Probability in the Engineering and Informational Sciences (2026), **40:**4 516–556
doi:10.1017/S0269964826100266

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Multiserver-job response time under multilevel scaling

Isaac Grosof[1] and Hayriye Ayhan[2]

[1]Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, USA
[2]H. Milton School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA
Corresponding author: Isaac Grosof; Email: izzy.grosof@northwestern.edu

Keywords: applied probability; queueing theory; stochastic modeling

We study the multiserver-job setting in the load-focused multilevel scaling limit, where system load approaches capacity much faster than the growth of the number of servers n. We consider the "1 and n" system, where each job requires either one server or all n. Within the multilevel scaling limit, we examine three regimes: load dominated by n-server jobs, 1-server jobs, or balanced. In each regime, we characterize the asymptotic growth rate of the boundary of the stability region and the scaled mean queue length. We demonstrate that mean queue length peaks near balanced load via theory, numerics, and simulation.

1. Introduction

Queueing theory primarily emphasizes models where only a single job can be served at a time, or where multiple jobs can be served but each requires the same amount of resources. Both classes of models, single-server and homogeneous multi-server models, capture some fraction of the space of real queueing systems while remaining conducive to theoretical study. However, homogeneous models do not reflect the behavior of many important systems, including many modern computing systems.

For example, in modern datacenters, such as those of Microsoft, Google, Amazon, etc., each job requests an amount of computing resources (cores, processors, etc.) specific to the needs of that job. In Google's recently published trace of its "Borg" computing cluster, given in [32], the requested resources of different jobs vary by a factor of 10^5, as shown in Figure 1. Models in which different jobs request different amounts of resources are also important for bandwidth-sharing applications, where a job's resource requirement is its bandwidth need, rather than computing resources, or in high-performance computing settings with dedicated computing nodes.

Throughout this paper, we therefore focus on a "multiserver-job model" (MSJ), in which each job requests some number of servers, and concurrently occupies that many servers throughout its time in service.

There has been recent progress in analyzing the performance of MSJ queueing models, both under basic First-Come First-Served (first-come, first-served) service by Grosof et al. [15], as well as under more advanced scheduling policies by Anselmi and Doncel [1], and Grosof et al. [13, 16]. The scheduling policies which are popular in today's datacenters focus on heuristics closely based on first-come, first-served scheduling, such as backfilling heuristics as discussed by Carastan-Santos et al. [30], Srinivasan et al. [5], and Wang and Guo [34]. Analyzing first-come, first-served scheduling is therefore important to understanding the behavior of these systems.

Recent results on mean response time in MSJ first-come, first-served models have focused on asymptotic limits, both the conventional heavy-traffic limit studied by Grosof et al. [15], and a dual-scaling

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-ShareAlike licence (http://creativecommons.org/licenses/by-sa/4.0), which permits re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited.

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

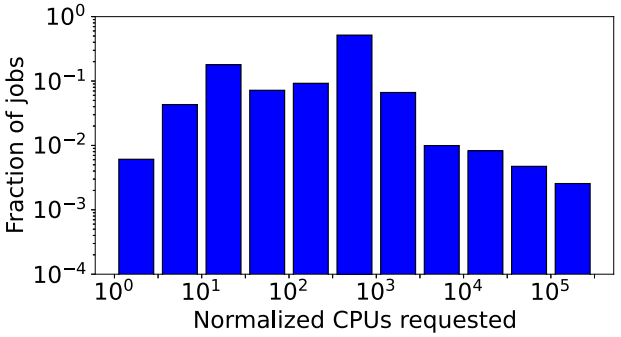


Figure 1. *The distribution of the number of CPUs requested in Google's Borg trace, published in Tirmazi et al. [32]. The number of CPUs is normalized to the size of the smallest request observed, not an absolute value. The peak of the distribution is around 500 normalized CPUs, and there is a significant probability mass anywhere from* 1 *to* 10^5 *normalized CPUs.*

limit, where both the load and the number of servers grow asymptotically studied by Wang et al. [35] and [18]. The latter work focuses on scalings where the number of servers grows so quickly relative to the load that the system experiences zero queueing or zero expected waiting time in the asymptotic limit. We call this limit the "SFMS" limit.

This recent work leaves open the problem of performance analysis in the "load-focused multilevel scaling" (LFMS) limit, where load approaches capacity much faster than the asymptotic growth of the number of servers. Stability results and numerical results exist in the LFMS limit, but no asymptotic scaling results exist. In this paper, we analyze the asymptotic scaling of the mean queue length of an MSJ system in the LFMS limit. To the best of our knowledge, this is the first such analysis. In particular, we inaugurate the study of the LFMS limit by focusing on a specific MSJ system. We study the "1 and n" system, where each job's server need is either a single server, or all n servers in the system, with first-come, first-served service. Our main objective in considering a system with two extreme server needs is to understand the impact of this variability in server needs on system performance. The insights derived from our setting are applicable in more general settings with highly variable server needs across more than two classes. As a result, our insights have practical implications since systems with extreme server needs arise in applications such as Large Language Model training studied by Wan et al. [33] and Ziheng et al. [36] and supercomputers such as Frontier, operated by Oak Ridge National Laboratory [23].

Our key insight, as discussed in Section 10, is that settings with load spread evenly across high and low-resource job types exhibit much higher mean queue lengths than settings with load concentrated on just high-resource jobs, or just low-resource jobs, controlling for system utilization. This insight generalizes beyond the 1-and-n system that we study, to cover general high-variability MSJ settings, such as AI training and supercomputing.

Within the LFMS limit in the 1 and n system, we encounter three regimes with distinct behavior, depending on the fraction of system load corresponding to jobs of each type. We call these regimes the 1-server dominated regime, the balanced regime, and the n-server dominated regime.

We provide the first analysis of mean queue length scaling with respect to the number of servers n in each of these three LFMS regimes.

Specifically, we characterize the asymptotic growth rate of the completion rate μ and the scaled mean queue length $\mathbb{E}[Q(1-\rho)]$, where Q is the queue length and $\rho = \lambda/\mu$ is the fraction of capacity in use. We characterize these growth rates in each of three LFMS regimes: The n-server dominated, balanced, and 1-server dominated regimes, in Theorems 4.1 and 4.2, respectively. Note that, as demonstrated in Theorem 4.1, the asymptotics are the same in the first two regimes. We also empirically evaluate our asymptotic expressions in comparison to prior fixed-n characterizations in Section 9.

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

The rest of the paper is organized as follows: Section 2 reviews relevant literature in relation to our results. Section 3 provides the details of our MSJ model and the related saturated system. Section 4 states our main results. Section 5 provides preliminary characterizations of system performance. Sections 6–8 prove our main results in the three regimes discussed above. Section 9 compares our asymptotic results against fixed-n formulas. Section 10 focuses on generalizations of our original 1 and n system. Finally, Section 11 concludes the paper.

2. Prior work

We review prior theoretical work on MSJ systems—for an initial overview, see [17]. We discuss the MSJ stability region under the First-Come First-Served (first-come, first-served) policy in Section 2.1, mean response time under first-come, first-served in Section 2.2, and results for other scheduling policies in Section 2.3.

2.1. Stability of the MSJ system under first-come, first-served

In the MSJ setting, the First-Come First-Served (first-come, first-served) policy with head-of-line blocking does not consistently keep all servers occupied, reducing its stability region relative to more optimized policies. Characterizing this stability region is an important line of work.

An important approach for characterizing the stability region of an MSJ system under first-come, first-served is the *saturated system* approach of [2]. This work provides a general framework under which one can relate the stability region of a queueing system that completes jobs in near- first-come, first-served order to the throughput of a *saturated* queueing system. In the MSJ system, jobs enter service in first-come, first-served order, causing the completion order to be very close to the first-come, first-served order. In the saturated system, an unlimited amount of jobs are available at all times, replacing the external arrival process.

This approach has been used to characterize the stability region of a variety of MSJ first-come, first-served systems by Grosof et al. [15], though often requiring the solution of a system of linear equations to obtain the stability region. More explicit symbolic results have been proven for special cases of job duration distribution and server need, in the works of Grosof et al. [14], Rumyantsev and Morozov [28], and [27].

These results have also been generalized to certain asymptotic limits, in the case of exponential duration and 1-or-n server need in the work of [24], specifically the limit as the n-server jobs complete much faster than the 1-server jobs, and the limit as the number of servers n diverges.

We study the same 1-and-n exponential setting as [24]. The asymptotic limits in that paper are simpler than our asymptotic limit—we allow both $n \to \infty$ and $p_n \to 0$, where p_n is the probability that a job needs all n servers, allowing us to explore a richer variety of asymptotic behavior. Our results cover both the stability region and the mean response time.

2.2. Response time of MSJ system under first-come, first-served

Much less is known about the MSJ first-come, first-served mean response time. The only setting in which mean response time has been exactly characterized is the setting with $n = 2$ servers and where 1-server jobs require an exponential service duration [3, 9].

A related but distinct model has been considered where a job's service must start at the same time at each of its servers, but may end at different times on different servers. If completion times are independent across the servers, exact results on mean waiting time and the distribution of waiting time are known, see Green [11] and Seila [29]. In contrast, the MSJ model requires that service ends at the same time across all of a job's servers.

In more general MSJ settings, one must turn either to asymptotic results or numerical algorithms and approximations.

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

As for numerical results for response time, the standard technique is to employ matrix geometric and matrix analytic methods, and we see no reason why off-the-shelf methods would not be applicable, such as those described by Latouche and Ramaswami [19]. These methods were also used to derive some of the stability results discussed above, such as [28]. Such numerical results are far more efficient than simulation, but are inherently restricted to examining a single set of parameters at a time. They cannot provide asymptotic results, which are the focus of this paper.

Mean response time, or equivalently mean waiting time, has been explored in several asymptotic limits. One type of limit that has received significant attention is a scaling limit where the number of servers, the largest server needed in the system, and the load all scale polynomially, mimicking the classical Halfin–Whitt limit, in the work of [35] and [18]. However, results in this limit have been limited to the case where the load is low enough to allow stability with the maximal number of wasted servers: If the largest server need of a job is $m_{\max}$, this line of work requires that the mean number of servers occupied is below $n - m_{\max}$. This limit becomes trivial when $m_{\max} = n$, as in this work. Results in this limit focus on the asymptotic decay of queueing probability in [35] or of mean waiting time in the scaling limit in [18].

We refer to this limit as the server-focused multilevel scaling (SFMS) limit. The SFMS limit has a much lighter load than the LFMS limit of this paper.

The remaining asymptotic work has focused on the fixed-n limit, as the load approaches capacity. This limit mirrors the classical heavy-traffic limit. Tight asymptotic characterization of mean response time is known for MSJ first-come, first-served in this limit, proven by Grosof et al. [15]. However, these results do not immediately generalize to any multilevel-scaling limit. This generalization is the contribution of this paper.

2.3. MSJ system under other scheduling policies

Outside of first-come, first-served scheduling, more results are known in the MSJ setting under nontrivial scheduling policies.

Throughput-optimality results have been established both in a preemptive setting, for the MaxWeight policy in the work of [22], and in a nonpreemptive setting, for the Randomized Timers policy introduced by Psychas and Ghaderi [25], though the Randomized Timers policy has extremely poor response time in certain settings. The recent Markovian Service Rate policy introduced by Chen et al. [6] uses a simple timer-based scheme to achieve throughput optimality, including in a nonpreemptive setting, with predictable and acceptable response time in all settings.

Moving to policies whose analyses primarily focus on mean response time, a significant line of work has focused on settings where all jobs' server needs are powers of 2 and n is a power of 2, and settings with related divisibility assumptions. First, the preemptive ServerFilling policy has been shown to match the response time of a resource-pooled M/G/1/ first-come, first-served, in the power-of-2 setting with a fixed number of servers n, studied by Grosof and Harchol-Balter [13], Grosof et al. [12]. In this setting, the ServerFilling-SRPT policy was shown by Grosof et al. [16] to achieve optimal mean response time in heavy traffic. Finally, in the power-of-2 setting, the Balanced Splitting dispatching policy was exactly analyzed by Anselmi and Doncel [1], with both fixed-n and scaling analysis results.

In the SFMS limit discussed in Section 2.2, the Smallest Need First policy has also been studied by Hong and Wang [18], resulting in an expanded set of asymptotic limits in which expected waiting time decays to 0.

Many practically used scheduling policies are backfilling policies, which can be seen as starting with the first-come, first-served scheduling policy, with additional steps taken to use capacity that first-come, first-served would otherwise leave unused, as discussed by Srinivasan et al. [30], Carastan-Santos et al. [5], and Wang and Guo [34]. These policies have been investigated in simulation and in practice, but little is known about their theoretical properties. We hope to strengthen the theoretical understanding of the first-come, first-served policy, to build toward a theoretical understanding of backfilling policies.

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

3. Model

We introduce the MSJ model in Section 3.1, and discuss the saturated system, a key theoretical tool for analyzing the MSJ system, in Section 3.2.

3.1. Multiserver-job system

We consider an MSJ system with n identical servers. Customers arrive according to a Poisson process with rate λ. We specifically consider a model in which there are two types of jobs: Jobs which need to be served by one server, and jobs that need all n servers. The probability that a job needs one server is p_1, and (as defined in Section 2.1) the probability that a job needs all n servers is $p_n = 1 - p_1$, i.i.d. The service time for a 1-server job is $Exp(\mu_1)$, and for an n-server job is $Exp(\mu_n)$, both independent of all other jobs. Jobs are served in First-Come First-Served (first-come, first-served) order with head-of-line blocking, meaning that jobs are admitted into service in first-come, first-served order until a job is reached that cannot be served in the remaining available servers. No further jobs are served. We define q to be the queue length at a particular point in time, and Q to be the stationary queue length random variable.

Note that the symbol p_n always refers to the fraction of jobs that use all of the servers, even as the number of servers n changes.

Prior work (see, e.g., [14, 27, 28], [15], or [24]) has demonstrated the existence of a stability threshold μ, where the system is stable if and only if $\lambda < \mu$. Note that μ has been explicitly characterized by several methods, including analysis of the corresponding saturated system, which we define and discuss further in Section 3.2. We build on these prior explicit results to establish the scaling of the stability threshold μ. We define $\rho = \lambda/\mu$ to be the fraction of the stability region that is in use.

A natural asymptotic limit is the $\lambda \to \mu$ scaling limit, known as the "heavy traffic" limit. The limiting value of the mean scaled queue length, $\lim_{\rho\to 1} \mathbb{E}[Q(1-\rho)]$, has been characterized by analyzing the saturated system, which we likewise discuss in Section 3.2.

However, the focus of this paper is the LFMS limit where we allow ρ to grow quickly enough that $\mathbb{E}[Q(1-\rho)]$ reaches its limiting value, and only then scale the number of servers to study the secondary effects of that scaling.

When scaling the number of servers, it is natural to also consider scaling the fraction of single-server jobs p_1 toward 1, as the number of single-server jobs that can be served at once grows asymptotically.

Our goal is to characterize asymptotic behavior of the mean scaled queue length for this system under two secondary asymptotic limits: $p_1 \to 1$, and $n \to \infty$. Both of these asymptotic limits are secondary to our load scaling.

In particular, it makes sense to consider taking both limits simultaneously as there is a natural separation, based on whether the system load is dominated by 1-server jobs, by n-server jobs, or is balanced between the two.

We refer to the inherent workload of a job as the product of its server need and its duration. We refer to the overall load as the rate at which this workload is arriving to the system. The overall load of 1-server jobs is $\lambda\frac{p_1}{\mu_1}$, and the overall load of n-server jobs is $\lambda\frac{np_n}{\mu_n}$. As a result, 1-server jobs require $\frac{p_1}{p_1+np_n}$ fraction of the total load.

We focus on the limit where $n \to \infty$ and $p_n \to 0$, holding μ_1 and μ_n constant, and adjusting λ as necessary to focus on the $\rho \to 1$ limit. As $p_1 \to 1$, the fraction of the total load required by 1 server jobs can be approximated as $\frac{1}{1+np_n}$.

In this paper, we define the notation $f(n) = o(g(n))$ and $f(n) = O(g(n))$ as follows:

$$f(n) = o(g(n)) \Leftrightarrow \lim_{n\to\infty} \frac{f(n)}{g(n)} = 0, \quad f(n) = O(g(n)) \Leftrightarrow \lim_{n\to\infty} \frac{f(n)}{g(n)} \in (-\infty, \infty).$$

Note that when $f(n) = o(g(n))$ or $f(n) = O(g(n))$, $f(n)$ may be negative while $g(n)$ is positive.

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

We define $\omega(\cdot), \Omega(\cdot)$, and $\theta(\cdot)$ similarly. If $f(n) = o(g(n))$, then $g(n) = \omega(f(n))$. If $f(n) = O(g(n))$, then $g(n) = \Omega(f(n))$. If $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$, then $f(n) = \theta(g(n))$.

If $np_n = \omega(1)$, or equivalently $p_n = \omega(\frac{1}{n})$, the fraction of load occupied by 1-server jobs converges to 0 as $p_n \to 0, n \to \infty$. We call this the "n-server dominated" regime.

If $p_n = c/n$, for an arbitrary positive constant c (which would imply that $np_n = \theta(1)$, or equivalently $p_n = \theta(\frac{1}{n})$), the fraction of load occupied by 1-server jobs converges to a value strictly between 0 and 1. We call this the "balanced" regime.

Finally, if $np_n = o(1)$, or equivalently $p_n = o(\frac{1}{n})$, the fraction of load occupied by 1-server jobs converges to 1. We call this the "1-server dominated" regime. We specifically focus on the polynomial-scaling subset of the 1-server dominated regime, where $p_n = 1/n^\alpha, \alpha > 1$.

Note that throughout the paper, all limits are taken as the number of servers n tends to infinity.

