ell+1)$, it suffices to compute the probability that $X(2\ell) = 0$ and $X(i) < -1$ for all $0 \le i \le 2\ell$; from there, with probability q, we stop and $\tau_{-1} = 2\ell + 1$. The number of paths for which this is the case is simply the Catalan number $C_\ell = \frac{1}{\ell+1}\binom{2\ell}{\ell}$, and the probability of each is $p^\ell q^\ell$. Thus,

$$\Pr(\tau_{-1} = 2\ell + 1) = q \cdot (pq)^\ell C_\ell.$$

Therefore,

$$\Pr(\tau_{-1} \le 2n+1) = 1 - \sum_{\ell=n+1}^{\infty} \Pr(\tau_{-1} = 2\ell + 1)$$

$$\geq \frac{q}{p}\left(1 - p\sum_{\ell=n+1}^{\infty}(qp)^{\ell}C_{\ell}\right)$$
$$= \frac{q}{p}\left(1 - p\sum_{\ell=n+1}^{\infty}(qp)^{\ell}\frac{1}{\ell+1}\frac{(2\ell)!}{\ell!\ell!}\right).$$

Applying Stirling's approximation gives

$$\Pr(\tau_{-1} \leq 2n+1) \geq \frac{q}{p}\left(1 - p\sum_{\ell=n+1}^{\infty}(qp)^{\ell}\frac{1}{\ell+1}\frac{1}{\sqrt{\pi\ell}}4^{\ell}\right)$$
$$= \frac{q}{p}\left(1 - \frac{p}{\sqrt{\pi}}\sum_{\ell=n+1}^{\infty}(4qp)^{\ell}\frac{1}{\ell+1}\frac{1}{\sqrt{\ell}}\right)$$
$$\geq \frac{q}{p}\left(1 - \frac{p}{\sqrt{\pi}}\sum_{\ell=n+1}^{\infty}\frac{1}{\ell^{3/2}}\right),$$

where we have used that $4qp \leq 1$ in the last inequality. Noting that $\sum_{\ell=n+1}^{\infty}\ell^{-3/2} \leq (n+1)^{-3/2} + \int_{n+1}^{\infty}\ell^{-3/2}\mathrm{d}\ell \leq \frac{3}{\sqrt{n+1}}$, we have

$$\Pr(\tau_{-1} \leq 2n+1) \geq \frac{q}{p}\left(1 - \frac{3p}{\pi(n+1)}\right),$$

as desired.

We derive the following lemma on a continuous-time random walk using Lemma B.1.

Lemma B.2. *For each $i \in \mathbb{Z}_+$ with $0 \leq i \leq R$, consider two independent Poisson processes $Y_a(t)$ of rate μR and $Y_d(t)$ of rate $\mu(R-i)$. Let the first passage time $\tau = \min\{t > 0 : Y_a(t) - Y_d(t) < 0\}$. Then, for any interval of length L,*

$$\frac{R-i}{R}\left(1 - \frac{3\sqrt{3}}{\sqrt{\pi\nu L}}\right) \leq \Pr(\tau \leq L) \leq \frac{R-i}{R},$$

where $\nu = 2\mu R - \mu i$ denotes the total rate of the two Poisson processes.

In particular, assume that $\frac{1/\alpha}{1/\mu} \geq 1000$ and $R \geq 128$. Then

$$\Pr\left(\tau \leq \frac{1}{\alpha}\right) \geq \frac{R-i}{R}e^{-\gamma},$$

where $\gamma = -\frac{1}{2}\ln(1-e^{-4}) > 0.009$.

Proof. The upper bound directly follows from the fact that $\Pr(\tau \leq L) \leq \Pr(\tau < \infty)$ and the classical result that $\Pr(\tau < \infty) \leq \frac{\mu(R-i)}{\mu R} = \frac{R-i}{R}$. We focus on proving the lower bound below.

Note that the superposition of the two Poisson processes $Y_a(t)$ and $Y_d(t)$ is a Poisson process with rate ν. We refer to events in the combined Poisson process as Poisson events. Conditioned on a Poisson event, the probability of it being an arrival (i.e. from $Y_a(t)$) is $p = \frac{R}{2R-i} \geq \frac{1}{2}$. Let $q = 1-p$. From here, note that, after conditioning on the number of Poisson events V during the interval $[0, L]$, we have, essentially, a discrete-time random walk with a finite number V of steps with upwards probability p. Recall the first passage time τ_{-1} for a finite random walk from Lemma B.1. It can be verified that

$$\Pr(\tau_{-1} \leq V \mid V = \ell) \geq \frac{q}{p}\left(1 - \frac{3\sqrt{3}p}{\sqrt{\pi}\cdot\sqrt{\ell+1}}\right) \geq \frac{q}{p}\left(1 - \frac{3\sqrt{3}}{\sqrt{\pi}\cdot\sqrt{\ell+1}}\right).$$