3.2. Saturated system

A key tool for analyzing the MSJ system is the "saturated system." The saturated system is a closed system with the same MSJ completion setup as the original open system. In the saturated system, whenever a server is unoccupied, and there is no job blocking the head of the line, jobs enter the system with the same i.i.d. probabilities p_1 and p_n of 1 and n-server jobs, respectively. There is no separate stochastic arrival process. Jobs enter in response to a completion until either all n servers are occupied, or an n-server job is generated that cannot be served immediately.

Saturated system analysis can be used to derive the stability region of the original MSJ system, because the stability region of the original system is bounded by the throughput of the saturated system, as demonstrated by Grosof et al. [15], Baccelli and Foss [2], and Foss and Konstantopoulos [10].

Saturated system analysis can also be used to derive the mean queue length of open MSJ systems under first-come, first-served service. To do so, [15] makes use of the *relative completions* function Δ, which maps states of the saturated system to real values. Note that the concept of relative arrivals and relative completions was also present in work prior to that result [7, 8].

To define Δ, we must first define several quantities in the saturated system. Let y denote a state of the saturated system and define μ_y as the completion rate in state y. Let ν_y denote the total transition rate out of state y, and $\nu_{y,y'}$ denote the specific transition rate from state y to another state y', where y' is an arbitrary state of the saturated system. Finally, define μ as the time-average completion rate of the saturated system (e.g., the throughput).

Letting y denote a state of the saturated system, $\Delta(y)$ is defined as the solution to the following system of equations:

$$\Delta(y) := \frac{\mu_y - \mu}{\nu_y} + \sum_{y'} \frac{\nu_{y,y'}}{\nu_y} \Delta(y'). \tag{1}$$

This equation (1) can be seen as the Poisson equation for a Markov Reward Process whose instantaneous reward is the completion rate μ_y. This relative completion function impacts mean queue length in the MSJ system similarly to the impact of job size variance in an M/G/1 queue: It effectively measures the variability in completion rate. If the system is in a state y with a large value of $\Delta(y)$, the system is likely to experience a high completion rate in the near future.

Note that the solution to this system of equations is only defined up to an additive constant. Letting Y denote the time-average random variable for the saturated system state, [15] adopts the convention that $\mathbb{E}[\Delta(Y)] = 0$, which results in a unique solution.

Let Y_d denote the state-average random variable corresponding to the embedded discrete-time Markov chain (DTMC), updating at completion epochs. [15, Theorem 4.2] proves the following result, for an arbitrary MSJ first-come, first-served system, in the heavy traffic limit:

$$\lim_{\rho \to 1} \mathbb{E}[Q(1-\rho)] = \mathbb{E}[\Delta(Y_d)] + 1. \tag{2}$$

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

Our goal in this paper is to analyze the asymptotic growth rate of $\mathbb{E}[\Delta(Y_d)]$ in each of our three regimes, within the surrounding LFMS context.

4. Main results

We study the heavy traffic limit under LFMS in each of three regimes:

(1) n-server dominated: $p_n = \omega(1/n)$,
(2) Balanced: $p_n = c/n$, where $c > 0$ is an arbitrary positive constant,
(3) 1-server dominated: $p_n = o(1/n)$, with a specific focus on the case where $p_n = 1/n^\alpha, \alpha > 1$.

The main results of the paper are characterizations of $\mathbb{E}[\Delta(Y_d)]$, which determines the scaled mean queue length $\mathbb{E}[Q(1-\rho)]$ as discussed in Section 3.2, and μ, which determines the stability region, in the limit as $p_n \to 0$ and $n \to \infty$, in each of these three regimes.

We now state our results for each of these regimes:

Theorem 4.1. *In regime 1 which is the n-server dominated regime, namely $p_n = \omega(1/n)$, as $p_n \to 0, n \to \infty$, and in regime 2, which is the balanced regime, namely $p_n = c/n$, for some positive constant c, as $p_n \to 0, n \to \infty$,*

$$\mu = (1 + o(1))\frac{\mu_1}{p_n \ln(1/p_n)} \text{ and}$$
$$\mathbb{E}[\Delta(Y_d)] = (1 + o(1))\frac{1}{2p_n}.$$

Proof. The proof for regime 1 is given in Section 6, split into Lemma 6.1 and Theorem 6.1 and the proof for regime 2 is given in Section 7, split into Lemma 7.1 and Theorem 7.1. □

Intuitively, in the n-server dominated regime, our result shows that the throughput is dominated by 1-server jobs, which are served while there is an n-server job at the head of the queue, and most of the servers are idle. This is why μ is far below $n\mu_1$ (the service rate when all servers are busy with 1-server jobs). This gives rise to periodic fluctuation in service rate with cycles of length $\frac{1}{p_n}$ (the number of 1-server jobs between each n-server job). These fluctuations control the size of $\mathbb{E}[\Delta(Y_d)]$, and hence the scaled mean queue length.

Note that the asymptotic growth rates in the balanced regime do not depend on c. The intuition behind this is that the balanced load regime exhibits the same behavior as regime 1, namely, throughput and scaled mean queue length are dominated by 1-server jobs served while there is an n-server job at the head of the queue, and most servers are idle. Regime 2 represents the boundary of parameters exhibiting this behavior, whereas in regime 3, we show that a distinct behavior emerges.

Phrased another way, regime 2 represents the peak of mean queue length, when controlling for utilization: $\mathbb{E}[\Delta(Y_d)]$ rises until regime 2 is reached, and falls thereafter in regime 3. See Section 10 for an exploration of this mean-queue-length peak via numerical evaluation and simulation, and its generalization into further settings.

In regime 3, the 1-server dominated regime, we primarily focus on the setting where p_n is a polynomial function of n, namely $p_n = 1/n^\alpha$ for some constant $\alpha > 1$. This setting can be seen as equivalent to the generalized Halfin–Whitt regime, which has been explored, for instance, in a load-balancing setting by Liu and Ying [20, 21].

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

Theorem 4.2. *In regime 3, which is the 1-server dominated regime, namely* $p_n = 1/n^\alpha$ *for some constant* $\alpha > 1$, *as* $p_n \to 0, n \to \infty$,

$$\mu = \mu_1 \left(n - n^{2-\alpha} (\ln n + \mu_1/\mu_n - 1 + o(1)) \right) \text{ and}$$

$$\mathbb{E}[\Delta(Y_d)] = (1 + o(1)) \frac{1}{2} n^{2-\alpha} \ln^2 n.$$

Proof. Proof is given in Section 8.2, split into Lemma 8.3 and Theorem 8.1 □

In the 1-server dominated regime, we see a different behavior than in the two previous regimes. Here, throughput nearly matches $n\mu_1$, the service rate when all n servers are serving 1-server jobs. The service that occurs when all n servers are serving 1-server jobs dominates the behavior of the system. Correspondingly, we see smaller values of $\mathbb{E}[\Delta(Y_d)]$. Specifically, in this regime, the expected number of small jobs between a pair of large jobs is still $1/p_n$, but that period is spent predominantly with a consistent $n\mu_1$ service rate, causing $\mathbb{E}[\Delta(Y_d)]$ to fall from the peak of $\frac{1}{2}\frac{n}{c}$ reached in the balanced load regime, with slower and slower growth rates for larger values of $\alpha > 1$.

We also prove more general but less detailed results in the general 1-server dominated regime, in Section 8.1 (see Lemma 8.1 and Lemma 8.2).

Section 5 provides some preliminary analysis and introduces the notation that will be used throughout the paper.

5. Preliminary analysis: Saturated system

In this section, we study the saturated MSJ system. There are three stationary distributions of importance in this system: the time-average stationary distribution P, and two distributions based on embedded Markov chains. Note that the random variable Y is distributed according to P. The first embedded DTMC that we consider is the transition-based DTMC, which updates on either arrivals or completions, giving rise to the transition-average stationary distribution π. Note that we separate each arrival as its own transition in this embedded DTMC, even if multiple arrivals occur at the same time. The second embedded DTMC is the completion-based DTMC, which only updates on completion, giving rise to the completion-average stationary distribution π^d. Note that the random variable Y_d is distributed according to π^d.

In this section, we will characterize π, P, π^d, and the throughput μ, for a given n. We will use this characterization to derive the asymptotic growth rates of μ and $\mathbb{E}[\Delta(Y_d)]$.

5.1. Transition-average stationary distribution

The state of the saturated system in the transition-based embedded DTMC can be captured by a pair (a, b) where a denotes the number of n-server jobs present, $a \in \{0, 1\}$, and b denotes the number of 1-server jobs present, $0 \le b \le n$.

There are two kinds of states in this embedded DTMC: states where jobs are completing, and states where jobs are arriving. If $a = 0$ and $b < n$, then the state is an arriving state, because room for additional jobs is available, while if $a = 1$ or $b = n$, the state is a completing state. If $a = 1$ and $b = 0$, then an n-server job is in service, while otherwise only 1-server jobs are in service.

The transition probabilities are as follows:

- From an arriving state $(0, b)$, where $b < n$, the arriving job is a 1-server job with probability p_1, resulting in a transition to $(0, b+1)$, or an n-server job with probability p_n, resulting in a transition to $(1, b)$.
- From a completing state of the form $(1, b)$, where $1 \le b < n$, b 1-server jobs are in service, resulting in a guaranteed transition to $(1, b-1)$.

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

- From the state $(1, 0)$, the n-server job is in service, resulting in a transition to $(0, 0)$.
- From the state $(0, n)$, n 1-server jobs are in service, resulting in a transition to $(0, n-1)$.

Now, we characterize the transition-average stationary distribution π:

Lemma 5.1. *The transition-average stationary distribution π can be written as:*

$$\begin{aligned}
\forall b < n-1, \pi_{0,b} &= \pi_{0,0} p_1^b, \\
\pi_{0,n-1} &= \pi_{0,0} \frac{p_1^{n-1}}{p_n}, \\
\pi_{0,n} &= \pi_{0,0} \frac{p_1^n}{p_n}, \\
\forall b < n, \pi_{1,b} &= \pi_{0,0} p_1^b.
\end{aligned}$$

Proof. Note that the arrivals occur in states $(0, b)$ for $0 \le b < n$, and job service completions occur in states $(0, n)$ and $(1, b)$ for $0 \le b < n$. Furthermore, in state $(1, 0)$, the n-server job is in service. In all other job service completion states, only 1-server jobs are in service. Then we have the following balance equations

$$\begin{aligned}
\pi_{0,0} &= \pi_{1,0}, \\
\forall b, 0 < b < n-1, \pi_{0,b} &= \pi_{0,b-1} p_1, \\
\pi_{0,n-1} &= \pi_{0,n-2} p_1 + \pi_{0,n}, \\
\pi_{0,n} &= \pi_{0,n-1} p_1, \\
\forall b, 0 \le b < n-1, \pi_{1,b} &= \pi_{1,b+1} + \pi_{0,b} p_n, \\
\pi_{1,n-1} &= \pi_{0,n-1} p_n,
\end{aligned}$$

and the result follows. □

5.2. *Time-average stationary distribution*

We now use our results from Section 5.1 to characterize the time-average stationary distribution.

Lemma 5.2. *The time-average stationary distribution P is given as*

$$\begin{aligned}
P_{1,0} &= C' \frac{1}{\mu_n}, \\
\forall b, 1 < b < n, P_{1,b} &= \frac{1}{b\mu_1} = C' \frac{1}{b\mu_1} p_1^b, \\
P_{0,n} &= C' \frac{1}{n\mu_1} \frac{p_1^n}{p_n},
\end{aligned}$$

where μ_1 and μ_n is the service completion rate of size 1 and size n jobs, respectively, and

$$C' = \left(\frac{1}{\mu_n} + \frac{1}{n\mu_1} \frac{p_1^n}{p_n} + \sum_{b=1}^{n-1} \frac{p_1^b}{b\mu_1} \right)^{-1}.$$

Proof. The time-average stationary distribution is proportional to the transition-average stationary distribution, rescaled by the time spent in each state. The only states in which time is spent are completion

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

states, and the average time per visit is the inverse of the completion rate. Thus, we define a proportionality constant C, such that $P_y = C\frac{\pi_y}{\mu_y}$, where recall that μ_y is the completion rate in state y. We also define $C' = C\pi_{0,0}$. Now, applying Lemma 5.1, we have

$$\begin{aligned} P_{1,0} &= C\pi_{1,0}\frac{1}{\mu_n} = C'\frac{1}{\mu_n}, \\ \forall b, 1 < b < n, P_{1,b} &= C\pi_{1,b}\frac{1}{b\mu_1} = C'\frac{1}{b\mu_1}p_1^b, \\ P_{0,n} &= C\pi_{0,n}\frac{1}{n\mu_1} = C'\frac{1}{n\mu_1}\frac{p_1^n}{p_n}, \end{aligned}$$

and C' can be computed by using the fact that P is a distribution:

$$1 = \sum_y P_y = C'\left(\frac{1}{\mu_n} + \frac{1}{n\mu_1}\frac{p_1^n}{p_n} + \sum_{b=1}^{n-1}\frac{p_1^b}{b\mu_1}\right)$$

□

Lemma 5.3. *The explicit expression for μ is given as*

$$\mu = \frac{1}{p_n}\left(\frac{1}{\mu_n} + \frac{1}{n\mu_1}\frac{p_1^n}{p_n} + \sum_{b=1}^{n-1}\frac{p_1^b}{b\mu_1}\right)^{-1}.$$

Proof. We apply Lemma 5.2. Note that we have

$$\begin{aligned} \mu &= P_{1,0}\mu_n + P_{0,n}n\mu_1 + \sum_{b=1}^{n-1} P_{1,b}b\mu_1 \\ &= C'(1 + \frac{p_1^n}{p_n} + \sum_{b=1}^{n-1} p_1^b) = C'(1 + \frac{p_1^n}{p_n} + \frac{p_1 - p_1^n}{1 - p_1}) = C'\frac{1}{p_n}. \end{aligned}$$

Plugging in the closed-form expression of C' gives the desired result. □

Note that semantically, C' is the rate of completion of n-server jobs, and $\frac{1}{C'}$ is the expected time per n-server job completion. We have

$$\frac{1}{C'} = \frac{1}{\mu_n} + \frac{1}{n\mu_1}\frac{p_1^n}{p_n} + \sum_{b=1}^{n-1}\frac{p_1^b}{b\mu_1}.$$

Intuitively, one can see that there are three phases to the completion of an n-server job, namely, time spent in state $(1, 0)$, time spent in state $(0, n)$, and time spent in the states $(1, b)$ for $0 < b < n$. The expected lengths of these phases are the first, second, and third terms, respectively, in the above expression.

5.3. Completion-average stationary distribution

Now we characterize the completion-average distribution π^d. Note that there are only $n + 1$ possible states in which a completion can occur, with 0 to n 1-server jobs in service.

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

Lemma 5.4. *The completion-average stationary distribution π^d is given as*

$$\forall b < n,\ \pi^d_{1,b} = p_1^b p_n,$$
$$\pi^d_{0,n} = p_1^n.$$

Proof. Note that the distribution π^d is equal to the transition-average stationary distribution π, conditioned on the saturated system being in a completion state. In particular, there exists a normalizing constant C'' such that for any completing state s,

$$\pi^d_s = C''\pi_s.$$

Using the expression for π given in Lemma 5.1, we can solve for the proportionality constant C'' as simply $\frac{p_n}{\pi_{0,0}}$, giving rise to the stated expression for π^d. □

5.4. Relative completions

We now characterize the relative completions in the saturated system, which is key to the characterization of the scaling behavior of mean queue length in the original open system (see Section 3.2 for details).

Lemma 5.5. *The explicit formula for $\mathbb{E}[\Delta(Y_d)]$ is*

$$\mathbb{E}[\Delta(Y_d)] = p_1^n\left(1 - \frac{\mu}{n\mu_1}\right)\left(n - 1 + \frac{1}{p_n} - \frac{\mu}{\mu_1}H_{n-1} - \frac{\mu}{np_n\mu_1}\right) + \sum_{i=1}^{n-1} p_1^i p_n\left(1 - \frac{\mu}{i\mu_1}\right)\left(i - \frac{\mu}{\mu_1}H_i\right).$$

where $H_i = \sum_{k=1}^{i}\frac{1}{k}$ is the i^{th} harmonic number.

Proof. For notational convenience, we will now switch to a one-dimensional state representation. More specifically, define the states of the time-average system and the completion-average embedded DTMC to be the integers $\{0, 1, 2, \ldots, n\}$, where the state is the number of 1-server jobs in the system. Note that we ignore the arrival states $(0, 0), (0, 1), \ldots, (0, n-1)$, as the system does not spend time in these states and completions cannot occur in them.

To characterize $\mathbb{E}[\Delta(Y_d)]$, we first characterize Δ up to an offset. Let $\widetilde{\Delta}(i) := \Delta(i) - \Delta(0)$, for any state i. We will use the following identity:

$$\mathbb{E}[\widetilde{\Delta}(Y)] = \mathbb{E}[\Delta(Y)] - \mathbb{E}[\Delta(0)] = -\mathbb{E}[\Delta(0)]. \tag{3}$$

The second equality holds because $\mathbb{E}[\Delta(Y)]$ is defined to be 0, as discussed in Section 3.2. Thus, we can characterize $\mathbb{E}[\Delta(Y_d)]$ as follows:

$$\begin{aligned}&\mathbb{E}[\widetilde{\Delta}(Y_d)] = \mathbb{E}[\Delta(Y_d)] - \mathbb{E}[\Delta(0)] = \mathbb{E}[\Delta(Y_d)] + \mathbb{E}[\widetilde{\Delta}(Y)]\\ \implies\ &\mathbb{E}[\Delta(Y_d)] = \mathbb{E}[\widetilde{\Delta}(Y_d)] - \mathbb{E}[\widetilde{\Delta}(Y)].\end{aligned} \tag{4}$$

Thus, it suffices to characterize $\widetilde{\Delta}$ and use (4) to characterize $\mathbb{E}[\Delta(Y_d)]$.