Note that $\tau \le L$ is equivalent to $\tau_{-1} \le V$. So

$$\begin{aligned}\Pr(\tau \le L) &= \Pr(\tau_{-1} \le V) \\ &\ge \sum_{\ell=1}^{\infty} \frac{q}{p}\left(1 - \frac{3\sqrt{3}}{\sqrt{\pi}\cdot\sqrt{\ell+1}}\right)\cdot \Pr(V=\ell) \\ &= \frac{q}{p}\left(1 - \frac{3\sqrt{3}}{\sqrt{\pi}\cdot \mathbb{E}\left[\sqrt{V+1}\right]}\right).\end{aligned}$$

It then suffices to compute an upper bound for $\mathbb{E}\left[\frac{1}{\sqrt{V+1}}\right]$ when $V \sim \text{Poisson}(\nu L)$:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\sqrt{V+1}}\right] &= \sum_{j=0}^{\infty} e^{-\nu L}\frac{(\nu L)^j}{j!}\frac{1}{\sqrt{j+1}} \\ &= \frac{1}{\nu L}\sum_{j=0}^{\infty} e^{-\nu L}\frac{(\nu L)^{j+1}}{(j+1)!}\sqrt{j+1} \\ &= \frac{1}{\nu L}\sum_{j=1}^{\infty} e^{-\nu L}\frac{(\nu L)^{j}}{(j)!}\sqrt{j} \\ &= \frac{1}{\nu L}\mathbb{E}\left[\sqrt{V}\right] \\ &\le \frac{1}{\sqrt{\nu L}},\end{aligned}$$

where we have used the concavity of the square root function. It follows that

$$\Pr(\tau \le L) \ge \frac{q}{p}\left(1 - \frac{3\sqrt{3}}{\sqrt{\pi \nu L}}\right).$$

as desired.

Finally, we set $L = \frac{1}{\alpha}$ to derive the lower bound on $\Pr\left(\tau \le \frac{1}{\alpha}\right)$. Note that under the conditions on $\frac{1/\alpha}{1/\mu}$ and R, the inequality $1 - x \ge e^{-\frac{2.8}{2.5}x}$ holds for $x = \frac{3\sqrt{3}}{\sqrt{\pi \nu \frac{1}{\alpha}}} \le \frac{3\sqrt{3}}{\sqrt{\pi \mu R \frac{1}{\alpha}}}$ and $\frac{3.3}{\sigma} < \gamma$. Thus,

$$\begin{aligned}\Pr\left(\tau \le \frac{1}{\alpha}\right) &\ge \frac{R-i}{R}\left(1 - \frac{3\sqrt{3}}{\sqrt{\pi \nu \frac{1}{\alpha}}}\right) \\ &\ge \frac{R-i}{R}e^{-\frac{2.8}{2.5}\frac{3\sqrt{3}}{\sqrt{\pi\mu R\frac{1}{\alpha}}}} \\ &\ge \frac{R-i}{R}e^{-\frac{3.3}{\sigma}} \\ &\ge \frac{R-i}{R}e^{-\gamma}.\end{aligned}$$

Lemma B.3. *For each $i \in \mathbb{Z}_+$ with $0 \le i \le R$, let $Y_a(t)$ be a Poisson process of rate $k\lambda$, and let $Y_d(t)$ be a Poisson process of rate $\mu(R-i)$. Let $\tilde{N}(t) = Y_a(t) - Y_d(t)$ be the difference between these processes. Then*

$$\mathbb{E}\left[\int_0^{\frac{1}{\alpha}} \tilde{N}(t)\,dt \;\middle|\; \tilde{N}(\ell) \ge 0, \forall \ell \in \left[0, \frac{1}{\alpha}\right]\right] \ge \frac{1}{2}\left(\frac{1}{\alpha}\right)^2 \mu i,$$

and

$$\mathsf{E}\left[\tilde{N}\left(\frac{1}{\alpha}\right) \mid \tilde{N}(\ell) \geq 0, \forall \ell \in \left(0, \frac{1}{\alpha}\right)\right] \geq \frac{\mu i}{\alpha}.$$