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

We can rewrite (4) more explicitly as

$$\mathbb{E}[\Delta(Y_d)] = \sum_{i=0}^{n} (\pi_i^d - P_i)\widetilde{\Delta}(i), \tag{5}$$

where note that $P_i = P(Y = i)$, and $\pi_i^d = P(Y_d = i)$. Recall from Lemma 5.4 that the distribution of π^d is:

$$\forall i < n, \pi_i^d = p_1^i p_n,$$
$$\pi_n^d = p_1^n.$$

Rewriting the formula for P_i from Lemma 5.2, we have

$$P_0 = \mu \frac{p_n}{\mu_n},$$
$$\forall 1 \le i < n, P_i = \mu \frac{p_1^i p_n}{i\mu_1},$$
$$P_n = \mu \frac{p_1^n}{n\mu_1}.$$

Note that $\pi_i^d = \frac{\mu}{\mu_i} P_i$, where μ_i is the instantaneous completion rate in state i. Applying (5),

$$\begin{aligned}\mathbb{E}[\Delta(Y_d)] &= (p_1^n - \mu \frac{p_1^n}{n\mu_1})\widetilde{\Delta}(n) + \sum_{i=1}^{n-1} (p_1^i p_n - \mu \frac{p_1^i p_n}{i\mu_1})\widetilde{\Delta}(i) \\ &= p_1^n (1 - \frac{\mu}{n\mu_1})\widetilde{\Delta}(n) + \sum_{i=1}^{n-1} p_1^i p_n (1 - \frac{\mu}{i\mu_1})\widetilde{\Delta}(i). \end{aligned} \tag{6}$$

Next, we calculate $\widetilde{\Delta}(i)$. Using the Poisson equation (1), which characterizes $\Delta(i)$ up to an absolute constant, we obtain

$$\begin{aligned} &\forall 1 \le i < n, \Delta(i) = 1 - \frac{\mu}{i\mu_1} + \Delta(i-1) \\ \implies &\forall 1 \le i < n, \Delta(i) - \Delta(i-1) = 1 - \frac{\mu}{i\mu_1} \\ \implies &\forall 1 \le i < n, \widetilde{\Delta}(i) = i - \frac{\mu}{\mu_1} H_i. \end{aligned}$$

Note that state n is a special case:

$$\begin{aligned} &\Delta(n) = 1 - \frac{\mu}{n\mu_1} + p_1 \Delta(n) + p_n \Delta(n-1) \\ \implies &p_n \Delta(n) = 1 - \frac{\mu}{n\mu_1} + p_n \Delta(n-1) \\ \implies &\Delta(n) - \Delta(n-1) = \frac{1}{p_n} - \frac{\mu}{n p_n \mu_1} \\ \implies &\widetilde{\Delta}(n) = n - 1 + \frac{1}{p_n} - \frac{\mu}{\mu_1} H_{n-1} - \frac{\mu}{n p_n \mu_1}. \end{aligned}$$

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

Plugging the expressions of $\widetilde{\Delta}(i)$ for $i = 1, \ldots, n$ into (6), we have

$$\mathbb{E}[\Delta(Y_d)] = p_1^n(1 - \frac{\mu}{n\mu_1})(n - 1 + \frac{1}{p_n} - \frac{\mu}{\mu_1}H_{n-1} - \frac{\mu}{np_n\mu_1}) + \sum_{i=1}^{n-1} p_1^i p_n(1 - \frac{\mu}{i\mu_1})(i - \frac{\mu}{\mu_1}H_i).$$

□

In the next three sections, we will characterize the asymptotic growth rate of the μ expression given in Lemma 5.3 and the $\mathbb{E}[\Delta(Y_d)]$ expression given in Lemma 5.5 for the three regimes discussed in Section 4.

6. Regime 1: $p_n = \omega(1/n)$: n-server dominated regime

This section focuses on regime 1, namely the n-server dominated regime, where the load of n-server jobs asymptotically dominates the load of 1-server jobs. To prove Theorem 4.1, we start with Lemma 6.1, which characterizes the throughput μ in this regime.

Lemma 6.1. *In regime 1, which is the n-server dominated regime, namely $p_n = \omega(1/n)$, as $p_n \to 0, n \to \infty$,*

$$\mu = (1 + o(1))\frac{\mu_1}{p_n \ln 1/p_n}.$$

Proof. From Lemma 5.3, we have the following expression for μ:

$$\mu = \frac{1}{p_n}\left(\frac{1}{\mu_n} + \frac{1}{n\mu_1}\frac{p_1^n}{p_n} + \sum_{b=1}^{n-1}\frac{p_1^b}{b\mu_1}\right)^{-1}.$$

Note that as $n \to \infty$, $p_1^n = (1 - p_n)^n$ can only converge to a nonzero constant if $p_n = \theta(1/n)$. In this setting, $p_n = \omega(1/n)$, so $p_1^n = o(1)$. Note that $\frac{p_1^n}{np_n} = \frac{o(1)}{\omega(1)} = o(1)$. Then

$$\begin{aligned}\mu &= \frac{1}{p_n}\left(\frac{1}{\mu_n} + \frac{1}{n\mu_1}\frac{p_1^n}{p_n} + \sum_{b=1}^{n-1}\frac{p_1^b}{b\mu_1}\right)^{-1} \\ &= \frac{1}{p_n}\left(\frac{1}{\mu_n} + o(1) + \sum_{b=1}^{n-1}\frac{p_1^b}{b\mu_1}\right)^{-1} + o(1) \\ &= (1 + o(1))\frac{1}{p_n}\left(\frac{1}{\mu_n} + \sum_{b=1}^{n-1}\frac{p_1^b}{b\mu_1}\right)^{-1}.\end{aligned}$$

Here we use the fact that $\frac{1}{\mu_n}$ is a constant, so the $o(1)$ term has a negligible multiplicative effect. Moreover, we will show that $\sum_{b=1}^{n-1}\frac{p_1^b}{b\mu_1} = \theta(\ln 1/p_n)$, so $\frac{1}{\mu_n}$ is also multiplicatively negligible.

$$(1 + o(1))\frac{1}{p_n}\left(\frac{1}{\mu_n} + \sum_{b=1}^{n-1}\frac{p_1^b}{b\mu_1}\right)^{-1} = (1 + o(1))\frac{\mu_1}{p_n}\left(\sum_{b=1}^{n-1}\frac{p_1^b}{b}\right)^{-1}. \quad (7)$$

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

We next focus on $\sum_{b=1}^{n-1} \frac{p_1^b}{b}$. First, we change from a finite summation to an infinite summation, and second, we change to an integral.

$$\sum_{b=1}^{n-1} \frac{p_1^b}{b} = (1+o(1)) \sum_{b=1}^{\infty} \frac{p_1^b}{b} = (1+o(1)) \int_{b=1}^{\infty} \frac{p_1^b}{b} db. \tag{8}$$

To justify the change to an infinite summation, note that $p_n = \omega(1/n)$ and hence, $1/p_n = o(n)$. At index $b = \lfloor 1/p_n \rfloor$, $p_1^b \sim 1/e$. In particular, all indices from $\lfloor 1/p_n \rfloor$ to $\lfloor 2/p_n \rfloor$ sum to at most $1/(p_n e)$:

$$\sum_{b=\lfloor 1/p_n \rfloor}^{\lfloor 2/p_n \rfloor} \frac{p_1^b}{b} \leq \frac{1}{p_n} \frac{p_n}{e} = \frac{1}{e}.$$

By a similar argument, indices from $\lfloor 2/p_n \rfloor$ to $\lfloor 3/p_n \rfloor$ sum to at most $\frac{1}{2e^2}$, and in general for any constant k,

$$\sum_{b=\lfloor k/p_n \rfloor}^{\lfloor (k+1)/p_n \rfloor} \frac{p_1^b}{b} \leq \frac{1}{p_n} \frac{k p_n}{e^k} = \frac{k}{e^k}.$$

Summing this infinite series, we find that

$$\sum_{b=\lfloor 1/p_n \rfloor}^{\infty} \frac{p_1^b}{b} \leq \sum_{k=1}^{\infty} \frac{k}{e^k} = \frac{e}{(e-1)^2}.$$

Thus, the sum over indices $b \geq \lfloor 1/p_n \rfloor$ is $O(1)$. As a result, the sum over indices $b \geq n$ must also be $O(1)$, as n is much larger than $1/p_n$. This $O(1)$ difference is negligible compared to the $\theta(\ln 1/p_n)$ growth rate of the integral, which we will demonstrate.

To justify the transformation from the summation to the integral, note that the difference can be rewritten as follows:

$$\sum_{b=1}^{\infty} \frac{p_1^b}{b} - \int_{b=1}^{\infty} \frac{p_1^b}{b} db = \sum_{b=1}^{\infty} \left(\frac{p_1^b}{b} - \int_{x=b}^{b+1} \frac{p_1^x}{x} dx \right).$$

Next, note that $\frac{p_1^b}{b}$ is decreasing as a function of b, so we can upper bound the difference:

$$\sum_{b=1}^{\infty} \left(\frac{p_1^b}{b} - \int_{x=b}^{b+1} \frac{p_1^x}{x} dx \right) \leq \sum_{b=1}^{\infty} \left(\frac{p_1^b}{b} - \frac{p_1^{b+1}}{b+1} \right).$$

Now, we can upper bound the term inside the summation:

$$\frac{p_1^b}{b} - \frac{p_1^{b+1}}{b+1} \leq \frac{p_1^b}{b} - \frac{p_1^{b+1}}{b} \leq p_1^b - p_1^{b+1}$$

$$\implies \sum_{b=1}^{\infty} \left(\frac{p_1^b}{b} - \frac{p_1^{b+1}}{b+1} \right) \leq \sum_{b=1}^{\infty} \left(p_1^b - p_1^{b+1} \right) = p_1 < 1.$$

Thus, this difference is negligible compared to the $\theta(\ln 1/p_n)$ growth rate of the integral, which we will demonstrate.

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

We now continue by simplifying the integral from (8),

$$\int_{b=1}^{\infty} \frac{p_1^b}{b} db = Gamma[0, -\ln(p_1)] = Gamma[0, -\ln(1-p_n)],$$

where Gamma is the incomplete gamma function. Note that

$$\lim_{p_n \to 0} Gamma[0, -\ln(1-p_n)] + \ln p_n = \gamma,$$

where $\gamma \approx 0.57726$ is the Euler–Mascheroni constant. Thus,

$$\int_{b=1}^{\infty} \frac{p_1^b}{b} db = -\ln p_n + O(1) = \ln \frac{1}{p_n} + O(1). \tag{9}$$

We have demonstrated the desired growth rate for the summation, and plugging that growth rate into (7) completes the proof. □

We next characterize the limiting scaled response time $\mathbb{E}[\Delta(Y_d)]$, given in (2), in this regime, to complete the proof of Theorem 4.1.

Theorem 6.1. *In regime 1, which is the n-server dominated regime, namely $p_n = \omega(1/n)$, as $p_n \to 0, n \to \infty$, then*

$$\mathbb{E}[\Delta(Y_d)] = (1 + o(1)) \frac{1}{2p_n}.$$

Proof. First, we have the explicit formula for $\mathbb{E}[\Delta(Y_d)]$, from Lemma 5.5,

$$\begin{aligned} \mathbb{E}[\Delta(Y_d)] = {} & p_1^n \left(1 - \frac{\mu}{n\mu_1}\right)\left(n - 1 + \frac{1}{p_n} - \frac{\mu}{\mu_1} H_{n-1} - \frac{\mu}{np_n\mu_1}\right) \\ & + \sum_{i=1}^{n-1} p_1^i p_n \left(1 - \frac{\mu}{i\mu_1}\right)\left(i - \frac{\mu}{\mu_1} H_i\right). \end{aligned} \tag{10}$$

We first rewrite expression (10) in terms of the difference of two infinite summations:

$$\mathbb{E}[\Delta(Y_d)] = p_1^n \left(1 - \frac{\mu}{n\mu_1}\right)\left(n - 1 + \frac{1}{p_n} - \frac{\mu}{\mu_1} H_{n-1} - \frac{\mu}{np_n\mu_1}\right) \tag{11}$$

$$+ \sum_{i=1}^{\infty} p_1^i p_n \left(1 - \frac{\mu}{i\mu_1}\right)\left(i - \frac{\mu}{\mu_1} H_i\right) \tag{12}$$

$$- \sum_{i=n}^{\infty} p_1^i p_n \left(1 - \frac{\mu}{i\mu_1}\right)\left(i - \frac{\mu}{\mu_1} H_i\right). \tag{13}$$

We will show that (11) and (13) are each $o(1/p_n)$, and thus negligible. We will then split up (12) into four terms:

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

$$p_n \sum_{i=1}^{\infty} p_1^i i \tag{14}$$

$$-\frac{\mu p_n}{\mu_1} \sum_{i=1}^{\infty} p_1^i \tag{15}$$

$$-\frac{\mu p_n}{\mu_1} \sum_{i=1}^{\infty} p_1^i H_i \tag{16}$$

$$+\frac{\mu^2 p_n}{\mu_1^2} \sum_{i=1}^{\infty} p_1^i \frac{1}{i} H_i \tag{17}$$

and show that the asymptotic growth rates of these terms are:

(1) $p_1^i i$ term (14): $\frac{1}{p_n} + o(1/p_n)$.
(2) p_1^i term (15): $o(1/p_n)$.
(3) $p_1^i H_i$ term (16): $-\frac{1}{p_n} + o(1/p_n)$.
(4) $p_1^i H_i / i$ term (17): $\frac{1}{2p_n} + o(1/p_n)$.

These growth rates sum to $\frac{1}{2p_n} + o(1/p_n)$, as desired.

First, we will show that (11) is negligible:

$$\begin{aligned}
&p_1^n(1 - \frac{\mu}{n\mu_1})(n - 1 + \frac{1}{p_n} - \frac{\mu}{\mu_1} H_{n-1} - \frac{\mu}{np_n\mu_1}) \\
&= \theta(p_1^n(n - \frac{\mu}{np_n})) \\
&= \theta(p_1^n(n - o(\frac{1}{p_n \log n}))) \\
&= \theta(np_1^n).
\end{aligned}$$

Next we verify that $np_1^n = o(1/p_n)$ which is equivalent to showing that $np_n = o(1/p_1^n) = o((\frac{1}{1-p_n})^n)$. Note that $\frac{1}{1-p_n} = 1 + p_n + p_n^2 + \ldots \geq 1 + p_n$. Thus,

$$\left(\frac{1}{1 - p_n}\right)^n \geq (1 + p_n)^n \geq p_n^2 \frac{n(n-1)}{2} = \theta(p_n^2 n^2).$$

Note that in regime 1, $np_n = \omega(1)$. Thus, $p_n^2 n^2 = \omega(np_n)$, and equivalently $np_n = o(1/p_1^n)$, as desired and (11) is negligible.

Next, we will show that (13) is negligible:

$$\sum_{i=n}^{\infty} p_1^i p_n (1 - \frac{\mu}{i\mu_1})(i - \frac{\mu}{\mu_1} H_i).$$

Let us start by bounding the term corresponding to $i = n$ of (13), namely,

$$p_1^n p_n (1 - \frac{\mu}{n\mu_1})(n - \frac{\mu}{\mu_1} H_n).$$

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

Recall that we have shown that $p_1^n = o(\frac{1}{np_n})$. Then

$$\begin{aligned}
&p_1^n p_n (1 - \frac{\mu}{n\mu_1})(n - \frac{\mu}{\mu_1} H_n) \\
&= o(1/n)(1 - \frac{\mu}{n\mu_1})(n - \frac{\mu}{\mu_1} H_n) \\
&= o(1/n)(1 - \frac{1}{np_n \log 1/p_n})(n - \frac{\log n}{p_n \log 1/p_n}) \\
&= o(1/n)(n - \frac{\log n}{p_n \log 1/p_n}) \\
&= o(1) - o(\frac{\log n}{np_n \log 1/p_n}).
\end{aligned}$$

Note that the function $x \ln 1/x$ is increasing in x in the interval $(0, 1/e)$. Recall that $p_n = \omega(1/n)$. Thus, $p_n \log 1/p_n = \Omega(1/n \log n)$. We therefore find that

$$\frac{\log n}{np_n \log 1/p_n} = O(\frac{\log n}{n(1/n \log n)}) = O(1).$$

As a result, we find that the term corresponding to $i = n$ of (13) is $o(1)$.

Next, consider the first $\lfloor 1/p_n \rfloor$ terms of (13), ranging from $i = n$ to $i = n + \lfloor 1/p_n \rfloor$. By the same argument, each of these terms is $o(1)$, and the sum is bounded accordingly:

$$\sum_{i=n}^{i=n+\lfloor 1/p_n \rfloor} p_1^i p_n (1 - \frac{\mu}{i\mu_1})(i - \frac{\mu}{\mu_1} H_i) = o(1/p_n).$$

Call these $\lfloor 1/p_n \rfloor$ terms the first *block* of terms, and similarly, define the kth block of terms:

$$b(k) := \sum_{i=n+\lfloor (k-1)/p_n \rfloor}^{i=n+\lfloor k/p_n \rfloor} p_1^i p_n (1 - \frac{\mu}{i\mu_1})(i - \frac{\mu}{\mu_1} H_i).$$

We can bound the kth block's ratio with respect to the first term of (13). The p_1^i term has shrunk by a factor of p_1^{k/p_n}, which is $(1 + o(1))e^{-k}$. The $1 - \frac{\mu}{i\mu_1}$ term differs negligibly from 1 and the $i - \frac{\mu}{\mu_1} H_i$ term is $O(i)$ as shown above. As a result, the ratio between this term in the kth block and the first block is at most k. Thus, we can bound the ratio of the blocks:

$$\begin{aligned}
&b(k)/b(1) \le e^{-k}(k+1) \\
&\sum_{k=1}^{\infty} b(k) \le b(1) \sum_{k=1}^{\infty} e^{-k}(k+1) \le 3b(1) = o(1/p_n).
\end{aligned}$$

As a result, the infinite sum in (13) is $o(1/p_n)$, as desired.