We first establish a shorthand and derive a sufficient condition for the above, which we then prove. For brevity, define the event $\uparrow s$ as

$$\uparrow s \triangleq \left[\tilde{N}(\ell) \geq 0, \forall \ell \in [0, s]\right].$$

Also, note that

$$\mathsf{E}\left[\tilde{N}(t)\right] = \mu i t,$$
$$\int_0^{\frac{1}{\alpha}} \mathsf{E}\left[\tilde{N}(t)\right] \mathrm{d}t = \frac{1}{2}\left(\frac{1}{\alpha}\right)^2 \mu i.$$

Applying Fubini's theorem,

$$\mathsf{E}\left[\int_0^{\frac{1}{\alpha}} \tilde{N}(t)\mathrm{d}t \mid \uparrow \frac{1}{\alpha}\right] = \int_0^{\frac{1}{\alpha}} \mathsf{E}\left[\tilde{N}(t) \mid \uparrow \frac{1}{\alpha}\right] \mathrm{d}t.$$

Thus, it suffices to show that, for all $t \in \left(0, \frac{1}{\alpha}\right)$,

$$\mathsf{E}\left[\tilde{N}(t) \mid \uparrow \frac{1}{\alpha}\right] \geq \mathsf{E}\left[\tilde{N}(t)\right]. \tag{24}$$

To do so, we go through $\mathsf{E}\left[\tilde{N}(t) \mid \uparrow t\right]$. In particular, we reduce (24) to the following two claims.

Claim B.1. *For all* $t \in \left(0, \frac{1}{\alpha}\right)$,

$$\mathsf{E}\left[\tilde{N}(t) \mid \uparrow \frac{1}{\alpha}\right] \geq \mathsf{E}\left[\tilde{N}(t) \mid \uparrow t\right].$$

Claim B.2. *For all* $t \in \left(0, \frac{1}{\alpha}\right)$,

$$\mathsf{E}\left[\tilde{N}(t) \mid \uparrow t\right] \geq \mathsf{E}\left[\tilde{N}(t)\right].$$

Proof of Claim B.2

We prove the latter claim first. We first cast the problem in the language of stopping times. Let $\tau_s = \min\left\{\ell > 0 : \tilde{N}(\ell) = -1\right\}$ be the (possibly infinite) $(0 \to -1)$ first passage time in this continuous-time random walk. Note that the event $\uparrow t$ is equivalent to the event $\{\tau_s > t\}$. Thus, we must show that for any $t \in \left(0, \frac{1}{\alpha}\right)$,

$$\mathsf{E}\left[\tilde{N}(t) \mid \tau_s > t\right] \geq \mathsf{E}\left[\tilde{N}(t)\right]. \tag{25}$$

Fix any $t \in \left(0, \frac{1}{\alpha}\right)$. To prove (25), we apply the Optional Stopping Theorem on an appropriate martingale. Define the martingale $M(\cdot)$ as

$$M(\ell) \triangleq \tilde{N}(\ell) - \mu i \ell;$$

one can verify this is a martingale via the independent increments property of a Poisson process. Note that for $\{M(\ell) : \ell \in \mathsf{R}_+\}$, the time $\tau_s \wedge t$ is a trivially almost surely bounded stopping time. Thus, by the Optional Stopping theorem,

$$\begin{aligned}
\mathsf{E}[M(0)] &= 0 \\
&= \mathsf{E}[M(t \wedge \tau_s)] \\
&= \mathsf{E}\left[\tilde{N}(t \wedge \tau_s)\right] - \mu i \mathsf{E}[t \wedge \tau_s] \\
&= \Pr(\tau_s > t)\,\mathsf{E}\left[\tilde{N}(t) \mid \tau_s > t\right] + \Pr(\tau_s \leq t)(-1) - \mu i \mathsf{E}[t \wedge \tau_s].
\end{aligned}$$

This implies

$$\begin{aligned}
\mathsf{E}\left[\tilde{N}(t) \mid \tau_s > t\right] &= \frac{1}{\Pr(\tau_s > t)}\left[\Pr(\tau_s \le t) + \mu i \mathsf{E}[t \wedge \tau_s]\right] \\
&\ge \frac{\mu i t \Pr(\tau_s > t)}{\Pr(\tau_s > t)} \\
&= \mu i t \\
&= \mathsf{E}\left[\tilde{N}(t)\right],
\end{aligned}$$

as desired.

Proof of Claim B.1

We wish to show that

$$\mathsf{E}\left[\tilde{N}(t) \uparrow \frac{1}{\alpha}\right] \ge \mathsf{E}\left[\tilde{N}(t) \uparrow t\right].$$

We prove this via stochastic dominance, i.e. we show that for any $y \ge 0$,

$$\Pr\left[\tilde{N}(t) > y \uparrow \frac{1}{\alpha}\right] \ge \Pr\left[\tilde{N}(t) > y \uparrow t\right]. \tag{26}$$

For brevity, let the event Y , $\tilde{N}(t) > y$. We reduce (26) to an observation and a claim. First, observe that the event $\uparrow \frac{1}{\alpha}$ implies the event $\uparrow t$, i.e.

$$\uparrow \frac{1}{\alpha} = \uparrow \frac{1}{\alpha} \cap \uparrow t.$$

Second, we make the following claim, whose proof is deferred to later.

Claim B.3. *We have*

$$\Pr\left[\uparrow \frac{1}{\alpha} Y, \uparrow t\right] \ge \Pr\left[\uparrow \frac{1}{\alpha} \uparrow t\right].$$

To complete the proof of Claim B.1, the rest is algebra:

$$\begin{aligned}
\Pr\left[Y \uparrow \frac{1}{\alpha}\right] &= \frac{\Pr\left[Y \cap \uparrow \frac{1}{\alpha}\right]}{\Pr\left[\uparrow \frac{1}{\alpha}\right]} \\
&= \frac{\Pr\left[Y \cap \uparrow t \cap \uparrow \frac{1}{\alpha}\right]}{\Pr\left[\uparrow \frac{1}{\alpha}\right]} && (27) \\
&= \frac{\Pr(\uparrow t)\Pr(Y \mid \uparrow t)\Pr\left[\uparrow \frac{1}{\alpha} \uparrow t, Y\right]}{\Pr\left[\uparrow \frac{1}{\alpha}\right]} \\
&\ge \frac{\Pr(\uparrow t)\Pr(Y \mid \uparrow t)\Pr\left[\uparrow \frac{1}{\alpha} \uparrow t\right]}{\Pr\left[\uparrow \frac{1}{\alpha}\right]} && (28) \\
&= \Pr(Y \mid \uparrow t),
\end{aligned}$$

where (27) is an application of our previous observation and (28) is an application of Claim B.3. After proving Claim B.3, we will be done.

Proof of Claim B.3

We want to show that

$$\Pr\left\{\uparrow \frac{1}{\alpha} Y, \uparrow t\right\} \geq \Pr\left\{\uparrow \frac{1}{\alpha} \uparrow t\right\}.$$

We argue this by conditioning on the value of $\tilde{N}(t)$, then arguing via stochastic dominance. To begin, we note that

$$\begin{aligned}
\Pr\left\{\uparrow \frac{1}{\alpha} Y, \uparrow t\right\} &= \sum_{x>y} \Pr\left\{\uparrow \frac{1}{\alpha} \cap \tilde{N}(t) = x \quad Y, \uparrow t\right\} \\
&= \sum_{x>y} \Pr\left\{\uparrow \frac{1}{\alpha} \tilde{N}(t) = x, Y, \uparrow t\right\} \Pr\left\{\tilde{N}(t) = x \; Y, \uparrow t\right\} \\
&= \sum_{x>y} \Pr\left\{\uparrow \frac{1}{\alpha} - t \quad \tilde{N}(0) = x\right\} \Pr\left\{\tilde{N}(t) = x \; Y, \uparrow t\right\},
\end{aligned}$$

where the last step is an application of the Markov property.

We now interpret this summation as an expectation. Let

$$f(x) \; , \Pr\left\{\uparrow \frac{1}{\alpha} - t \quad \tilde{N}(0) = x\right\}.$$

Then

$$\begin{aligned}
\Pr\left\{\uparrow \frac{1}{\alpha} Y, t\right\} &= \sum_{x>y} f(x) \Pr\left\{\tilde{N}(t) = x \; Y, \uparrow t\right\} \\
&= \mathsf{E}\left[f\left(\tilde{N}(t)\right) \quad Y, \uparrow t\right],
\end{aligned}$$

and, likewise,

$$\Pr\left\{\uparrow \frac{1}{\alpha} \uparrow t\right\} = \mathsf{E}\left[f\left(\tilde{N}(t)\right) \quad \uparrow t\right].$$

Now, note that $f(x)$ is increasing in x (from a straightforward coupling argument). Since $\tilde{N}(t) \mid Y, \uparrow t$ stochastically dominates $\tilde{N}(t) \mid \uparrow t$, it follows that

$$\mathsf{E}\left[f\left(\tilde{N}(t)\right) \quad Y, \uparrow t\right] \geq \mathsf{E}\left[f\left(\tilde{N}(t)\right) \quad \uparrow t\right],$$

as desired.