Now, we are ready to characterize the four subterms of the main term (12), starting with (14), confirming Item 1:

$$p_n \sum_{i=1}^{\infty} p_1^i i = p_n \frac{p_1}{p_n^2} = \frac{p_1}{p_n} = \frac{1}{p_n} + o(1/p_n).$$

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

Next, we have (15), confirming Item 2:

$$-\frac{\mu p_n}{\mu_1}\sum_{i=1}^{\infty} p_1^i = -\frac{\mu p_n}{\mu_1}\frac{p_1}{p_n} = -\frac{\mu p_1}{\mu_1} = \theta(\mu) = \theta(\frac{1}{p_n \log(1/p_n)}) = o(1/p_n).$$

Next, we consider (16), in which we replace H_i with $\ln i + O(1)$:

$$-\frac{\mu p_n}{\mu_1}\sum_{i=1}^{\infty} p_1^i H_i = -\frac{\mu p_n}{\mu_1}\sum_{i=1}^{\infty} p_1^i(\ln i + O(1)).$$

We next switch from a summation to an integral, and bound the difference from this change:

$$\sum_{i=1}^{\infty} p_1^i \ln i - \int_{i=1}^{\infty} p_1^i \ln i di = \sum_{i=1}^{\infty}\left(p_1^i \ln i - \int_{j=i}^{i+1} p_1^j \ln j dj\right).$$

Because p_1^j is decreasing in j and $\ln j$ is increasing in j, we can upper bound this difference as

$$\sum_{i=1}^{\infty}\left(p_1^i \ln i - \int_{j=i}^{i+1} p_1^j \ln j dj\right) \le \sum_{i=1}^{\infty}\left(p_1^i \ln i - p_1^{i+1}\ln i\right) = \sum_{i=1}^{\infty}\left((p_1^i - p_1^{i+1})\ln i\right)$$
$$= (1-p_1)\sum_{i=1}^{\infty}\left(p_1^i \ln i\right).$$

We have shown that the difference between the summation and the integral is equal to the summation multiplied by a $1 - p_1 = o(1)$ term. We can thus relate (16) to the integral as follows:

$$-\frac{\mu p_n}{\mu_1}\sum_{i=1}^{\infty} p_1^i H_i = -\frac{\mu p_n}{\mu_1}\sum_{i=1}^{\infty} p_1^i(\ln i + O(1)) = -\frac{\mu p_n}{\mu_1}\left(O(\frac{1}{p_n}) + \sum_{i=1}^{\infty} p_1^i \ln i\right)$$
$$= -\frac{\mu p_n}{\mu_1}\left(O(\frac{1}{p_n}) + (1+o(1))\int_{i=1}^{\infty} p_1^i \ln i di\right).$$

Now, let us focus on the integral:

$$\int_{i=1}^{\infty} p_1^i \ln(i) di = \frac{LogIntegral[p_1]}{\ln(p_1)},$$

where LogIntegral is the logarithmic integral function $\int_0^x \frac{dt}{\ln t}$.

We can take the following limits:

$$\lim_{p_n\to 0^+} LogIntegral[1-p_n] - \ln p_n = \gamma \implies LogIntegral[1-p_n] = \ln p_n + O(1) \tag{18}$$

$$\lim_{p_n\to 0^+} \frac{\ln(1-p_n)}{p_n} = -1 \implies \ln(1-p_n) = -p_n + o(p_n) \tag{19}$$

$$\implies \frac{LogIntegral[p_1]}{\ln(p_1)} = \frac{\ln p_n + O(1)}{-p_n + o(p_n)} = -(1+o(1))\frac{\ln p_n}{p_n}.$$

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

Combining it all together, we get

$$-\frac{\mu p_n}{\mu_1}\sum_{i=1}^{\infty} p_1^i H_i = -\frac{\mu p_n}{\mu_1}\left(-(1+o(1))\frac{\ln p_n}{p_n} + O(1/p_n)\right)$$
$$= (1+o(1))\frac{1}{\ln(1/p_n)}\frac{\ln(p_n)}{p_n}$$
$$= -(1+o(1))\frac{1}{p_n}.$$

This confirms Item 3, the specified growth rate of (16).

We now finish with (17). We start by transforming H_i to $\ln i$ and the summation to an integral in the same fashion as for (15), with the same error bound:

$$\frac{\mu^2 p_n}{\mu_1^2}\sum_{i=1}^{\infty} p_1^i \frac{1}{i} H_i = \frac{\mu^2 p_n}{\mu_1^2}\left((1+o(1))\int_{i=1}^{\infty} p_1^i \frac{1}{i}\ln i di + O(1/p_n)\right).$$

We evaluate this integral using a symbolic algebraic package (specifically, Mathematica [26]) and obtain

$$\int_{i=1}^{\infty} p_1^i \ln(i)/i di = \frac{1}{2}\ln(1/p_n)^2 + O(\ln(1/p_n)). \tag{20}$$

Putting it all together:

$$p_n\frac{\mu^2}{\mu_1^2}\sum_{i=1}^{\infty} p_1^i \frac{H_i}{i} = p_n\frac{\mu^2}{\mu_1^2}\left(\frac{1}{2}\ln(1/p_n)^2 + O(\ln(1/p_n))\right)$$
$$= (1+o(1))\frac{1}{p_n \ln(1/p_n)^2}\frac{1}{2}\ln(1/p_n)^2$$
$$= (1+o(1))\frac{1}{2}\frac{1}{p_n}.$$

This confirms Item 4, the desired growth rate of (17) and completes the proof. □

7. Regime 2: $p_n = c/n$: Balanced load

This section focuses on regime 2, namely, the balanced regime, where the load of n-server jobs and 1-server jobs remain in fixed proportion, for a general proportionality constant c. To prove the result in Theorem 4.1 for regime 2, we start by characterizing the throughput μ in this regime. Lemma 7.1 proves the first equation of Theorem 4.1.

Lemma 7.1. *In regime 2, which is the balanced regime, namely $p_n = c/n$, for some positive constant c, as $p_n \to 0, n \to \infty$,*

$$\mu = (1+o(1))\frac{\mu_1}{p_n \ln 1/p_n}.$$

Proof. From Lemma 5.3, we have the following expression for μ:

$$\mu = \frac{1}{p_n}\left(\frac{1}{\mu_n} + \frac{1}{n\mu_1}\frac{p_1^n}{p_n} + \sum_{b=1}^{n-1}\frac{p_1^b}{b\mu_1}\right)^{-1}.$$

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

Note that as $n \to \infty$, $p_1^n = (1 - p_n)^n = (1 + o(1))e^{-c}$. Then

$$\mu = \frac{1}{p_n}\left(\frac{1}{\mu_n} + (1 + o(1))\frac{e^{-c}}{c\mu_1} + \sum_{b=1}^{n-1}\frac{p_1^b}{b\mu_1}\right)^{-1}$$
$$= \frac{1}{p_n}\left(O(1) + \sum_{b=1}^{n-1}\frac{p_1^b}{b\mu_1}\right)^{-1}. \tag{21}$$

We will show that $\sum_{b=1}^{n-1}\frac{p_1^b}{b} = \omega(1)$, and hence that the $O(1)$ term is negligible.

To do so, we first change from a finite summation to an infinite summation, and then change to an integral:

$$\sum_{b=1}^{n-1}\frac{p_1^b}{b} = (1 + o(1))\sum_{b=1}^{\infty}\frac{p_1^b}{b} = (1 + o(1))\int_{b=1}^{\infty}\frac{p_1^b}{b}db.$$

To justify the change to an infinite summation, note that $p_n = c/n$ and hence, $1/p_n = n/c$. At index $b = \lfloor 1/p_n \rfloor$, $p_1^b \approx 1/e$, and $p_1^b/b \approx p_n/e$. All indices from $\lfloor 1/p_n \rfloor$ to $\lfloor 2/p_n \rfloor$ sum to at most $1/e$, and indices from $\lfloor 2/p_n \rfloor$ to $\lfloor 3/p_n \rfloor$ sum to at most $2/e^2$. In general,

$$\sum_{b=\lfloor k/p_n \rfloor}^{\lfloor (k+1)/p_n \rfloor}\frac{p_1^b}{b} \le \frac{k}{e^k}$$
$$\implies \sum_{b=\lfloor 1/p_n \rfloor}^{\infty}\frac{p_1^b}{b} \le \sum_{k=1}^{\infty}\frac{k}{e^k} = \frac{e}{(e-1)^2} = O(1).$$

As $1/p_n = \theta(n)$, we similarly find that the sum from index n to ∞ is $O(1)$. As stated above, we will show that this $O(1)$ term is negligible.

The transformation from the summation to the integral holds for the same reason as given in Lemma 6.1, again with a negligible $O(1)$ difference.

Next, as shown in (9), we have

$$\int_{b=1}^{\infty}\frac{p_1^b}{b}db = -\ln p_n + O(1).$$

Note that the above integral is thus $\omega(1)$, so the prior $O(p_n)$ term is negligible as specified above. Substituting everything into (21), we have

$$\mu = \frac{1}{p_n}(O(1) - \ln p_n)^{-1} = (1 + o(1))\frac{1}{p_n \ln 1/p_n},$$

as desired. □

We next characterize the limiting scaled response time $\mathbb{E}[\Delta(Y_d)]$ for regime 2 to complete the proof of Theorem 4.1 in the balanced regime.

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

Theorem 7.1. *In regime 2, which is the balanced regime, namely $p_n = c/n$, for some positive constant c, as $p_n \to 0, n \to \infty$,*

$$\mathbb{E}[\Delta(Y_d)] = (1 + o(1))\frac{1}{2p_n}.$$

Proof. We start with the formula for $\mathbb{E}[\Delta(Y_d)]$, from Lemma 5.5, which we split into five terms:

$$p_1^n(1 - \frac{\mu}{n\mu_1})(n - 1 + \frac{1}{p_n} - \frac{\mu}{\mu_1}H_{n-1} - \frac{\mu}{np_n\mu_1}) \tag{22}$$

$$+p_n \sum_{i=1}^{n-1} p_1^i i \tag{23}$$

$$-\frac{\mu p_n}{\mu_1} \sum_{i=1}^{n-1} p_1^i \tag{24}$$

$$-\frac{\mu p_n}{\mu_1} \sum_{i=1}^{n-1} p_1^i H_i \tag{25}$$

$$+\frac{\mu^2 p_n}{\mu_1^2} \sum_{i=1}^{n-1} p_1^i \frac{1}{i} H_i \tag{26}$$

We will demonstrate the following asymptotic growth rates for each of these terms:

(1) Non-summation term (22): $\frac{ce^{-c}}{p_n} + o(1/p_n)$.
(2) $p_1^i i$ term (23): $\frac{1-e^{-c}(1+c)}{p_n} + o(1/p_n)$.
(3) p_1^i term (24): $o(1/p_n)$.
(4) $p_1^i H_i$ term (25): $-\frac{1-e^{-c}}{p_n} + o(1/p_n)$.
(5) $p_1^i H_i/i$ term (26): $\frac{1}{2p_n} + o(1/p_n)$.

Note that the non-summation term (22) is now significant, which was not the case in the n-server dominated regime in Section 6.

These growth rates sum to $\frac{1}{2p_n} + o(1/p_n)$, as desired. Furthermore, c does not appear in the result, as its contributions cancel out.

Now, to verify Items 1 to 5, we start with the non-summation term (22), and simplify. First, we note that $p_1^n \to e^{-c}$:

$$\begin{aligned} &p_1^n(1 - \frac{\mu}{n\mu_1})(n - 1 + \frac{1}{p_n} - \frac{\mu}{\mu_1}H_{n-1} - \frac{\mu}{np_n\mu_1}) \\ &= (1 + o(1))e^{-c}(1 - \frac{\mu}{n\mu_1})(n - 1 + \frac{1}{p_n} - \frac{\mu}{\mu_1}H_{n-1} - \frac{\mu}{np_n\mu_1}). \end{aligned}$$

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

Next, we remove negligible terms. Note that μ/n is negligible, because $\mu = O(\frac{1}{p_n \ln(1/p_n)})$, and $n = O(1/p_n)$.

$$(1+o(1))e^{-c}(1-\frac{\mu}{n\mu_1})(n-1+\frac{1}{p_n}-\frac{\mu}{\mu_1}H_{n-1}-\frac{\mu}{np_n\mu_1})$$
$$=(1+o(1))e^{-c}(\frac{c+1}{p_n}-\frac{\mu}{\mu_1}H_{n-1}).$$

We now apply Lemma 7.1, and use the facts that that $|H_{n-1}-\ln n| \le 1$ by standard harmonic series bounds, and $\ln n - \ln 1/p_n = \ln c$. Thus,

$$(1+o(1))e^{-c}(\frac{c+1}{p_n}-\frac{\mu}{\mu_1}H_{n-1}) = (1+o(1))e^{-c}(\frac{c}{p_n}).$$

This verifies Item 1, the claimed asymptotic growth rate for (22).

Next, we examine Item 2, the claimed asymptotic growth rate for (23). This summation has a simple closed form:

$$p_n\sum_{i=1}^{n-1}p_1^i i = p_n\frac{p_1+p_1^{n+1}(-1-n+np_1)}{p_n^2} = (1+o(1))\frac{1+e^{-c}(-1-n(1-p_1))}{p_n}$$
$$=(1+o(1))\frac{1-e^{-c}(1+c)}{p_n}.$$

Next, we examine Item 3, the claimed asymptotic growth rate for (24). This summation also has a simple closed form:

$$\sum_{i=1}^{n-1}p_1^i = \frac{1-p_1^n}{1-p_1} = \frac{1-e^{-c}}{p_n}$$
$$\implies -\frac{\mu p_n}{\mu_1}\sum_{i=1}^{n-1}p_1^i = -(1+o(1))\frac{1-e^{-c}}{p_n\ln(1/p_n)} = -(1+o(1))\frac{1-e^{-c}}{p_n\ln(1/p_n)}$$

Note that $\frac{1}{p_n\ln(1/p_n)} = o(1/p_n)$, confirming the growth rate given in Item 3.

Next, we examine Item 4, the claimed asymptotic growth rate for (25):

$$-\frac{\mu p_n}{\mu_1}\sum_{i=1}^{n-1}p_1^i H_i.$$

We claim that replacing H_i by $\ln(i)$ has a negligible impact on the sum, as does transforming the sum into an integral. In particular, $H_i - \ln(i) \in [0,1]$. Similar to the argument in Lemma 6.1, replacing the sum by an integral changes the growth rate by at most a bounded constant per term. Both have at most an additive impact per term, resulting in net change on the order of $\frac{\mu p_n}{\mu_1}\sum_{i=1}^{n-1}p_1^i$, which matches the formula from (24), and is similarly negligible. As a result, we have an integral that can be evaluated explicitly. Thus,

$$\sum_{i=1}^{n-1}p_1^i H_i = (1+o(1))\int_{i=1}^{n}p_1^i\ln(i)di$$
$$\int_{i=1}^{n}p_1^i\ln i di = \frac{-ExpIntegralEi[n\ln p_1]+p_1^n\ln n+LogIntegral[p_1]}{\ln p_1}, \tag{27}$$

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

where $ExpIntegralEi[x] := -\int_{-x}^{\infty} \frac{e^{-t}}{t} dt$ is the exponential integral function.

Taking limits, we find that

$$\lim_{n\to\infty} ExpIntegralEi[n \ln(1 - c/n)] = ExpIntegralEi[c]$$
$$\text{equivalently, } ExpIntegralEi[n \ln p_1] = \theta(1)$$

Recall the asymptotic expression from Section 6, in equations (18) and (19). There, we used the following limits:

$$LogIntegral[p_1] = \ln p_n + \theta(1) = -\ln n + \theta(1),$$
$$\ln p_1 = -p_n(1 + o(1)).$$

It is also simple to calculate that:

$$p_1^n \ln n = e^{-c} \ln n + o(\ln n),$$
$$\frac{\mu p_n}{\mu_1} = (1 + o(1)) \frac{1}{\ln 1/p_n} = \frac{1}{\ln n} + o(1/\ln n).$$

As a result,

$$-ExpIntegralEi[n \ln p_1] + p_1^n \ln n + LogIntegral[p_1] = (e^{-c} - 1) \ln n + o(\ln n)$$
$$\implies \sum_{i=1}^{n-1} p_1^i H_i = (1 + o(1)) \frac{(1 - e^{-c}) \ln n}{p_n}$$
$$\implies \frac{\mu p_n}{\mu_1} \sum_{i=1}^{n-1} p_1^i H_i = \frac{1 - e^{-c}}{p_n} + o(1/p_n).$$

This confirms Item 4, claimed asymptotic growth rate for (25).

Finally, we have Item 5, the claimed asymptotic growth rate for (26):

$$\frac{\mu^2 p_n}{\mu_1^2} \sum_{i=1}^{n-1} p_1^i \frac{1}{i} H_i.$$

Let us focus on the leading multiplier first:

$$\frac{\mu^2 p_n}{\mu_1^2} = \frac{1}{p_n \ln(1/p_n)^2}.$$

Now, we turn to the summation. Note that a component of the summation must be of the order of $\ln(1/p_n)^2$ to be non-negligible.