C BOUNDING A GAUSSIAN SUM

We show the following lemma and a corollary.

Lemma C.1. *For any $a > 0$ and $b \geq 0$, we can lower-bound a Gaussian sum as*

$$\sum_{i=1}^{R} e^{-ai^2 - bi} \geq e^{\frac{b^2}{2a}} \left(\frac{1}{2}\sqrt{\frac{\pi}{a}} - \frac{e^{-a(R+\frac{b}{2a})^2}}{a(R+\frac{b}{2a})} - \frac{b}{2a} + 1\right).$$

Proof. We begin with some algebraic manipulations.

$$\begin{aligned}
\sum_{i=1}^{R} e^{-ai^2-bi} &= \sum_{i=1}^{R} e^{-a(i+\frac{b}{2a})^2 + \frac{b^2}{4a}} = e^{\frac{b^2}{4a}} \sum_{i=1}^{R} e^{-a(i+\frac{b}{2a})^2} \\
&\geq e^{\frac{b^2}{4a}} \int_{1}^{R} e^{-a(x+\frac{b}{2a})^2} \mathrm{d}x
\end{aligned} \tag{29}$$

$$= e^{\frac{b^2}{4a}} \left(\int_{-\frac{b}{2a}}^{\infty} e^{-a(x+\frac{b}{2a})^2} dx - \int_{-\frac{b}{2a}}^{1} e^{-a(x+\frac{b}{2a})^2} dx - \int_{R}^{\infty} e^{-a(x+\frac{b}{2a})^2} dx \right),$$

where (29) follows because the summand is monotone decreasing in i. We bound each of these terms separately. For the first term, we observe from the well-known Gaussian integral formula that

$$\int_{-\frac{b}{2a}}^{\infty} e^{-a(x+\frac{b}{2a})^2} dx = \frac{1}{2}\sqrt{\frac{\pi}{a}}.$$

For the second term, we note that the integrand is ≤ 1, so that

$$\int_{-\frac{b}{2a}}^{1} e^{-a(x+\frac{b}{2a})^2} dx \leq \frac{b}{2a} + 1.$$

For the third term, we note that the integrand is decreasing in x and bound the remainder with the integral of an exponential:

$$\begin{aligned}
\int_{R}^{\infty} e^{-a(x+\frac{b}{2a})^2} dx &\leq \int_{R}^{\infty} e^{-a(x+\frac{b}{2a})(R+\frac{b}{2a})} dx \\
&= e^{-a(\frac{b}{2a})(R+\frac{b}{2a})} \int_{R}^{\infty} e^{-a(R+\frac{b}{2a})x} dx \\
&= e^{-a(\frac{b}{2a})(R+\frac{b}{2a})} \left(-\frac{e^{-a(R+\frac{b}{2a})x}}{a(R+\frac{b}{2a})} \right)_{R}^{\infty} \\
&= \frac{e^{-a(R+\frac{b}{2a})^2}}{a(R+\frac{b}{2a})}.
\end{aligned}$$

This completes the proof.

We also have the following corollary.

Corollary C.1.1. *Recall that* $\sigma = \sqrt{\mu R/\alpha}$. *When* $a = \frac{1}{R}$ *and* $b = \frac{3.3}{\sigma}$, *this gives*

$$\begin{aligned}
\sum_{i=1}^{R-1} e^{-\frac{i^2}{R} - \frac{3.3}{\sigma} i} &\geq e^{5.4\frac{\alpha}{\mu}} \left(\frac{\sqrt{\pi}}{2}\sqrt{R} - \frac{e^{-R(1+\frac{1.65}{\sigma})^2}}{1+\frac{1.65}{\sigma}} - 1.65\sqrt{\frac{\alpha}{\mu}}\sqrt{R} + 1 \right) \\
&\geq \frac{\sqrt{\pi}}{2} - \frac{1+e^{-R}}{\sqrt{R}} - 1.65\sqrt{\frac{\alpha}{\mu}}\sqrt{R},
\end{aligned}$$

so that, for $R \geq 128$ *and* $\frac{1/\alpha}{1/\mu} \geq 1000$,

$$e^{-\frac{1}{12}} \sum_{i=1}^{R-1} e^{-\frac{i^2}{R} - \frac{3.3}{\sigma} i} \geq 0.745\sqrt{R} \geq \frac{1}{2}\sqrt{R}. \tag{30}$$

Received August 2022; revised October 2022; accepted November 2022.