We perform the same transformation as in (27), replacing the summation by an integral and the H_i with $\ln i$, with the same justification:

$$\sum_{i=1}^{n-1} p_1^i \frac{H_i}{i} \approx \int_{i=1}^{n} p_1^i \frac{H_i}{i} di = \int_{i=1}^{c/p_n} (1 - p_n)^i \frac{H_i}{i} di \approx \int_{i=1}^{c/p_n} (1 - p_n)^i \frac{\ln i}{i} di.$$

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

Note that this differs from (20) because the upper limit of the integral is $n = c/p_n$, not ∞. Nonetheless, using Mathematica [26], we obtain the same asymptotics:

$$\int_{i=1}^{c/p_n} (1-p_n)^i \frac{\ln i}{i} di = \frac{1}{2}\ln(1/p_n)^2 + \theta(\log(1/p_n))$$

Thus the entire term satisfies the claim:

$$\frac{\mu^2 p_n}{\mu_1^2} \sum_{i=1}^{n-1} p_1^i \frac{1}{i} H_i = \frac{1}{2p_n} + o(1/p_n).$$

All five claims, namely Items 1 to 5, are verified, completing the proof. □

8. Regime 3: $p_n = o(1/n)$: Load dominated by 1-server jobs

This section focuses on regime 3, namely, the 1-server dominated regime, where the load of 1-server jobs asymptotically overwhelms the load of n-server jobs. We start in Section 8.1 in the general setting, covering all asymptotic relationships for $p_n = o(1/n)$. However, in Section 8.2, we will primarily focus on a setting of polynomial scaling, where $p_n = 1/n^\alpha, \alpha > 1$.

8.1. Regime 3, general setting: Any growth rate $p_n = o(1/n)$

This setting contains a wide variety of behaviors, so while we characterize asymptotic behavior of the throughput μ in significant detail, our result on scaled mean queue length, $\mathbb{E}[\Delta(Y_d)]$, is less tight. Our tighter results are in the more specific setting of Section 8.2.

We start by characterizing the throughput μ in this regime.

Lemma 8.1. *If* $p_n = o(1/n\ln n)$, *then*

$$\mu = (1+o(1))n\mu_1.$$

On the other hand, if $p_n = \omega(1/(n\ln n))$ *and* $p_n = o(1/n)$, *then*

$$\mu = (1+o(1))\frac{\mu_1}{p_n \ln n}.$$

Proof. We start with Lemma 5.3:

$$\begin{aligned}\mu &= \frac{1}{p_n}\left(\frac{1}{\mu_n} + \frac{1}{n\mu_1}\frac{p_1^n}{p_n} + \sum_{b=1}^{n-1}\frac{p_1^b}{b\mu_1}\right)^{-1} \\ &= \frac{1}{p_n}\left(\frac{1}{\mu_n} + (1+o(1))\frac{1}{n\mu_1 p_n} + \sum_{b=1}^{n-1}\frac{p_1^b}{b\mu_1}\right)^{-1}.\end{aligned}$$

This step holds because $p_n = o(1/n)$, so $p_1^n = 1 + o(1)$. As a result, we can similarly simplify the summation term—note that $p_1^n \le p_1^b \le 1$. Furthermore, $np_n = o(1)$, so the $1/\mu_n$ term is negligible.

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

Then

$$\mu = \frac{1+o(1)}{p_n}\left(\frac{1}{n\mu_1 p_n} + \sum_{b=1}^{n-1}\frac{1}{b\mu_1}\right)^{-1} = \frac{1+o(1)}{p_n}\left(\frac{1}{n\mu_1 p_n} + \frac{H_{n-1}}{\mu_1}\right)^{-1}$$
$$= \frac{1+o(1)}{p_n}\left(\frac{1}{n\mu_1 p_n} + \frac{\ln n}{\mu_1}\right)^{-1}$$

Either of these two terms can be larger in asymptotic limit, depending on the asymptotic behavior of p_n. If $p_n = o(1/(n\ln n))$, then $\frac{1}{n\mu_1 p_n}$ is larger. If $p_n = \omega(1/(n\ln n))$, then $\frac{\ln n}{\mu_1}$ is larger. In the former case,

$$\mu = \frac{1+o(1)}{p_n}\left(\frac{1}{n\mu_1 p_n}\right)^{-1} = (1+o(1))\frac{n\mu_1 p_n}{p_n} = (1+o(1))n\mu_1.$$

In the latter case,

$$\mu = \frac{1+o(1)}{p_n}\left(\frac{\ln n}{\mu_1}\right)^{-1} = (1+o(1))\frac{\mu_1}{p_n \ln n}.$$

Finally, if $p_n = \theta(1/(n\ln n))$, then both terms $\frac{1}{n\mu_1 p_n}$ and $\frac{\ln n}{\mu_1}$ are within a constant factor of each other in the asymptotic limit, and both contribute to the asymptotic behavior of μ. □

We now switch to $\mathbb{E}[\Delta(Y_d)]$, where we only focus on the case where $p_n = o(1/(n\ln n))$, and only provide an upper bound on the asymptotic growth of $\mathbb{E}[\Delta(Y_d)]$. Our stronger results for $\mathbb{E}[\Delta(Y_d)]$ are in Section 8.2.

Lemma 8.2. *If $p_n = o(1/(n\ln n))$, then*

$$\mathbb{E}[\Delta(Y_d)] = o(1/p_n).$$

Proof. We start with the expression of $\mathbb{E}[\Delta(Y_d)]$ from Lemma 5.5, which we separate into two terms.

$$\mathbb{E}[\Delta(Y_d)] = p_1^n(1-\frac{\mu}{n\mu_1})(n-1+\frac{1}{p_n}-\frac{\mu}{\mu_1}H_{n-1}-\frac{\mu}{np_n\mu_1}) \tag{28}$$

$$+\sum_{i=1}^{n-1} p_1^i p_n(1-\frac{\mu}{i\mu_1})(i-\frac{\mu}{\mu_1}H_i). \tag{29}$$

First consider (28). Note that by Lemma 8.1, in the setting where $p_n = o(1/(n\ln n))$, we have $\mu = (1+o(1))n\mu_1$, and hence $1-\frac{\mu}{n\mu_1} = o(1)$. Furthermore, $p_1^n \le 1$. Thus,

$$p_1^n(1-\frac{\mu}{n\mu_1})(n-1+\frac{1}{p_n}-\frac{\mu}{\mu_1}H_{n-1}-\frac{\mu}{np_n\mu_1})$$
$$= o(1)(n-1+\frac{1}{p_n}-\frac{\mu}{\mu_1}H_{n-1}-\frac{\mu}{np_n\mu_1}).$$

Note that $\frac{1}{p_n} = \omega(n\ln n)$, so the dominant terms are the two terms involving $\frac{1}{p_n}$.

$$(n-1+\frac{1}{p_n}-\frac{\mu}{\mu_1}H_{n-1}-\frac{\mu}{np_n\mu_1}) = \frac{1-\frac{\mu}{n\mu_1}}{p_n} + O(n\ln n) = o(1/p_n).$$

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

Note that we only derive an upper bound on the growth rate of (28), namely $o(1/p_n)$. We do not have any lower bound on the growth rate of this term, due to the cancellation of the large quantities $1/p_n$ and $\mu/np_n\mu_1$. Since these terms are of the same magnitude, we can only upper bound the result.

Now, consider (29). Note that $p_1^i = 1 + o(1)$ for any $i \le n$ and $-\frac{n}{i} \le 1 - \frac{\mu}{i\mu_1} \le 1$. Specifically, $1 - \frac{\mu}{i\mu_1} = O(n/i)$. Furthermore, $-n \ln i \le i - \frac{\mu}{\mu_1} h_i \le i$. This gives the bound $i - \frac{\mu}{\mu_1} H_i = O(n \ln n)$.

Thus, substituting these bounds into (29), we find that

$$\sum_{i=1}^{n-1} p_1^i p_n (1 - \frac{\mu}{i\mu_1})(i - \frac{\mu}{\mu_1} H_i) \le O(p_n) \sum_{i=1}^{n-1} \frac{n^2 \ln n}{i}$$
$$= O(p_n) n^2 \ln n H_{n-1} = O(p_n n^2 \ln^2 n) = o(n \ln n) = o(1/p_n).$$

□

8.2. *Regime 3, specific setting: Polynomial scaling* $p_n = 1/n^\alpha, \alpha > 1$

In this section, we focus on the specific setting of polynomial p_n scaling where $p_n = 1/n^\alpha, \alpha > 1$. This allows us to obtain much stronger results. To prove Theorem 4.2, we start by characterizing the asymptotic behavior of throughput μ in Lemma 8.3:

Lemma 8.3. *If $p_n = 1/n^\alpha$ for $\alpha > 1$, then*

$$\mu = n\mu_1 - \mu_1 n^{2-\alpha} \ln n - (1 + o(1))\mu_1 (\frac{\mu_1}{\mu_n} - 1) n^{2-\alpha}.$$

Proof. Consider the general expression for μ from Lemma 5.3:

$$\mu = \frac{1}{p_n} \left(\frac{1}{\mu_n} + \frac{1}{n\mu_1} \frac{p_1^n}{p_n} + \sum_{b=1}^{n-1} \frac{p_1^b}{b\mu_1} \right)^{-1}.$$

From Lemma 8.1, we know that $\mu = (1 + o(1))n\mu_1$. Note that for sufficiently large n, $\mu_n < n\mu_1$, and as a result $\mu < n\mu_1$, because the completion rate μ_i in state i is at most $n\mu_1$ for all states i, for all $n \ge \frac{\mu_n}{\mu_1}$.

We therefore examine $\frac{n\mu_1}{\mu} - 1$:

$$\frac{n\mu_1}{\mu} - 1 = \left(n\mu_1 p_n \left(\frac{1}{\mu_n} + \frac{1}{n\mu_1} \frac{p_1^n}{p_n} + \sum_{b=1}^{n-1} \frac{p_1^b}{b\mu_1} \right) - 1 \right)$$
$$= \frac{n\mu_1 p_n}{\mu_n} + \frac{n\mu_1 p_n}{n\mu_1} \frac{p_1^n}{p_n} + n\mu_1 p_n \sum_{b=1}^{n-1} \frac{p_1^b}{b\mu_1} - 1$$
$$= \frac{n^{1-\alpha}\mu_1}{\mu_n} + p_1^n - 1 + n^{1-\alpha} \sum_{b=1}^{n-1} \frac{p_1^b}{b}. \tag{30}$$

We next expand $p_1^n = (1 - n^{-\alpha})^n$ as

$$(1 - n^{-\alpha})^n = 1 - nn^{-\alpha} + \frac{n(n-1)}{2} n^{-2\alpha} \ldots$$

Because $\alpha > 1$, we can summarize this expansion as:

$$p_1^n = 1 - n^{1-\alpha} + O(n^{2-2\alpha}).$$

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

Substituting the asymptotic behavior of p_1^n into (30), we find that

$$\frac{n^{1-\alpha}\mu_1}{\mu_n} + p_1^n - 1 + n^{1-\alpha}\sum_{b=1}^{n-1}\frac{p_1^b}{b} = \frac{n^{1-\alpha}\mu_1}{\mu_n} - n^{1-\alpha} + O(n^{2-2\alpha}) + n^{1-\alpha}\sum_{b=1}^{n-1}\frac{p_1^b}{b}.$$

Note also that $p_1^b \in [p_1^n, 1]$ for all $b \le n$. In particular, $p_1^b = 1 + O(n^{1-\alpha})$ (uniformly) for all $b \le n$. We can therefore simplify the summation as

$$n^{1-\alpha}\sum_{b=1}^{n-1}\frac{p_1^b}{b} = n^{1-\alpha}(1 + O(n^{1-\alpha}))\sum_{b=1}^{n-1}\frac{1}{b} = n^{1-\alpha}\ln n + O(n^{2-2\alpha}\ln n)).$$

Putting it together, we find that

$$\begin{aligned}\frac{n\mu_1}{\mu} - 1 &= \frac{n^{1-\alpha}\mu_1}{\mu_n} - n^{1-\alpha} + O(n^{2-2\alpha}) + n^{1-\alpha}\ln n + O(n^{2-2\alpha}))\\ &= n^{1-\alpha}\left(\ln n + \frac{\mu_1}{\mu_n} - 1\right) + O(n^{2-2\alpha}\ln n). \end{aligned} \tag{31}$$

Let x denote the growth rate of the final expression (31), and use it to solve for μ:

$$\frac{n\mu_1}{\mu} - 1 = x \implies n\mu_1 - \mu = x\mu \implies n\mu_1 = (1+x)\mu \implies \frac{n\mu_1}{1+x} = \mu.$$

Note that for any x' near zero, $\frac{1}{1+x'} = 1 - x' + O(x'^2)$. Furthermore, since $\alpha > 1$, x is in fact near 0. Thus, we have

$$\begin{aligned}\mu = \frac{n\mu_1}{1+x} &= n\mu_1(1 - x + O(x^2))\\ &= n\mu_1 - n\mu_1\left(n^{1-\alpha}\left(\ln n + \frac{\mu_1}{\mu_n} - 1\right) + O(n^{2-2\alpha}\ln n) + O(n^{2-2\alpha}\ln^2 n)\right)\\ &= n\mu_1 - \mu_1 n^{2-\alpha}\ln n - \mu_1\left(\frac{\mu_1}{\mu_n} - 1\right)n^{2-\alpha} + O(n^{3-2\alpha}\ln^2 n).\end{aligned}$$

Since $\alpha > 1$, we have $3 - 2\alpha < 2 - \alpha$, so our final expression matches our theorem statement with a more specific error bound. □

We are now ready to prove the remainder of Theorem 4.2 by characterizing the asymptotic mean queue length $\mathbb{E}[\Delta(Y_d)]$.

Theorem 8.1. *If $p_n = 1/n^\alpha$ for $\alpha > 1$, then*

$$\mathbb{E}[\Delta(Y_d)] = (1 + o(1))\frac{1}{2}n^{2-\alpha}\ln^2 n.$$

Proof. We start by splitting up $\mathbb{E}[\Delta(Y_d)]$ into five terms, as before:

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

$$\mathbb{E}[\Delta(Y_d)] = p_1^n(1 - \frac{\mu}{n\mu_1})(n - 1 + \frac{1}{p_n} - \frac{\mu}{\mu_1}H_{n-1} - \frac{\mu}{np_n\mu_1}) \tag{32}$$

$$+ p_n \sum_{i=1}^{n-1} p_1^i i \tag{33}$$

$$- \frac{\mu p_n}{\mu_1} \sum_{i=1}^{n-1} p_1^i \tag{34}$$

$$- \frac{\mu p_n}{\mu_1} \sum_{i=1}^{n-1} p_1^i H_i \tag{35}$$

$$+ \frac{\mu^2 p_n}{\mu_1^2} \sum_{i=1}^{n-1} p_1^i \frac{1}{i} H_i \tag{36}$$

We will demonstrate the following asymptotic growth rates for each of these terms:

(1) Non-summation term (32): $(1 + o(1))(\frac{\mu_1}{\mu_n} - \gamma)n^{2-\alpha} \ln n$.
(2) $p_1^i i$ term (33): $O(n^{2-\alpha})$.
(3) p_1^i term (34): $-(1 + o(1))p_1 n^{2-\alpha}$.
(4) $p_1^i H_i$ term (35): $-n^{2-\alpha} \ln n + O(n^{2-\alpha})$.
(5) $p_1^i H_i / i$ term (36): $\frac{1}{2}n^{2-\alpha} \ln^2 n + O(n^{2-\alpha} \ln n)$.

Note that only Item 1, the asymptotic growth rate for (36), is non-negligible. As a result, these growth rates sum to $\frac{1}{2}n^{2-\alpha} \ln^2 + O(n^{2-\alpha} \ln n)$, as desired.

We start with the non-summation term (32):

$$p_1^n(1 - \frac{\mu}{n\mu_1})(n - 1 + \frac{1}{p_n} - \frac{\mu}{\mu_1}H_{n-1} - \frac{\mu}{np_n\mu_1}).$$

We will consider the three multiplicative terms one by one, left to right. First, note that $p_1^n = 1 + o(1)$. Then, we expand $1 - \frac{\mu}{n\mu_1}$. Using Lemma 8.3:

$$1 - \frac{\mu}{n\mu_1} = \frac{n\mu_1 - \mu}{n\mu_1} = n^{1-\alpha} \ln n + (1 + o(1))(\frac{\mu_1}{\mu_n} - 1)n^{1-\alpha}.$$

Next expand the final (rightmost) parenthesis of (32). Again using Lemma 8.3, we have:

$$\begin{aligned} &n - 1 + \frac{1}{p_n} - \frac{\mu}{\mu_1}H_{n-1} - \frac{\mu}{np_n\mu_1} = n - 1 - \frac{\mu}{\mu_1}H_{n-1} + n^{\alpha}(1 - \frac{\mu}{n\mu_1}) \\ &= n - 1 - \frac{\mu}{\mu_1}H_{n-1} + n \ln n + (1 + o(1))n(\frac{\mu_1}{\mu_n} - 1) \\ &= n - 1 - (n - \frac{\ln n}{n^{2-\alpha}} - (1 + o(1))(\frac{\mu_1}{\mu_n} - 1))n^{2-\alpha})H_{n-1} + n \ln n + (1 + o(1))n(\frac{\mu_1}{\mu_n} - 1). \end{aligned} \tag{37}$$

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

Note that $2-\alpha<1$. Let $\epsilon>0$ be a constant such that $2-\alpha<1-\epsilon<1$. As a result, we can find the asymptotic growth rate of (37) as follows:

$$\begin{aligned}
&n-1-(n-n^{2-\alpha}\ln n-(1+o(1))(\frac{\mu_1}{\mu_n}-1))n^{2-\alpha})H_{n-1}+n\ln n+(1+o(1))n(\frac{\mu_1}{\mu_n}-1)\\
&=n-1-(n-O(n^{1-\epsilon}))H_{n-1}+n\ln n+(1+o(1))n(\frac{\mu_1}{\mu_n}-1)\\
&=n-1-(n-O(n^{1-\epsilon}))(\ln n+\gamma+O(1/n))+n\ln n+(1+o(1))n(\frac{\mu_1}{\mu_n}-1))\\
&=n-1+O(n^{1-\epsilon})(\ln n+\gamma+O(1/n))-n(\gamma+O(1/n))+(1+o(1))n(\frac{\mu_1}{\mu_n}-1)\\
&=n(\frac{\mu_1}{\mu_n}-\gamma)+o(n).
\end{aligned}$$

Combining these results, we find that the overall growth rate of (32) is

$$(1+o(1))(\frac{\mu_1}{\mu_n}-\gamma)n^{2-\alpha}\ln n,$$

confirming Item 1.

We next consider (33) which can be quantified exactly as:

$$\begin{aligned}
\sum_{i=1}^{n-1}p_1^i p_n i &= p_n\frac{p_1-np_1^n+(n-1)p_1^{1+n}}{(1-p_1)^2}=\frac{p_1-np_1^n+np_1p_1^n-p_1p_1^n}{p_n}\\
&=\frac{p_1(1-p_1^n)-np_1^n(1-p_1)}{p_n}=\frac{p_1(1-p_1^n)-np_1^n(1-p_1)}{p_n}.
\end{aligned}$$

We now expand $p_1^n=(1-p_n)^n$ using a binomial expansion, and recalling that $np_n=o(1)$,

$$\begin{aligned}
&\frac{p_1(1-p_1^n)-np_1^n(1-p_1)}{p_n}=\frac{p_1(np_n+O(n^2p_n^2))-np_n(1-np_n+O(n^2p_n^2))}{p_n}\\
&=\frac{-(1-p_1)np_n+O(n^2p_n^2)}{p_n}=\frac{-np_n^2+O(n^2p_n^2)}{p_n}=\frac{O(n^2p_n^2)}{p_n}=O(n^2p_n)=O(n^{2-\alpha}),
\end{aligned}$$

confirming Item 2.

Next consider (34) which again has an exact expression

$$\begin{aligned}
-\sum_{i=1}^{n-1}p_1^i p_n\frac{\mu}{\mu_1}&=-\frac{\mu}{\mu_1}p_n\frac{p_1-p_1^n}{1-p_1}=-\frac{\mu}{\mu_1}(p_1-p_1^n)=-\frac{\mu}{\mu_1}p_1(1-p_1^{n-1})\\
&=-\frac{\mu}{\mu_1}p_1(1+o(1))n^{1-\alpha}=-p_1(1+o(1))n^{2-\alpha},
\end{aligned}$$

confirming Item 3.

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

We next consider (35):

$$\begin{aligned}-\sum_{i=1}^{n-1} p_1^i p_n \frac{\mu}{\mu_1} H_i &= -\sum_{i=1}^{n-1} p_1^i p_n \frac{\mu}{\mu_1} \ln i + O(n^{2-\alpha}) \\ &= -p_n \frac{\mu}{\mu_1} \int_{i=1}^{n} p_1^i \ln i di + O(n^{2-\alpha}) \\ &= -p_n \frac{\mu}{\mu_1} \int_{i=1}^{n} (1-n^{-\alpha})^i \ln i di + O(n^{2-\alpha}). \end{aligned} \tag{38}$$

As usual, we convert the H_i to $\ln i$ and the summation to an integral. To bound the change caused by switching from H_i to $\ln i$, note that $|H_i - \ln i| \leq 1$, which gives rise to an error bounded by (34), which is negligible compared to the growth rate we will establish for this term.

The change caused by switching from the summation to the integral is similarly bounded using the fact that $\ln i - \ln(i-1) \leq 1$ (which again introduces an error bounded by (34)).

Now, we upper and lower bound the integral as follows:

$$\begin{aligned}\int_{i=1}^{n} p_1^i \ln i di &\leq \int_{i=1}^{n} \ln i di = n \ln n - n + 1, \\ \int_{i=1}^{n} p_1^i \ln i di &\geq \int_{i=1}^{n} p_1^n \ln i di = (n \ln n - n + 1) p_1^n. \end{aligned}$$

Using a binomial expansion of p_1^n, we have

$$p_1^n = (1-n^{-\alpha})^n = 1 - nn^{-\alpha} + \frac{n(n-1)}{2} n^{-2\alpha} - \ldots = 1 - O(n^{1-\alpha}).$$

We can therefore analyze the bound above as:

$$\begin{aligned}(n \ln n - n + 1) p_1^n &= (n \ln n - n + 1)(1 - O(n^{1-\alpha})) \\ &= n \ln n - n - O(n^{2-\alpha} \ln n) = n \ln n - O(n) \\ \Longrightarrow \int_{i=1}^{n} p_1^i \ln i di &= n \ln n - O(n). \end{aligned}$$

Substituting the integral into (38), multiplying by $-p_n \mu / \mu_1 = n^{1-\alpha} + O(n^{2-2\alpha} \ln n)$, and using Lemma 8.3, we get

$$-\sum_{i=1}^{n-1} p_1^i p_n \frac{\mu}{\mu_1} H_i = -n^{2-\alpha} \ln n + O(n^{2-\alpha}),$$

which matches Item 4, the asymptotic growth rate for (35).

Finally, consider (36). As usual, we convert the H_i to $\ln i$ and the summation to an integral:

$$\sum_{i=1}^{n-1} p_1^i p_n \frac{\mu^2}{i \mu_1^2} H_i = p_n \frac{\mu^2}{\mu_1^2} \int_{i=1}^{n} p_1^i \frac{\ln i}{i} di + O(n^{2-\alpha} \ln n). \tag{39}$$

To justify the error term, note that both conversions introduce an error in the order of

$$p_n \frac{\mu^2}{\mu_1} \sum_{i=1}^{n-1} p_1^i \frac{1}{i}.$$

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

Making use of our bound from Lemma 8.3, which states that $\mu = n\mu_1 + O(n^{1-\alpha} \ln)$, and the earlier shown result $p_1^i = 1 + O(n^{1-\alpha})$, we have

$$p_n \frac{\mu^2}{\mu_1} \sum_{i=1}^{n-1} p_1^i \frac{1}{i} = O(n^{2-\alpha} \ln n),$$

which is the error term we used in (39).

We upper and lower bound the integral from (39) as follows:

$$\int_{i=1}^{n} p_1^i \frac{\ln i}{i} di \leq \int_{i=1}^{n} \frac{\ln i}{i} di = \frac{\ln^2 n}{2},$$
$$\int_{i=1}^{n} p_1^i \frac{\ln i}{i} di \geq \int_{i=1}^{n} p_1^n \frac{\ln i}{i} di = p_1^n \frac{\ln^2 n}{2}. \tag{40}$$

Using our prior result $p_1^n = 1 + O(n^{1-\alpha})$, we get

$$\int_{i=1}^{n} p_1^i \frac{\ln i}{i} di \geq (1 + O(n^{1-\alpha})) \frac{\ln^2 n}{2} = \frac{\ln^2 n}{2} + O(n^{1-\alpha} \frac{\ln^2 n}{2}).$$

Combining this upper bound with our lower bound in (40), we obtain

$$\int_{=1}^{n} p_1^i \frac{\ln i}{i} di = \frac{\ln^2 n}{2} + O(n^{1-\alpha} \frac{\ln^2 n}{2}). \tag{41}$$

We now need to incorporate the $p_n \frac{\mu^2}{\mu_1^2}$ term. From Lemma 8.3, we have $\frac{\mu}{\mu_1} = n + O(n^{2-\alpha} \ln n)$. Then

$$p_n \frac{\mu^2}{\mu_1^2} = n^{-\alpha} (n + O(n^{2-\alpha} \ln n))^2 = n^{-\alpha} (n^2 + O(n^{3-\alpha} \ln n)) = n^{2-\alpha} + O(n^{3-2\alpha} \ln n). \tag{42}$$

We now combine (41) and (42) to get our result:

$$\sum_{i=1}^{n-1} p_1^i p_n \frac{\mu^2}{i\mu_1^2} H_i = \frac{1}{2} n^{2-\alpha} \ln^2 n + O(n^{2-\alpha} \ln n),$$

which matches the claimed asymptotic growth rate for (36), and completing the proof. □

9. Comparison: Asymptotic growth rate versus fixed-*n* formula

We have asymptotically characterized throughput μ and scaled mean queue length $\mathbb{E}[\Delta(Y_d)]$ for each of our three regimes of interest, namely, n-server dominated, balanced load, and 1-server dominated, in Theorems 4.2 and 4.1, within the load-focused multiserver scaling limit. In this section, we numerically evaluate our asymptotic results by comparing our asymptotic formulas with the explicit summation-based formulas in Lemmas 5.3 and 5.5. In Section 10, we derive practical insights from our asymptotic expressions and demonstrate that those insights generalize to settings beyond the ones studied in this paper, via numerical evaluation and simulation.

We numerically evaluate these formulas at numbers of servers up to $n = 10^8$. We choose this upper limit because modern exascale datacenters reach about this number of CPU cores. For instance, the Citadel Campus [31] reports 650 MW of planned total energy consumption, corresponding to about

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

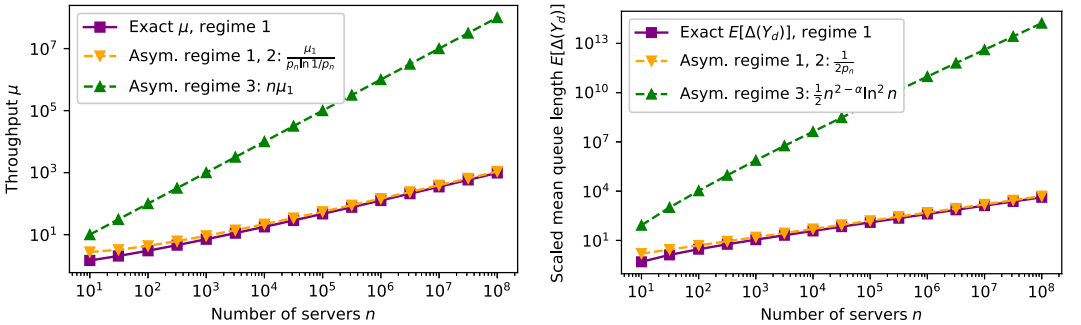


Figure 2. *Exact versus asymptotic formulas in the n-server dominated load regime for μ and $\mathbb{E}[\Delta(Y_d)]$ as functions of n. Parametrization: $p_n = n^{-0.5}$ and $\mu_1 = \mu_n = 1$.*

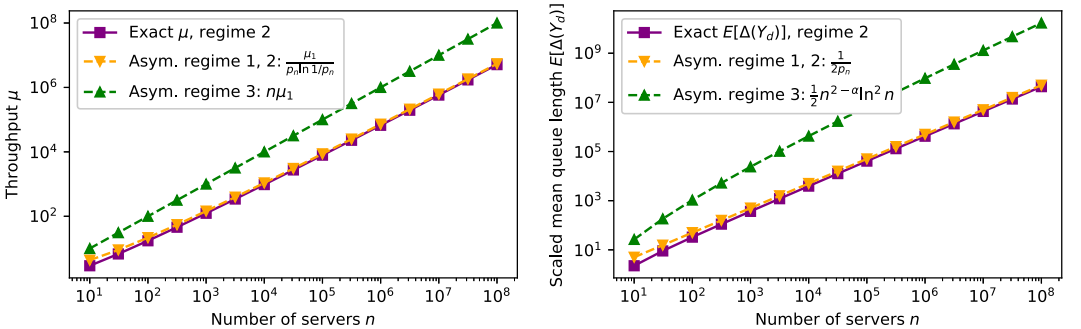


Figure 3. *Exact versus asymptotic formulas in the balanced load regime for μ and $\mathbb{E}[\Delta(Y_d)]$ as functions of n. Parametrization: $p_n = n^{-1}$ and $\mu_1 = \mu_n = 1$.*

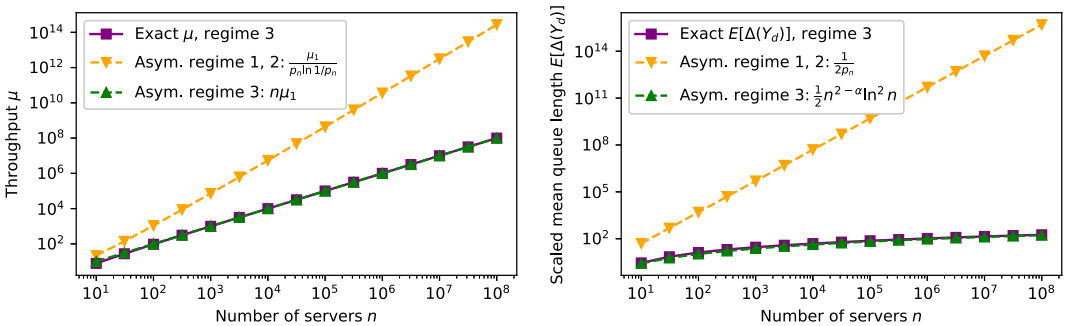


Figure 4. *Exact versus asymptotic formulas in the 1-server dominated regime for μ and $\mathbb{E}[\Delta(Y_d)]$ as functions of n. Parametrization: $p_n = n^{-2}$ and $\mu_1 = \mu_n = 1$.*

1-2 million 1U server racks and about 100 million CPU cores. Values of n around 10^8 thus represent the approximate scale of the largest MSJ queueing systems currently in operation or construction.

We explore three parameter families, setting $p_n = n^{-\alpha}$ for $\alpha \in \{0.5, 1, 2\}$. These three values of α correspond to our three regimes. In Figure 2, we compare the exact values of μ and $\mathbb{E}[\Delta(Y_d)]$ with the asymptotic growth rates in the $\alpha = 0.5$ parametrization, which is in the n-server dominated regime. In Figure 3, we examine the $\alpha = 1$ parametrization, which is in the balanced load regime. In Figure 4, we examine the $\alpha = 2$ parametrization, which is in the 1-server dominated regime. In this section, we specifically focus on the $\mu_1 = \mu_n = 1$ parameterization. See Appendix A for further parameterizations of μ_1 and μ_n.

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

For each parameter family, we compare the exact formulas against each of our two asymptotic growth rates. In each case, the exact values converge to one of the two asymptotic growth rates in the $n \to \infty$ limit, as predicted by our theoretical results. Specifically, in Figure 2, where we examine the n-server dominated regime, and in Figure 3, where we examine the balanced regime, we compare the exact value of μ, given in purple, with the asymptotic formula $\frac{\mu_1}{p_n \ln 1/p_n}$, in orange, and exact value of $\mathbb{E}[\Delta(Y_d)]$ with the asymptotic formula $1/2p_n$. In Theorem 4.1 we proved that these two formulas, the purple and orange curves, differ by a ratio of $1 + o(1)$ as $n \to \infty, p_n \to 0$. We plot these formulas on a log-log plot, so this multiplicative convergence corresponds to a converging vertical gap.

Throughout the paper, for the benefit of monochromatic readers, the green curves use upward-triangle markers, the orange curves use downward-triangle markers, and the purple curves use square markers.

The convergence is clearly apparent over the interval $n \in [10^1, 10^8]$ spanned by our plots, confirming our theoretical results in both settings. For comparison, we also plot the formulas $n\mu_1$ against μ and $\frac{1}{2}n^{2-\alpha} \ln^2 n$ against $\mathbb{E}[\Delta(Y_d)]$, in green, which are the formulas that we derived for the 1-server dominated regime. As expected, this green curve diverges from the true value and from the correct asymptotic curve.

In Figure 4, we examine the 1-server dominated regime, and as expected from Theorem 4.1, the tight asymptotics are the green curves in this setting: $n\mu_1$ for μ and $\frac{1}{2}n^{2-\alpha} \ln^2 n$ for $\mathbb{E}[\Delta(Y_d)]$. In fact, throughout the range of n that we examine, there is nearly no perceptible separation between the purple and green curves, indicating not just multiplicative convergence, but a tight multiplicative approximation across all n and p_n.

10. Empirical evaluation of extension models

Our asymptotic theoretical results imply two practical insights in the 1-and-n model, as the mix of jobs changes from n-server focused to 1-server focused:

(1) The throughput of the system rises monotonically, with an inflection point around regime 2, where the loads of both types of jobs are similar.
(2) The mean response time of the system, holding fraction-of-throughput constant, is a bell curve, rising for n-server-dominated systems, peaking for a balanced mixture, and falling again for 1-server dominated systems.

We will show that these insights hold not just in the 1-and-n system that is the primary focus of this paper, but also in the following three generalizations:

- A system with an additional scaling behavior: 1-server jobs complete at rate $\mu_1 = n$, while $\mu_n = 1$, so the n-server jobs have much longer duration than the 1-server jobs, in addition to their larger server need (see Section 10.2).
- A system where the large-resource jobs require $n/2$ servers, rather than n servers (see Section 10.3).
- A system with three classes of jobs: n-server, $n/2$-server and 1-server jobs (see Section 10.4).

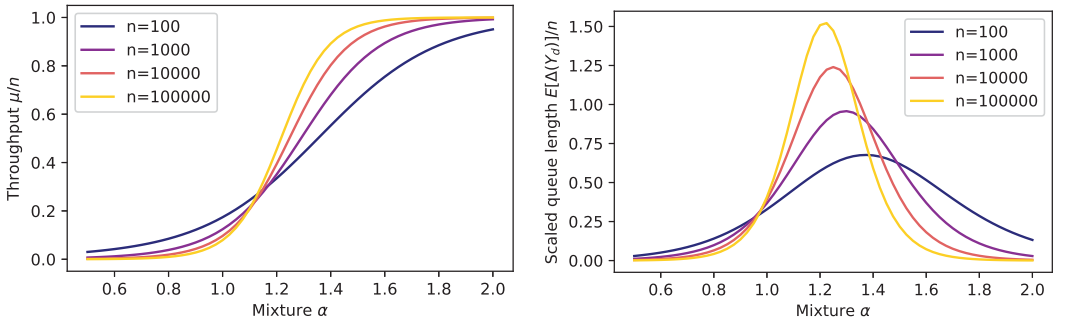
To investigate these systems, we use a combination of exact calculations, using the formulas proven in Lemmas 5.3 and 5.5 where applicable, as well as simulation. In all cases, we show that the same two practical insights apply, appropriately tailored to the relevant systems.

We start in Section 10.1 by investigating the original 1-and-n system via exact calculation and simulation, to obtain a baseline for the behavior corresponding to our insights, then turn to the extensions.

10.1. Calculations and simulation for 1-and-n system

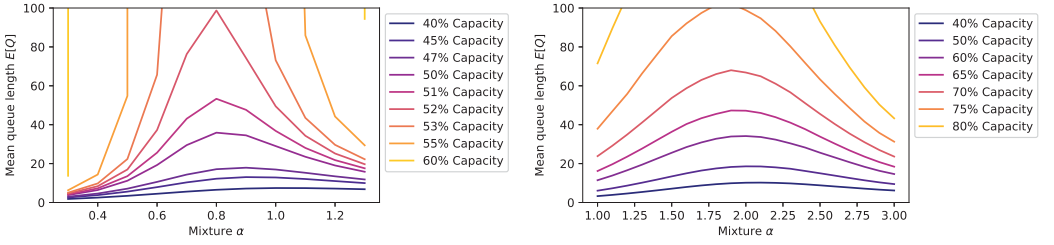
In Figure 5, we see both insights verified in the 1-and-n system, via direct calculation of throughput μ and the scaled queue length $\mathbb{E}[\Delta(Y_d)]$ using Lemmas 5.3 and 5.5. Each curve represents a fixed

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press



(a) (b)

Figure 5. *Main setting with server needs* 1 *and* n, $\mu_1 = \mu_n = 1$, *and* $p_n = 1/n^\alpha$. *(a) Throughput* μ/n, *calculated using Lemma 5.3. (b) Scaled mean queue length* $\mathbb{E}[\Delta(Y_d)]/n$, *calculated using Lemma 5.5.*



(a) (b)

Figure 6. *Simulation results (*10^8 *jobs per data point) for mean queue length when* $p_{10} = 1/n^\alpha$ *and* λ *is chosen so that a constant fraction of capacity is utilized. (a) Main setting with server needs* 1 *and 10,* $n = 10$ *servers and* $\mu_1 = \mu_{10} = 1$. *Equal-area setting with server needs* 1 *and* 10, $n = 10$ *servers and* $\mu_1 = 1, \mu_{10} = 10$.

number of servers n and a varying parameter α controlling the fraction of large jobs $p_n = \frac{1}{n^\alpha}$. We scale throughput μ and scaled queue length $\mathbb{E}[\Delta(Y_d)]$ by a factor of n to make the curves comparable.

In Figure 5a, we observe that throughput increases monotonically with α, with an inflection point around $\alpha \approx 1.2$, as predicted by Theorems 4.1 and 4.2.

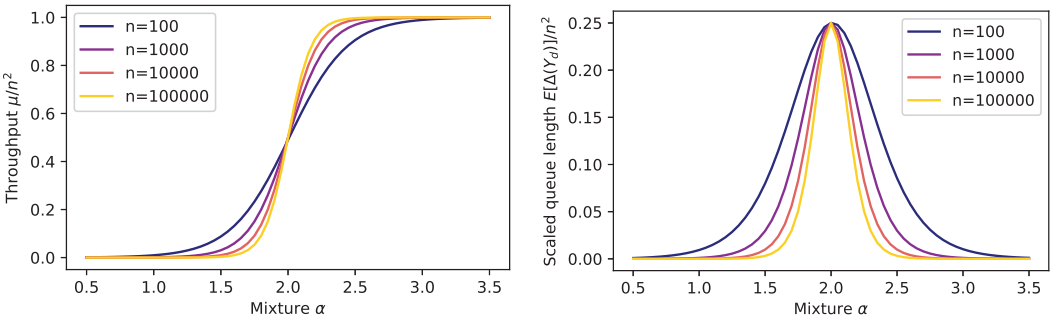
In Figure 5b, we observe that scaled mean queue length increases with α to a peak around $\alpha = 1.2$ and then decreases back to 0 as α gets large, also as predicted by Theorems 4.1 and 4.2.

In Figure 6a, we observe via simulation that these insights are not restricted to calculations or to very large n. Here, we simulate $n = 10$ servers, with each curve representing a different fraction of capacity demanded, using the following formula:

$$\text{Fraction of capacity} := \frac{\mu_1 p_1 + \mu_n p_n n}{n}$$

The bell-curve behavior of mean queue length is clearly visible, peaking around $\alpha = 0.8$, and is especially clear for capacity fractions around the boundary of the stability region. Our bell-curve insight thus applies not only for scaled mean queue length $\mathbb{E}[\Delta(Y_d)]$, but also for mean queue length $\mathbb{E}[Q]$ itself, including in systems with as few as $n = 10$ servers.

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press



(a) (b)

Figure 7. *Scaled service rate setting with server needs* 1 *and* n, $\mu_1 = n, \mu_n = 1$ *and* $p_n = 1/n^\alpha$. *(a) Throughput* μ/n^2, *calculated using Lemma 5.3. (b) Scaled mean queue length* $\mathbb{E}[\Delta(Y_d)]/n^2$, *calculated using Lemma 5.5.*

10.2. Calculations and simulation for duration-scaling system

In this section, we investigate the realistic scenario where low-resource jobs also require much lower service duration than high-resource jobs. This is realistic in datacenters, for instance, where a single-server analytics job might only take a few seconds, while an LLM-training job might take most of the datacenter for a period of months. To model this scenario, we set the service rate μ_1 of the 1-server jobs equal to n, the number of servers, while leaving the server rate μ_n of the n-server jobs constant.

In Figure 7, we again observe that both of our insights are verified even more clearly in the duration-scaling system where throughput μ and the scaled queue length $\mathbb{E}[\Delta(Y_d)]$ are computed using Lemmas 5.3 and 5.5. Note that we use the scaling μ/n^2 and $\mathbb{E}[\Delta(Y_d)]/n^2$ to account for the fact that the 1-server jobs have both $1/n$ the mean duration and $1/n$ the resource demand of the n-server jobs.

Figure 7a shows that throughput μ has an inflection point at $\alpha = 2$, and Figure 7b shows that scaled mean queue length has a particularly symmetrical bell curve centered on $\alpha = 2$ as well. We expect the inflection point to be around $\alpha = 2$, because this is where the load of n-server and 1-server jobs is equalized. Thus, our insights are verified in this setting as well, despite leaving the domain of applicability of asymptotic results.

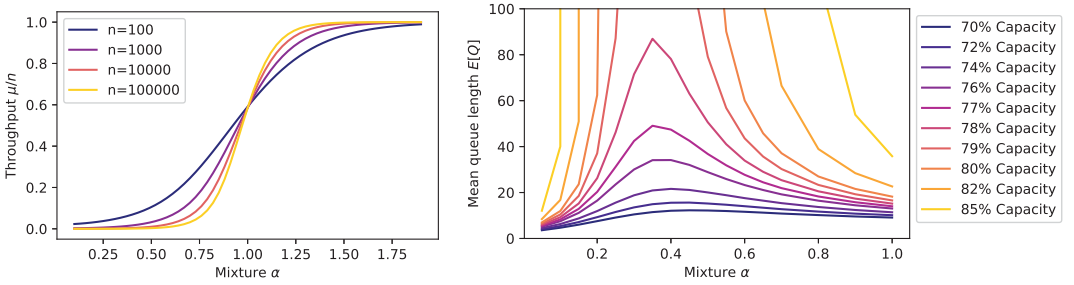
Note that Figure 6b shows that our insights again carry over into simulation and into low-n settings, with mean queue length $\mathbb{E}[Q]$ peaking around $\alpha = 1.9$, across a wide range of fractions of capacity for $n = 10$ servers. Hence, the bell-curve behavior is even smoother and more consistent in this duration-scaling setting than in Section 10.1.

10.3. Calculations and simulation for half-size large jobs

In this section, we investigate the case where the large-resource jobs require $n/2$ servers, rather than n servers as in our asymptotic results. This mimics the definition of leadership-class jobs on the Frontier supercomputer at Oak Ridge National Labs [23], which consists of jobs that require at least 60% of the capacity of the supercomputer, giving top priority to such large-resource jobs and setting policy goals to attain a large fraction of load devoted to leadership-class jobs.

With half-size large jobs, we can no longer use the exact results in Lemmas 5.3 and 5.5 to compute throughput and scaled queue length. Instead, we compute the throughput μ using the product-form stationary distribution formula for the MSJ 2-class saturated system from [14, Section 5], and [4, Section 5.1]. Figure 8a shows the result of this calculation, with throughput increasing monotonically with increasing α, as the mixture of load moves from n-server dominated to 1-server dominated, with an inflection point around $\alpha = 1$.

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press



(a) (b)

Figure 8. *Calculations and simulation for the half-size large jobs setting with server needs* 1 *and* $n/2$, $\mu_1 = \mu_{n/2} = 1$ *and* $p_n = 1/n^\alpha$. *(a) Throughput* μ/n, *calculated using [14, Section 5], [4, Section 5.1]. (b) Simulation results (*10^8 *jobs per data point) for mean queue length when* λ *is chosen so that a constant fraction of capacity is utilized and* $n = 10$ *servers.*

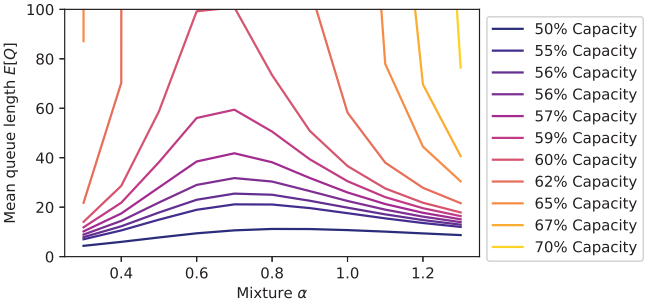


Figure 9. *Simulation results (*10^8 *jobs per data point) for three-class setting with server needs* $1, 5,$ 10, $n = 10$ *servers,* $\mu_1 = \mu_5 = \mu_{10} = 1$, *and* $p_5 = p_{10} = n^\alpha/2$.

Since we do not have an exact formula for limiting scaled mean queue length $\mathbb{E}[\Delta(Y_d)]$, we instead turn to simulation in Figure 8b, again with $n = 10$ servers. Here we see the same bell-curve behavior of mean queue length $\mathbb{E}[Q]$ near the boundary of the stability region, with a peak around $\alpha = 0.4$. Note that in this setting, the fraction of capacity in use is significantly higher than in any of the other settings around the boundary of the stability region for α near that peak—around 78% of capacity in this setting, compared to $50 - 60\%$ of capacity in the other settings. This follows from the fact that this system will never leave $n/2$ or more servers empty, while settings with jobs that demand n servers may leave up to $n - 1$ servers empty. Note also that in this setting, the peak of mean queue length $\mathbb{E}[Q]$ does not occur at the same α value as the inflection point of throughput, in contrast to Sections 10.2 and 10.1. We leave the analysis of this behavior to future work.

10.4. Calculations and simulation for three-class system

In this section, we investigate the case where there are three classes of jobs, requiring $1, n/2$, and n servers, respectively. This is the most realistic setting because in practice, queueing systems where jobs require varying resources are not limited to two possibilities, but instead vary widely.

In this setting, it is more difficult to calculate throughput μ than in the half-size large jobs setting of Section 10.3. There is an exact product-formula stationary distribution formula for this setting, given by Rumyantsev and Morozov [28], as well as [14, Section 3]. However, the number of terms in this formula is exponential in n, while the formula in Lemma 5.3 and the formula used in Section 10.3 each contained a number of terms that was linear in n. We leave the problem of efficiently calculating μ in

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

this setting to future work—a formula with a number of terms that is linear in n is likely doable. No exact formula for $\mathbb{E}[\Delta(Y_d)]$ is known in this setting.

We now turn to simulation in Figure 9, again with $n = 10$ servers. There is a larger parameter space of probability distributions over the three classes, but we choose arbitrarily to equalize the fraction of jobs in each of the two larger-resource classes. Here we see the same bell-curve behavior of mean queue length $\mathbb{E}[Q]$ near the boundary of the stability region, with a peak around $\alpha = 0.7$, between the peaks in the original setting in Section 10.1 and the half-size large jobs setting in Section 10.3. The fraction of capacity in use is likewise between those two settings around the boundary of the stability region near that peak. Even for this more complicated setting, our key bell-curve insight is again confirmed.

11. Conclusion

In this paper, we studied the MSJ setting, and specifically the 1-and-n system, in the LFMS limit. Within this limit, we provided the first analysis of the asymptotic growth rate of throughput μ and scaled mean queue length $\mathbb{E}[Q(1-\rho)] \to \mathbb{E}[\Delta(Y_d)]$ in each of three regimes: the n-server dominated load regime, balanced load regime, and 1-server dominated load regimes. Our Theorems 4.1 and 4.2 demonstrate that the balanced load regime represents a tipping point with respect to throughput: In the n-server dominated regime, throughput is dominated by 1-server jobs that run while an n-server job is blocking the head of the queue, while in the 1-server dominated regime, throughput is dominated by 1-server jobs running on all n servers. Similarly, we prove that the balanced load regime exhibits the largest scaled response time, indicating the most variation in service rates: $\mathbb{E}[\Delta(Y_d)] = \theta(n)$ in this regime, with slower growth rates as a function of n in the other two regimes.

Furthermore, our numerical and simulation results indicate that these insights can be generalized to duration-scaled systems where low-resource jobs require much lower service duration than high-resource jobs, systems where the large-resource jobs require $n/2$ servers rather than n servers, and systems with three classes where jobs require $1, n/2$, and n servers. A natural direction for future research would be to prove asymptotic results in these settings.

Competing interests declaration. The authors declare none.

Acknowledgements. The work of the first author was supported by a Tennenbaum Postdoctoral Fellowship at Georgia Tech and by AFOSR Grant FA9550-24-1-0,002. The work of the second author was supported by the National Science Foundation Grant CMMI-2,127,778.

The authors thank the reviewer and editor for helpful comments, which improved the paper.

References

[1] Anselmi, J. & Doncel, J. (2025). Balanced splitting: A framework for achieving zero-wait in the multiserver-job model. *IEEE Transactions on Parallel and Distributed Systems* 36(1): 43–54. doi: 10.1109/TPDS.2024.3493631.

[2] Baccelli, F. & Foss, S. (1995). On the saturation rule for the stability of queues. *Journal of Applied Probability* 32(2): 494–507. doi: 10.2307/3215303.

[3] Brill, P.H. & Green, L. (1984). Queues in which customers receive simultaneous service from a random number of servers: A system point approach. *Management Science* 30(1): 51–68.

[4] Céline, C. & Isaac, G. Graph-based product form. arXiv preprint arXiv:2502.13890 2025.

[5] Carastan-Santos, D., De Camargo, R.Y., Trystram, D. & Zrigui, S. (2019). One can only gain by replacing EASY backfilling: A simple scheduling policies case study. In *2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*. Larnaca, Cyprus: IEEE, pp. 1–10. doi: 10.1109/CCGRID.2019.00010.

[6] Chen Z., Grosof I., Berg B. (2024) Analyzing practical policies for multiresource job scheduling. *arXiv preprint* arXiv:2412.08915. Accepted to ACM SIGMETRICS 2025.

[7] Dimitrov, M. (2011). Single-server queueing system with Markov-modulated arrivals and service times. *PLISKA Studia Mathematica Bulgarica* 20: 53–62.

[8] Falin, G. & Falin, A. (1999). Heavy traffic analysis of M/G/1 type queueing systems with Markov-modulated arrivals. *Top* 7(2): 279–291.

[9] Filippopoulos, D. & Karatza, H. (2007). An M/M/2 parallel system model with pure space sharing among rigid jobs. *Mathematical and Computer Modelling* 45(5): 491–530.

[10] Foss, S. & Konstantopoulos, T. (2004). An overview of some stochastic stability methods. *Journal of the Operations Research Society of Japan* 47(4): 275–303.

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

[11] Green, L. (1980). A queueing system in which customers require a random number of servers. *Operations Research* 28(6): 1335–1346. doi: 10.1287/opre.28.6.1335.
[12] Grosof, I. & Harchol-Balter, M. (2023). Invited paper: ServerFilling: A better approach to packing multiserver jobs. In Y.A. Liu & E.M. Schiller (eds), *Proceedings of the 5th Workshop on Advanced Tools, Programming Languages, and PLatforms for Implementing and Evaluating Algorithms for Distributed Systems*, ApPLIED 2023. New York, NY, USA: Association for Computing Machinery, pp. 1–5. doi: 10.1145/3584684.359 7264.
[13] Grosof, I., Harchol-Balter, M. & Scheller-Wolf, A. (2022a). WCFS: A new framework for analyzing multiserver systems. *Queueing Systems* 102: 143–174.
[14] Grosof, I., Harchol-Balter, M. & Scheller-Wolf, A. (2023a). New stability results for multiserver-job models via product-form saturated systems. *MAthematical Performance Modeling and Analysis (MAMA)* 51(2): 6–8.
[15] Grosof, I., Hong, Y., Harchol-Balter, M. & Scheller-Wolf, A. (2023b). The RESET and MARC techniques, with application to multiserver-job analysis. *Performance Evaluation* 162(102378): 0166–5316. doi: 10.1016/j.peva.2023.102378.
[16] Grosof, I., Scully, Z., Harchol-Balter, M. & Scheller-Wolf, A. (2022b). Optimal scheduling in the multiserver-job model under heavy traffic. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 6(3). doi: 10.1145/3570612.
[17] Harchol-Balter, M. (2022). The multiserver job queueing model. *Queueing Systems* 100(3): 201–203.
[18] Hong, Y. & Wang, W. (2024). Sharp waiting-time bounds for multiserver jobs. *Stochastic Systems* 14(4): 455–478.
[19] Latouche, G. & Ramaswami, V. (1999). *ASA-SIAM Series on Statistics and Applied Mathematics*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics. doi: 10.1137/1.9780898719734.
[20] Liu, X. & Ying, L. (2020). Steady-state analysis of load-balancing algorithms in the sub-Halfin–Whitt regime. *Journal of Applied Probability* 57(2): 578–596. doi: 10.1017/jpr.2020.13.
[21] Liu, X. & Ying, L. (2022). Universal scaling of distributed queues under load balancing in the super-Halfin-Whitt regime. *IEEE/ACM Transactions on Networking* 30(1): 190–201. doi:10.1109/TNET.2021.3105480.
[22] Maguluri, S.T., Srikant, R. & Ying, L. (2012). Stochastic models of load balancing and scheduling in cloud computing clusters. In A. Greenberg & K. Sohraby (eds), *2012 Proceedings IEEE Infocom*. IEEE, pp. 702–710.
[23] Oak Ridge National Laboratory. (2025). Frontier user guide. https://docs.olcf.ornl.gov/systems/frontier_user_guide.html#job-priority-by-node-count.
[24] Olliaro, D., Marsan, M.A., Balsamo, S. & Marin, A. (2023). The saturated multiserver job queuing model with two classes of jobs: Exact and approximate results. *Performance Evaluation* 162(102370): 0166–5316. doi: 10.1016/j.peva.2023.102370.
[25] Psychas, K. & Ghaderi, J. (2018). Randomized algorithms for scheduling multi-resource jobs in the cloud. *IEEE/ACM Transactions on Networking* 26(5): 2202–2215.
[26] Wolfram Research Inc. (2024) Mathematica, Version 14.1, Champaign, IL. https://www.wolfram.com/mathematica.
[27] Rumyantsev, A. (2020). Stability of multiclass multiserver models with automata-type phase transitions. In *Proceedings of the Second International Workshop on Stochastic Modeling and Applied Research of Technology (SMARTY 2020)*. 2792: 213–225.
[28] Rumyantsev, A. & Morozov, E. (2017). Stability criterion of a multiserver model with simultaneous service. *Annals of Operations Research* 252(1): 29–39.
[29] Seila, A.F. (1984). Technical note — on waiting times for a queue in which customers require simultaneous service from a random number of servers. *Operations Research* 32(5): 1181–1184. doi: 10.1287/opre.32.5.1181.
[30] Srinivasan, S., Kettimuthu, R., Subramani, V. & Sadayappan, P. (2002). Characterization of backfilling strategies for parallel job scheduling. In S. Olariu (ed), *Proceedings. International Conference on Parallel Processing Workshop*. Vancouver, BC, Canada: IEEE, pp. 514–519. doi: 10.1109/ICPPW.2002.1039773.
[31] Switch Tahoe Reno colocation | Nevada colocation | Switch, May 2025. https://www.switch.com/tahoe-reno/.
[32] Tirmazi, M., Barker, A., Deng, N., Haque, M.E., Qin, Z.G., Hand, S., Harchol-Balter, M. & Borg, J.W. (2020). Borg: The next generation. In A. Bilas, K. Magoutis, & E. Markatos (eds), *Proceedings of the Fifteenth European Conference on Computer Systems*. New York, NY, USA: Association for Computing Machinery, pp. 1–14. EuroSys'20.
[33] Wan, B., Liu, G., Song, Z., Wang, J., Zhang, Y., Shen, G., Wang, S., Wei, H., Wang, C., Lou, W., Yang, X., Zhang, M., Jiang, K., Ren, C., Zhi, X., Menghan, Y., Nan, Z., Zheng, Z., Zhong, B., Wang, Q., Huan, Y., Chi, J., Zhang, W., Yuhan, L., Zichian, D., Zhao, S., Zhang, Y., Tang, J., Liu, Z., Chuan, W., Peng, Y., Lin, H., Xiaio, W., Liu, X. & Xiang, L. (2025). Robust LLM training infrastructure at bytedance. In Y. Won & Y. Kwon (eds), *Proceedings of the ACM SIGOPS 31st Symposium on Operating Systems Principles*, SOSP'25. Association for Computing Machinery, New York, NY, USA. pp. 186–203.
[34] Wang, J. & Guo, W. (2009). The application of backfilling in cluster systems. In C.-X. Wang (ed), *2009 WRI International Conference on Communications and Mobile Computing*. Vol. 3. Kunming, China: IEEE, pp. 55–59. doi: 10.1109/CMC.2009.252.
[35] Wang, W., Xie, Q. & Harchol-Balter, M. (2021). Zero queueing for multi-server jobs. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 5(1), 1–25: doi: 10.1145/3447385.
[36] Ziheng, J., Haibin, L., Zhong, Y., Huang, Q., Yangrui, C., Zhi, Z., Yanghua, P., Xiang, L., Cong, X., Shibiao, N., Yulu, J., Sun, H., Hongmin, C., Zhihao, B., Hou, Q., Shipeng, Y., Ding, Z., Yiyao, S., Zhuo, J., Haohan, X., Haoran, W., Zhang, Z., Pengfei, N., Leqi, Z., Sida, Z., Liang, X., Zherui, L., Zhe, L., Xiaoying, J., Jianxi, Y., Xin, J. & Xin, L. (2024). Megascale: Scaling large language model training to more than 10,000 GPUs. https://www.usenix.org/conference/nsdi24/presentation/jiang-ziheng.

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

Appendix A. Additional numerical comparisons

In this section, we numerically confirm our theoretical results in Theorem 4.1 and 4.2, showing the convergence of μ and $\mathbb{E}[\Delta(Y_d)]$ to our theoretically predicted asymptotic behaviors. In Section 9, we demonstrated this convergence for the parametrization $\mu_1 = \mu_n = 1$. In this section, we explore the cases $\mu_1 = 10, \mu_n = 1$, where 1-server jobs have much shorter durations than n-server jobs, and $\mu_1 = 1, \mu_n = 10$, where 1-server jobs have much longer durations than n-server jobs.

In Figures A1–A3, we explore the case where $\mu_1 = 10, \mu_n = 1$, under the parametrizations $p_n = n^{-0.5}, p_n = n^{-1}$, and $p_n = n^{-2}$, respectively.

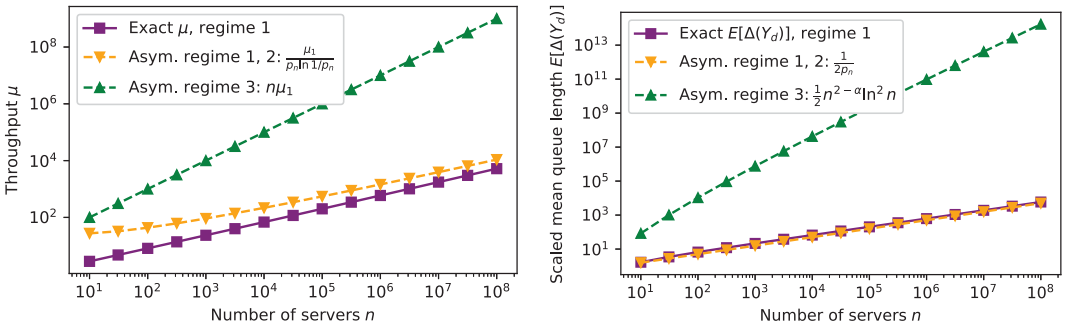


Figure A1. *Exact versus asymptotic formulas in the n-server dominated regime for μ and $\mathbb{E}[\Delta(Y_d)]$ as functions of n. Parametrization: $p_n = n^{-0.5}$ and $\mu_1 = 10, \mu_n = 1$.*

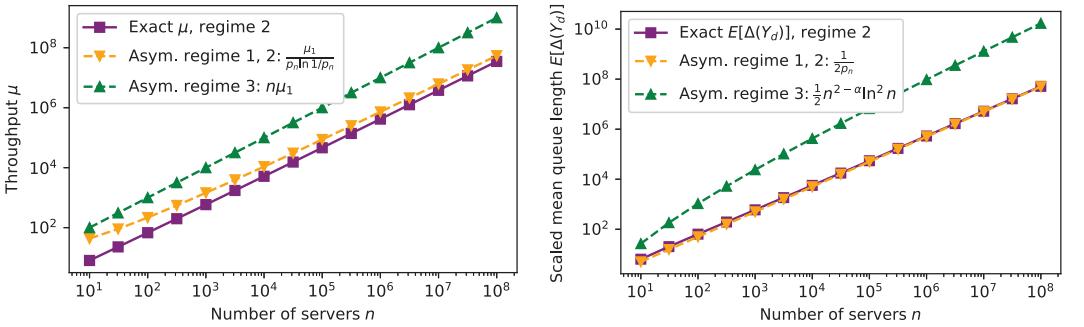


Figure A2. *Exact versus asymptotic formulas in the balanced load regime for μ and $\mathbb{E}[\Delta(Y_d)]$ as functions of n. Parametrization: $p_n = n^{-1}$ and $\mu_1 = 10, \mu_n = 1$.*

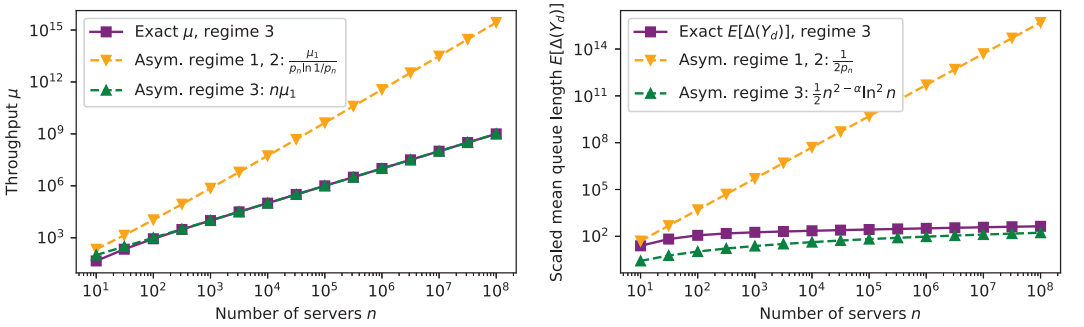


Figure A3. *Exact versus asymptotic formulas in the 1-server dominated regime for μ and $\mathbb{E}[\Delta(Y_d)]$ as functions of n. Parametrization: $p_n = n^{-2}$ and $\mu_1 = 10, \mu_n = 1$.*

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

These figures demonstrate the same multiplicative convergence shown in Section 9 and proven in Theorem 4.1 and 4.2: In Figures A1 and A2, the orange and purple curves converge, while in Figure A3, the green and orange curves converge.

Throughout the paper, for the benefit of monochromatic readers, the green curves use upward-triangle markers, the orange curves use downward-triangle markers, and the purple curves use square markers.

In some cases, the convergence is faster or slower than in the $\mu_1 = \mu_n = 1$ setting shown in Section 9. For instance, in Figure A1, the convergence for μ to its asymptotic curve is slower than it was in Figure 2, while the convergence for $\mathbb{E}[\Delta(Y_d)]$ is faster than in Figure 2, and in fact the asymptotic expression is a tight approximation across all n.

As for why the quality of approximation changes, for the μ plots, the two asymptotic formulas emphasize terms corresponding to periods of time when 1-server jobs are in service. The orange asymptotic curve emphasizes the period when 1-server jobs are emptying from service to allow an n-server job to be served, while the green asymptotic curve emphasizes the period when 1-server jobs are being served on all n servers. In the $\mu_1 = 10, \mu_n = 1$ parameterization, the 1-server jobs have shorter durations, raising the importance of the time when the n-server jobs are in service, which is not asymptotically relevant, but shows up in these plots for small n.

For the $\mathbb{E}[\Delta(Y_d)]$ plots, we do not have a clear interpretation. In Figures A1 and A2, the orange asymptotics are tighter than in Section 9, while in Figure A3, the green asymptotics are looser.

Likewise, in figures Figures A4 to A6, we explore the case where $\mu_1 = 1, \mu_n = 10$, under the parametrizations $p_n = n^{-0.5}, p_n = n^{-1}$, and $p_n = n^{-2}$, respectively.

As before, in Figures A4 and A5, the orange and purple curves converge, while in Figure A6, the green and orange curves converge.

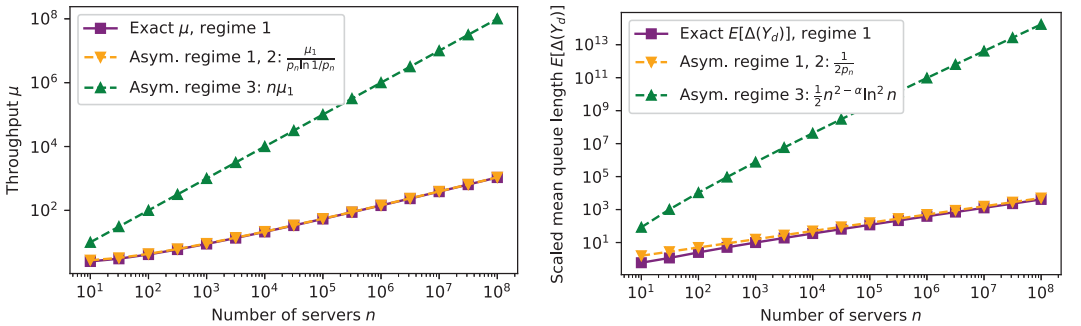


Figure A4. *Exact versus asymptotic formulas in the n-server dominated regime for μ and $\mathbb{E}[\Delta(Y_d)]$ as functions of n. Parametrization: $p_n = n^{-0.5}$ and $\mu_1 = 1, \mu_n = 10$.*

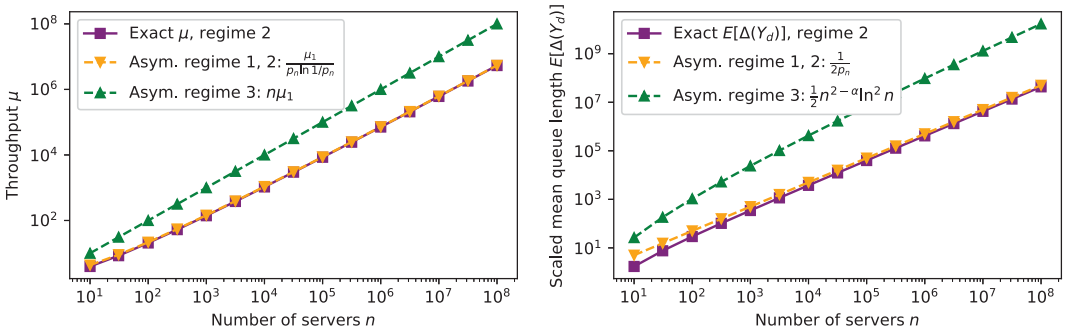


Figure A5. *Exact versus asymptotic formulas in the balanced load regime for μ and $\mathbb{E}[\Delta(Y_d)]$ as functions of n. Parametrization: $p_n = n^{-1}$ and $\mu_1 = 1, \mu_n = 10$.*

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press

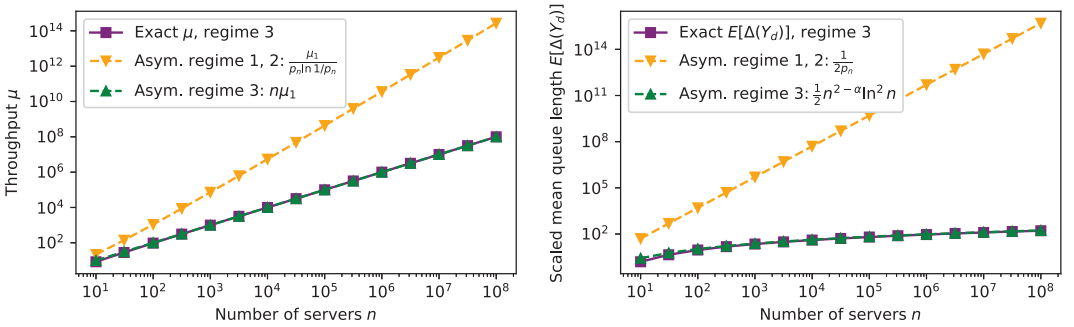


Figure A6. *Exact versus asymptotic formulas in the* 1*-server dominated regime for* μ *and* $\mathbb{E}[\Delta(Y_d)]$ *as functions of n. Parametrization:* $p_n = n^{-2}$ *and* $\mu_1 = 1, \mu_n = 10$.

As predicted by the discussion above, the exact formulas for μ are very tight in all three regimes, as this parametrization lengthens the periods when 1-server jobs are being served, improving the accuracy of the asymptotics at low n.

Again, it is unclear when our asymptotic formulas for $\mathbb{E}[\Delta(Y_d)]$ are tight for lower values of n. In this case, the green and purple curves are very close in Figure A6, the reverse of the situation in Figure A3.

Cite this article: Grosof I. and Ayhan H. (2026). Multiserver-job response time under multilevel scaling. *Probability in the Engineering and Informational Sciences* 40(4): 516–556. https://doi.org/10.1017/S0269964826100266

https://doi.org/10.1017/S0269964826100266 Published online by Cambridge University Press